PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ

ESCUELA DE POSGRADO

MAESTRÍA EN ESTADÍSTICA

TÉCNICAS DE MUESTREO

PRACTICA III

Alumnos : Gregory Cesar Valderrama Vilca, 20133303

: Angel Rivera Solis, 20183726

: Juan Pablo Moreano, 20184093

: Rafael Visa Flores , 20184041

Recursos : Scripts en R y Bases de datos en correo

Lima - Perú

Lista de ejercicios 3 Final

(Técnicas de muestreo)

Indicaciones: El siguiente ejercicio debe de entregarse como máximo el Sábado 7 de Julio, día del examen final. Podrá realizarse en grupos de hasta de un máximo de 4 personas y equivaldrá, si está correctamente resuelta, a un máximo de 5 puntos de la nota del examen final, el cual se diseñará sobre 15 puntos.

Este ejercicio está basado en una actual investigación sobre el uso de cierta metodología de construcción BIM en el país. El estudio se hizo en la ciudad de Lima con el fin de poder estimar a un nivel de confianza del 95% y un error de 0.05, la proporción de obras en Lima que hacían uso de esta metodología. El diseño empleado fue uno estratificado por conglomerados bietápico. Los estratos estuvieron conformados por las divisiones urbanas: Lima Top, Lima Moderna, Lima Centro, Lima Norte, Lima Sur y el Callao. Cada estrato se dividió en sectores A, B y C de acuerdo al nivel socio-económico y dependiendo también si habían obras en ellas. En algunos casos se colapsaron sectores. En cada estrato se tomó un MASs de sectores y dentro de cada sector un MASs de obras. La data del muestreo y el marco muestral (estimado por datos de Capeco) se encuentra en la intranet bajo el nombre de DATAEX. La variable principal de investigación aquí es BIM, que indica si la obra encuestada hace o no uso de esta metodología. Se pide entonces lo siguiente:

a) Estime la proporción de obras de construcción en Lima Metropolitana que hacen uso de la metodología BIM, reportando su intervalo de confianza al 95%. (2.0 puntos)

b) Halle la estimación de la proporción del número de obras de construcción en Lima Top que hacen uso de la metodología BIM, junto con su error estándar de estimación estimado. (1.0 punto)

c) Suponga que en lugar de haberse empleado este diseño para Lima Top, usted hubiese empleado un muestreo ppt de 4 sectores, para luego, encuestar a todas las obras de los sectores de Lima Top seleccionados. Implemente este diseño, reportando la proporción del número de obras de construcción en Lima Top que hacen uso de la metodología BIM, junto con su error estándar de estimación estimado. Compare finalmente los errores de estimación de este diseño con los del anteriormente tomado.

NOTA: Una vez que seleccione el sector, use la estimación de DATAEX tomada para este sector a fin de imputar su proporción del uso del BIM. En caso que el sector no halla sido seleccionado en DATAEX (es decir, cuando vea que el número de obras encuestadas es 0), impute esta proporción simulando ella de una distribución Beta de parámetros $\alpha = 2$ y $\beta = 8$. (2.0 puntos)

Todo el trabajo lo deben de hacer en R, adjuntando los códigos respectivos.

L.H.V.S

a) Estime la proporción de obras de construcción en Lima Metropolitana que hacen uso de la metodología BIM, reportando su intervalo de confianza al 95%. (2.0 puntos)

Primero agregamos la información del marco muestral a la base de datos

```
37  library(sampling)
38  library(survey)
39  library(data.table)
40  library(foreign)
41
42  dataset <- read.csv(file= "DATAEX.csv", header=TRUE, sep=";")
43
44  head(dataset)
45  str(dataset)
46
47  levels(dataset$ESTRATO)
48  levels(dataset$DISTRITO)
49  levels(dataset$SECTOR)
50
51  datatable = as.data.table(dataset)
52  datatable <- datatable[ order ( datatable$ESTRATO ) , ]
53
54  table(datatable$ESTRATO)
55  table(datatable$DISTRITO)
56  table(datatable$SECTOR)
57
58  datatable [ , NESTRATO := 0]
59  datatable [ , NDISTRITO := 0]
60  datatable [ , NSECTOR := 0]
61
```

Por lo tanto NESTRATO, NDISTRITO, NSECTOR contienen la información numérica de las obras.

```r
# ================ TOTAL DE OBRAS POR ESTRATO ==============================

NESTRATO_LIMA_TOP <- 423  # total obras por estrato
NESTRATO_LIMA_MODERNA <- 360  # total obras por estrato
NESTRATO_LIMA_CENTRO <- 98  # total  obras por estrato
NESTRATO_LIMA_ESTE <- 75  # total  obras por estrato
NESTRATO_LIMA_NORTE <- 134  # total  obras por estrato
NESTRATO_LIMA_SUR <- 94  # total  obras por estrato
NESTRATO_CALLAO <- 32  # total  obras por estrato
datatable[which(datatable$ESTRATO == "LIMA TOP", arr.ind=T), "NESTRATO" ] = NESTRATO_LIMA_TOP
datatable[which(datatable$ESTRATO == "LIMA MODERNA", arr.ind=T), "NESTRATO" ] = NESTRATO_LIMA_MODERNA
datatable[which(datatable$ESTRATO == "LIMA CENTRO", arr.ind=T), "NESTRATO" ] = NESTRATO_LIMA_CENTRO
datatable[which(datatable$ESTRATO == "LIMA ESTE", arr.ind=T), "NESTRATO" ] = NESTRATO_LIMA_ESTE
datatable[which(datatable$ESTRATO == "LIMA NORTE", arr.ind=T), "NESTRATO" ] = NESTRATO_LIMA_NORTE
datatable[which(datatable$ESTRATO == "LIMA SUR", arr.ind=T), "NESTRATO" ] = NESTRATO_LIMA_SUR
datatable[which(datatable$ESTRATO == "CALLAO", arr.ind=T), "NESTRATO" ] = NESTRATO_CALLAO
```

```r
# ================ TOTAL DE OBRAS POR DISTRITO ==============================
# DISTRITO LIMA TOP
datatable[which(datatable$ESTRATO == "LIMA TOP" & datatable$DISTRITO == "MIRAFLORES"), "NDISTRITO" ] = 133
datatable[which(datatable$ESTRATO == "LIMA TOP" & datatable$DISTRITO == "SAN ISIDRO"), "NDISTRITO" ] = 58
#  molina
datatable[which(datatable$ESTRATO == "LIMA TOP" & datatable$DISTRITO == "SURCO"), "NDISTRITO" ] = 136
datatable[which(datatable$ESTRATO == "LIMA TOP" & datatable$DISTRITO == "SAN BORJA"), "NDISTRITO" ] = 53
datatable[which(datatable$ESTRATO == "LIMA TOP" & datatable$DISTRITO == "BARRANCO"), "NDISTRITO" ] = 34
# DISTRITO LIMA MODERNA
datatable[which(datatable$ESTRATO == "LIMA MODERNA" & datatable$DISTRITO == "JESUS MARIA"), "NDISTRITO" ] = 60
datatable[which(datatable$ESTRATO == "LIMA MODERNA" & datatable$DISTRITO == "LINCE"), "NDISTRITO" ] = 54
datatable[which(datatable$ESTRATO == "LIMA MODERNA" & datatable$DISTRITO == "MAGDALENA"), "NDISTRITO" ] = 64
datatable[which(datatable$ESTRATO == "LIMA MODERNA" & datatable$DISTRITO == "PUEBLO LIBRE"), "NDISTRITO" ] = 53
datatable[which(datatable$ESTRATO == "LIMA MODERNA" & datatable$DISTRITO == "SAN MIGUEL"), "NDISTRITO" ] = 82
datatable[which(datatable$ESTRATO == "LIMA MODERNA" & datatable$DISTRITO == "SURQUILLO"), "NDISTRITO" ] = 47
# DISTRITO LIMA CENTRO
datatable[which(datatable$ESTRATO == "LIMA CENTRO" & datatable$DISTRITO == "CERCADO LIMA"), "NDISTRITO" ] = 24
datatable[which(datatable$ESTRATO == "LIMA CENTRO" & datatable$DISTRITO == "BRENIA"), "NDISTRITO" ] = 34
datatable[which(datatable$ESTRATO == "LIMA CENTRO" & datatable$DISTRITO == "LA VICTORIA"), "NDISTRITO" ] = 37
# san luis 0
# DISTRITO LIMA ESTE
datatable[which(datatable$ESTRATO == "LIMA ESTE" & datatable$DISTRITO == "ATE"), "NDISTRITO" ] = 36
# Cieneguilla # Chaclacayo # Lurigancho
datatable[which(datatable$ESTRATO == "LIMA ESTE" & datatable$DISTRITO == "SANTA ANITA"), "NDISTRITO" ] = 7
datatable[which(datatable$ESTRATO == "LIMA ESTE" & datatable$DISTRITO == "EL AGUSTINO"), "NDISTRITO" ] = 3
# San Juan de Lurigancho
# DISTRITO NORTE
datatable[which(datatable$ESTRATO == "LIMA NORTE" & datatable$DISTRITO == "CARABAYLLO"), "NDISTRITO" ] = 23
datatable[which(datatable$ESTRATO == "LIMA NORTE" & datatable$DISTRITO == "COMAS"), "NDISTRITO" ] = 42
datatable[which(datatable$ESTRATO == "LIMA NORTE" & datatable$DISTRITO == "LOS OLIVOS"), "NDISTRITO" ] = 57
# puente piedra
datatable[which(datatable$ESTRATO == "LIMA NORTE" & datatable$DISTRITO == "SMP"), "NDISTRITO" ] = 6
datatable[which(datatable$ESTRATO == "LIMA NORTE" & datatable$DISTRITO == "ANCON/INDEP"), "NDISTRITO" ] = 2
# DISTRITO SUR
datatable[which(datatable$ESTRATO == "LIMA SUR" & datatable$DISTRITO == "CHORRILLOS"), "NDISTRITO" ] = 35
datatable[which(datatable$ESTRATO == "LIMA SUR" & datatable$DISTRITO == "LURIN/PACH/VES"), "NDISTRITO" ] = 13
datatable[which(datatable$ESTRATO == "LIMA SUR" & datatable$DISTRITO == "SAN JUAN MIRAFLORES"), "NDISTRITO" ] = 11
#Pucusana
datatable[which(datatable$ESTRATO == "LIMA SUR" & datatable$DISTRITO == "PUNTA HERMOSA/NEGRA"), "NDISTRITO" ] = 13
datatable[which(datatable$ESTRATO == "LIMA SUR" & datatable$DISTRITO == "SAN BARTOLO"), "NDISTRITO" ] = 14
# Santa Maria del Mar
# DISTRITO CALLAO
datatable[which(datatable$ESTRATO == "CALLAO" & datatable$DISTRITO == "BELLAVISTA"), "NDISTRITO" ] = 10
datatable[which(datatable$ESTRATO == "CALLAO" & datatable$DISTRITO == "CALLAO"), "NDISTRITO" ] = 14
# La perla
datatable[which(datatable$ESTRATO == "CALLAO" & datatable$DISTRITO == "VENTANILLA"), "NDISTRITO" ] = 2
```

```
130
131 # ================ TOTAL DE OBRAS POR SECTOR ============================
132 # DISTRITO LIMA TOP
133 datatable[which(datatable$ESTRATO == "LIMA TOP" & datatable$DISTRITO == "MIRAFLORES" & datatable$SECTOR == "A" ), "NSECTOR" ] = 34
134 datatable[which(datatable$ESTRATO == "LIMA TOP" & datatable$DISTRITO == "MIRAFLORES" & datatable$SECTOR == "B" ), "NSECTOR" ] = 77
135 datatable[which(datatable$ESTRATO == "LIMA TOP" & datatable$DISTRITO == "MIRAFLORES" & datatable$SECTOR == "C" ), "NSECTOR" ] = 23
136 datatable[which(datatable$ESTRATO == "LIMA TOP" & datatable$DISTRITO == "SAN ISIDRO" & datatable$SECTOR == "A" ), "NSECTOR" ] = 49
137 datatable[which(datatable$ESTRATO == "LIMA TOP" & datatable$DISTRITO == "SAN ISIDRO" & datatable$SECTOR == "B" ), "NSECTOR" ] = 9
138 #  molina
139 datatable[which(datatable$ESTRATO == "LIMA TOP" & datatable$DISTRITO == "SURCO" & datatable$SECTOR == "A" ), "NSECTOR" ] = 66
140 datatable[which(datatable$ESTRATO == "LIMA TOP" & datatable$DISTRITO == "SURCO" & datatable$SECTOR == "B" ), "NSECTOR" ] = 50
141 datatable[which(datatable$ESTRATO == "LIMA TOP" & datatable$DISTRITO == "SURCO" & datatable$SECTOR == "C" ), "NSECTOR" ] = 20
142 datatable[which(datatable$ESTRATO == "LIMA TOP" & datatable$DISTRITO == "SAN BORJA" & datatable$SECTOR == "A" ), "NSECTOR" ] = 19
143 datatable[which(datatable$ESTRATO == "LIMA TOP" & datatable$DISTRITO == "SAN BORJA" & datatable$SECTOR == "B" ), "NSECTOR" ] = 34
144 datatable[which(datatable$ESTRATO == "LIMA TOP" & datatable$DISTRITO == "SAN BORJA" & datatable$SECTOR == "C" ), "NSECTOR" ] = 1
145 datatable[which(datatable$ESTRATO == "LIMA TOP" & datatable$DISTRITO == "BARRANCO" & datatable$SECTOR == "A" ), "NSECTOR" ] = 34
146 # DISTRITO LIMA MODERNA
147 datatable[which(datatable$ESTRATO == "LIMA MODERNA" & datatable$DISTRITO == "JESUS MARIA" & datatable$SECTOR == "A" ), "NSECTOR" ] = 31
148 datatable[which(datatable$ESTRATO == "LIMA MODERNA" & datatable$DISTRITO == "JESUS MARIA" & datatable$SECTOR == "B" ), "NSECTOR" ] = 24
149 datatable[which(datatable$ESTRATO == "LIMA MODERNA" & datatable$DISTRITO == "JESUS MARIA" & datatable$SECTOR == "C" ), "NSECTOR" ] = 6
150 datatable[which(datatable$ESTRATO == "LIMA MODERNA" & datatable$DISTRITO == "LINCE" & datatable$SECTOR == "A" ), "NSECTOR" ] = 54
151 datatable[which(datatable$ESTRATO == "LIMA MODERNA" & datatable$DISTRITO == "MAGDALENA" & datatable$SECTOR == "A" ), "NSECTOR" ] = 34
152 datatable[which(datatable$ESTRATO == "LIMA MODERNA" & datatable$DISTRITO == "MAGDALENA" & datatable$SECTOR == "C" ), "NSECTOR" ] = 30
153 datatable[which(datatable$ESTRATO == "LIMA MODERNA" & datatable$DISTRITO == "PUEBLO LIBRE" & datatable$SECTOR == "A" ), "NSECTOR" ] = 22
154 datatable[which(datatable$ESTRATO == "LIMA MODERNA" & datatable$DISTRITO == "PUEBLO LIBRE" & datatable$SECTOR == "B" ), "NSECTOR" ] = 17
155 datatable[which(datatable$ESTRATO == "LIMA MODERNA" & datatable$DISTRITO == "PUEBLO LIBRE" & datatable$SECTOR == "C" ), "NSECTOR" ] = 13
156 datatable[which(datatable$ESTRATO == "LIMA MODERNA" & datatable$DISTRITO == "SAN MIGUEL" & datatable$SECTOR == "A" ), "NSECTOR" ] = 19
157 datatable[which(datatable$ESTRATO == "LIMA MODERNA" & datatable$DISTRITO == "SAN MIGUEL" & datatable$SECTOR == "B" ), "NSECTOR" ] = 19
158 datatable[which(datatable$ESTRATO == "LIMA MODERNA" & datatable$DISTRITO == "SAN MIGUEL" & datatable$SECTOR == "C" ), "NSECTOR" ] = 44
159 datatable[which(datatable$ESTRATO == "LIMA MODERNA" & datatable$DISTRITO == "SURQUILLO" & datatable$SECTOR == "A" ), "NSECTOR" ] = 13
160 datatable[which(datatable$ESTRATO == "LIMA MODERNA" & datatable$DISTRITO == "SURQUILLO" & datatable$SECTOR == "B" ), "NSECTOR" ] = 6
161 datatable[which(datatable$ESTRATO == "LIMA MODERNA" & datatable$DISTRITO == "SURQUILLO" & datatable$SECTOR == "C" ), "NSECTOR" ] = 28
162 # DISTRITO LIMA CENTRO
163 datatable[which(datatable$ESTRATO == "LIMA CENTRO" & datatable$DISTRITO == "CERCADO LIMA" & datatable$SECTOR == "A" ), "NSECTOR" ] = 14
164 datatable[which(datatable$ESTRATO == "LIMA CENTRO" & datatable$DISTRITO == "CERCADO LIMA" & datatable$SECTOR == "B" ), "NSECTOR" ] = 10
165 datatable[which(datatable$ESTRATO == "LIMA CENTRO" & datatable$DISTRITO == "BRENIA" & datatable$SECTOR == "A" ), "NSECTOR" ] = 23
166 datatable[which(datatable$ESTRATO == "LIMA CENTRO" & datatable$DISTRITO == "BRENIA" & datatable$SECTOR == "B" ), "NSECTOR" ] = 4
167 datatable[which(datatable$ESTRATO == "LIMA CENTRO" & datatable$DISTRITO == "BRENIA" & datatable$SECTOR == "C" ), "NSECTOR" ] = 6
168 datatable[which(datatable$ESTRATO == "LIMA CENTRO" & datatable$DISTRITO == "LA VICTORIA" & datatable$SECTOR == "A" ), "NSECTOR" ] = 37
169 # san luis 0
```

```
169 # san luis 0
170 # DISTRITO LIMA ESTE
171 datatable[which(datatable$ESTRATO == "LIMA ESTE" & datatable$DISTRITO == "ATE" & datatable$SECTOR == "A" ), "NSECTOR" ] = 36
172 # Cieneguilla # Chaclacayo # Lurigancho
173 datatable[which(datatable$ESTRATO == "LIMA ESTE" & datatable$DISTRITO == "SANTA ANITA" & datatable$SECTOR == "A" ), "NSECTOR" ] = 7
174 datatable[which(datatable$ESTRATO == "LIMA ESTE" & datatable$DISTRITO == "EL AGUSTINO" & datatable$SECTOR == "A" ), "NSECTOR" ] = 3
175 # San Juan de Lurigancho
176 # DISTRITO NORTE
177 datatable[which(datatable$ESTRATO == "LIMA NORTE" & datatable$DISTRITO == "CARABAYLLO" & datatable$SECTOR == "A" ), "NSECTOR" ] = 23
178 datatable[which(datatable$ESTRATO == "LIMA NORTE" & datatable$DISTRITO == "COMAS" & datatable$SECTOR == "A" ), "NSECTOR" ] = 42
179 datatable[which(datatable$ESTRATO == "LIMA NORTE" & datatable$DISTRITO == "LOS OLIVOS" & datatable$SECTOR == "A" ), "NSECTOR" ] = 10
180 datatable[which(datatable$ESTRATO == "LIMA NORTE" & datatable$DISTRITO == "LOS OLIVOS" & datatable$SECTOR == "B" ), "NSECTOR" ] = 44
181 datatable[which(datatable$ESTRATO == "LIMA NORTE" & datatable$DISTRITO == "LOS OLIVOS" & datatable$SECTOR == "C" ), "NSECTOR" ] = 3
182 # puente piedra
183 datatable[which(datatable$ESTRATO == "LIMA NORTE" & datatable$DISTRITO == "SMP" & datatable$SECTOR == "A" ), "NSECTOR" ] = 6
184 datatable[which(datatable$ESTRATO == "LIMA NORTE" & datatable$DISTRITO == "ANCON/INDEP" & datatable$SECTOR == "A" ), "NSECTOR" ] = 4
185 # DISTRITO SUR
186 datatable[which(datatable$ESTRATO == "LIMA SUR" & datatable$DISTRITO == "CHORRILLOS" & datatable$SECTOR == "A" ), "NSECTOR" ] = 5
187 datatable[which(datatable$ESTRATO == "LIMA SUR" & datatable$DISTRITO == "CHORRILLOS" & datatable$SECTOR == "C" ), "NSECTOR" ] = 30
188 datatable[which(datatable$ESTRATO == "LIMA SUR" & datatable$DISTRITO == "LURIN/PACH/VES" & datatable$SECTOR == "A" ), "NSECTOR" ] = 13
189 datatable[which(datatable$ESTRATO == "LIMA SUR" & datatable$DISTRITO == "SAN JUAN MIRAFLORES" & datatable$SECTOR == "A" ), "NSECTOR" ] = 11
190 #Pucusana
191 datatable[which(datatable$ESTRATO == "LIMA SUR" & datatable$DISTRITO == "PUNTA HERMOSA/NEGRA" & datatable$SECTOR == "A" ), "NSECTOR" ] = 13
192 datatable[which(datatable$ESTRATO == "LIMA SUR" & datatable$DISTRITO == "SAN BARTOLO" & datatable$SECTOR == "A" ), "NSECTOR" ] = 14
193 # Santa Maria del Mar
194 # DISTRITO CALLAO
195 datatable[which(datatable$ESTRATO == "CALLAO" & datatable$DISTRITO == "BELLAVISTA" & datatable$SECTOR == "A" ), "NSECTOR" ] = 10
196 datatable[which(datatable$ESTRATO == "CALLAO" & datatable$DISTRITO == "CALLAO" & datatable$SECTOR == "A" ), "NSECTOR" ] = 14
197 datatable[which(datatable$ESTRATO == "CALLAO" & datatable$DISTRITO == "VENTANILLA" & datatable$SECTOR == "A" ), "NSECTOR" ] = 2
198 # La perla
199
200
```

Adicionalmente una columna NCONG que tiene el número de conglomerados por estrato.

```
datatable[which(datatable$ESTRATO == "LIMA TOP" ), "NCONG" ] = 14
datatable[which(datatable$ESTRATO == "LIMA MODERNA" ), "NCONG" ] = 15
datatable[which(datatable$ESTRATO == "LIMA CENTRO" ), "NCONG" ] = 7
datatable[which(datatable$ESTRATO == "LIMA ESTE" ), "NCONG" ] = 7
datatable[which(datatable$ESTRATO == "LIMA NORTE" ), "NCONG" ] = 8
datatable[which(datatable$ESTRATO == "LIMA SUR" ), "NCONG" ] = 8
datatable[which(datatable$ESTRATO == "CALLAO" ), "NCONG" ] = 4
```

Con esto hemos introducido a nuestra tabla toda la información de nuestro marco muestral

	A	B	C	D	E	F	G	H	I
1	LOCALIZACIÓN			Número de	Número de obras	Encuestas		Conglomerados	
2	SECTOR URBANO(SU)	DISTRITO	GRUPO	Obras por distrito	por sector	por grupo			
3	Lima Top	Miraflores	A	133	34	4	A	208	
4		Miraflores	B	133	77	21	B	172	
5		Miraflores	C	133	23	0	C	44	
6		San Isidro	A	58	49	20			
7		San Isidro	B	58	9	2			
8		La Molina	A	9	6	0			
9		La Molina	B	9	2	0			
10		Santiago de Surco	A	136	66	21			
11		Santiago de Surco	B	136	50	20			
12		Santiago de Surco	C	136	20	0			
13		San Borja	A	53	19	16			
14		San Borja	B	53	34	7			
15		San Borja	C	53	1	0			
16		Barranco** (AyB)	A	34	34	17			
17				423	424	128		424	

A continuación calculamos los Factores de corrección para poblaciones finitas (FPC) para nuestro muestreo por conglomerados bietápico parte de nuestro muestreo complejo. Como el valor de número de conglomerados por estrato y número de obras por sector.

A continuación definimos el diseño de nuestro muestreo.

```
213
214 datatable [ , FPC := NCONG ]
215 datatable [ , FPC2 := NSECTOR ]
216
217 design = svydesign(id=~CONG + NUM, fpc = ~FPC + FPC2 , strata = ~ESTRATO, data = datatable)
218 design
219
```

```
Stratified 2 - level Cluster Sampling design
With (42, 319) clusters.
svydesign(id = ~CONG + NUM, fpc = ~FPC + FPC2, strata = ~ESTRATO,
    data = datatable)
```

Estimamos la media para obtener la proporción de obras que usan la metodología BIM

```
# estaimacion por metodo tradicional linealizacion
mean = svymean(~BIM, design = design, deff = T)
mean
confint(mean)
```

```
> mean
          mean       SE   DEff
BIMNO 0.754909 0.030367 1.9999
BIMSI 0.245091 0.030367 1.9999
> confint(mean)
          2.5 %    97.5 %
BIMNO 0.6953899 0.8144282
BIMSI 0.1855718 0.3046101
>
```

b) Halle la estimación de la proporción del número de obras de construcción en Lima Top que hacen uso de la metodología BIM, junto con su error estándar de estimación estimado. (1.0 punto)

En la parte B como tenemos toda la informacion en la tabla de datos , solo tomamos la parte de la muestra referente a lima top

```
31
32  # PART B
33  top_sample = datatable[which(datatable$ESTRATO == "LIMA TOP" ),  ]
34  top_design = svydesign(id=~~CONG + NUM, fpc = ~FPC + FPC2 , data = top_sample)
35  top_mean = svymean(~BIM, design = top_design, deff = T)
36  top_mean
37  confint(top_mean)
38
```

```
          mean       SE   DEff
BIMNO 0.773261 0.037735 1.3244
BIMSI 0.226739 0.037735 1.3244
> confint(top_mean)
          2.5 %    97.5 %
BIMNO 0.6993006 0.8472208
BIMSI 0.1527792 0.3006994
>
```

c) Suponga que en lugar de haberse empleado este diseño para Lima Top, usted hubiese empleado un muestreo ppt de 4 sectores, para luego, encuestar a todas las obras de los sectores de Lima Top seleccionados. Implemente este diseño, reportando la proporción del número de obras de construcción en Lima Top que hacen uso de la metodología BIM, junto con su error estándar de estimación estimado. Compare finalmente los errores de estimación de este diseño con los del anteriormente tomado.

NOTA: Una vez que seleccione el sector, use la estimación de DATAEX tomada para este sector a fin de imputar su proporción del uso del BIM. En caso que el sector no halla sido seleccionado en DATAEX (es decir, cuando vea que el número de obras encuestadas es 0), impute esta proporción simulando ella de una distribución Beta de parámetros $\alpha = 2$ y $\beta = 8$. (2.0 puntos)

El marco muestral tiene algunos sectores en los distritos cuyas encuestas no han sido contestadas, por esta razón utilizamos valores desde la distribución beta para imputar dichos valores, a continuación completamos las muestras faltantes.

LOCALIZACIÓN			Número de	Número de obras	Encuestas
SECTOR URBANO(SU)	DISTRITO	GRUPO	Obras por distrito	por sector	por grupo
Lima Top	Miraflores	A	133	34	4
	Miraflores	B	133	77	21
	Miraflores	C	133	23	0
	San Isidro	A	58	49	20
	San Isidro	B	58	9	2
	La Molina	A	9	6	0
	La Molina	B	9	2	0
	Santiago de Surco	A	136	66	21
	Santiago de Surco	B	136	50	20
	Santiago de Surco	C	136	20	0
	San Borja	A	53	19	16
	San Borja	B	53	34	7
	San Borja	C	53	1	0
	Barranco** (AyB)	A	34	34	17
			423	424	128

Utilizamos la función Beta para calcular los valores y estimar la proporción faltante y completar los valores.

```
# imputation miraflores c 133 obras en el distrito, 23 obras en sector C
n_distrito_sector = 23
prop <- rbeta(1, 2, 8)
prop_yes <- ceiling(n_distrito_sector * prop)
prop_no <- n_distrito_sector - prop_yes
for (index in rep(1,prop_yes)) {
  ppt_sample = rbind(ppt_sample, list( nrow(ppt_sample) + 1, "SI", "LIMA TOP", "MIRAFLORES", "C", "MIRAFLORESC", 133, n_distrito_sector, 1 ) )
}
for (index in rep(1,prop_no)) {
  ppt_sample = rbind(ppt_sample, list( nrow(ppt_sample) + 1, "NO", "LIMA TOP", "MIRAFLORES", "C", "MIRAFLORESC", 133, n_distrito_sector, 1 ) )
}
```

```
# imputation molina A 9 obras en el distrito, 6 obras en sector A
n_distrito_sector = 6
prop <- rbeta(1, 2, 8)
prop_yes <- ceiling(n_distrito_sector * prop)
prop_no <- n_distrito_sector - prop_yes
for (index in rep(1,prop_yes)) {
  ppt_sample = rbind(ppt_sample, list( nrow(ppt_sample) + 1, "SI", "LIMA TOP", "MOLINA", "A", "MOLINAA", 9, n_distrito_sector, 1 ) )
}
for (index in rep(1,prop_no)) {
  ppt_sample = rbind(ppt_sample, list( nrow(ppt_sample) + 1, "NO", "LIMA TOP", "MOLINA", "A", "MOLINAA", 9, n_distrito_sector, 1 ) )
}
```

```
# imputation molina B 9 obras en el distrito, 2 obras en sector B
n_distrito_sector = 2
prop <- rbeta(1, 2, 8)
prop_yes <- ceiling(n_distrito_sector * prop)
prop_no <- n_distrito_sector - prop_yes
for (index in rep(1,prop_yes)) {
  ppt_sample = rbind(ppt_sample, list( nrow(ppt_sample) + 1, "SI", "LIMA TOP", "MOLINA", "B", "MOLINAB", 9, n_distrito_sector, 1 ) )
}
for (index in rep(1,prop_no)) {
  ppt_sample = rbind(ppt_sample, list( nrow(ppt_sample) + 1, "NO", "LIMA TOP", "MOLINA", "B", "MOLINAB", 9, n_distrito_sector, 1 ) )
}

```

```
# imputation santiago de surco C 136 obras en el distrito, 20 obras en sector C
n_distrito_sector = 20
prop <- rbeta(1, 2, 8)
prop_yes <- ceiling(n_distrito_sector * prop)
prop_no <- n_distrito_sector - prop_yes
for (index in rep(1,prop_yes)) {
  ppt_sample = rbind(ppt_sample, list( nrow(ppt_sample) + 1, "SI", "LIMA TOP", "SURCO", "C", "SURCOC", 136, n_distrito_sector, 1 ) )
}
for (index in rep(1,prop_no)) {
  ppt_sample = rbind(ppt_sample, list( nrow(ppt_sample) + 1, "NO", "LIMA TOP", "SURCO", "C", "SURCOC", 136, n_distrito_sector, 1 ) )
}
```

```r
# imputation san borja C 53 obras en el distrito, 1 obras en sector C
n_distrito_sector = 1
prop <- rbeta(1, 2, 8)
prop_yes <- ceiling(n_distrito_sector * prop)
prop_no <- n_distrito_sector - prop_yes
for (index in rep(1,prop_yes)) {
  ppt_sample = rbind(ppt_sample, list( nrow(ppt_sample) + 1, "SI", "LIMA TOP", "SAN BORJA", "C", "SAN BORJAC", 53, n_distrito_sector, 1 ) )
}
for (index in rep(1,prop_no)) {
  ppt_sample = rbind(ppt_sample, list( nrow(ppt_sample) + 1, "NO", "LIMA TOP", "SAN BORJA", "C", "SAN BORJAC", 53, n_distrito_sector, 1 ) )
}

```

Ahora calculamos las proporciones para cada distrito sector dentro de lima top

```r

ppt_sample$CONG = factor(ppt_sample$CONG)
ppt_sample <- ppt_sample[order( ppt_sample$CONG ) , ]
data.frame(table(ppt_sample$CONG))[ ,2 ]
DIS_SECTOR = sort(sapply(unique(ppt_sample$CONG), as.character))
TAM_SECTOR = vector(length = length(DIS_SECTOR))
BIM_SI_SECTOR = vector(length = length(DIS_SECTOR))
BIM_NO_SECTOR = vector(length = length(DIS_SECTOR))
PROP_SECTOR = vector(length = length(DIS_SECTOR))
i = 1
for (sector in DIS_SECTOR) {
  TAM_SECTOR[i] = dim(ppt_sample[ which(ppt_sample$CONG == sector), ])[1]
  BIM_SI_SECTOR[i] = dim(ppt_sample[ which(ppt_sample$CONG == sector & ppt_sample$BIM == "SI"), ])[1]
  BIM_NO_SECTOR[i] = dim(ppt_sample[ which(ppt_sample$CONG == sector & ppt_sample$BIM == "NO"), ])[1]
  PROP_SECTOR[i] =   BIM_SI_SECTOR[i]  / (BIM_SI_SECTOR[i] + BIM_NO_SECTOR[i]) # calculate proportions
  i = i + 1
}
```

Con lo cual tenemos una base de datos igual a :

	DIS_SECTOR	TAM_SECTOR	BIM_SI_SECTOR	BIM_NO_SECTOR	PROP_SECTOR
1	BARRANCOA	17	3	14	0.1764706
2	MIRAFLORESA	4	1	3	0.2500000
3	MIRAFLORESB	21	7	14	0.3333333
4	MIRAFLORESC	23	7	16	0.3043478
5	MOLINAA	6	2	4	0.3333333
6	MOLINAB	2	1	1	0.5000000
7	SAN BORJAA	16	1	15	0.0625000
8	SAN BORJAB	7	2	5	0.2857143
9	SAN BORJAC	1	1	0	1.0000000
10	SAN ISIDROA	20	3	17	0.1500000
11	SAN ISIDROB	2	2	0	1.0000000
12	SURCOA	21	3	18	0.1428571
13	SURCOB	20	3	17	0.1500000
14	SURCOC	20	3	17	0.1500000

Debido a que necesitamos hallar el error de estimación, utilizaremos la técnica de muestreo secuencial ppt que nos permite estimar la varianza del estimador HT.

Utilizaremos las probabilidades acumuladas y un valor aleatorio como semilla para la selección de la muestra de 4 distrito sectores. Hemos ordenado los distrito sectores alfabeticamente.

```
phi = TAM_SECTOR / sum(TAM_SECTOR)
phi2 =  cumsum(phi)

ppt_obras = data.frame(DIS_SECTOR, TAM_SECTOR, BIM_SI_SECTOR, BIM_NO_SECTOR,  PROP_SECTOR)
ppt_obras = as.data.table(ppt_obras)

ppt_obras[, phi :=  TAM_SECTOR / sum(TAM_SECTOR) ]
ppt_obras[, phi2 := cumsum(phi) ]
```

	DIS_SECTOR	TAM_SECTOR	BIM_SI_SECTOR	BIM_NO_SECTOR	PROP_SECTOR	phi	phi2
1:	BARRANCOA	17	3	14	0.1764706	0.094444444	0.09444444
2:	MIRAFLORESA	4	1	3	0.2500000	0.022222222	0.11666667
3:	MIRAFLORESB	21	7	14	0.3333333	0.116666667	0.23333333
4:	MIRAFLORESC	23	7	16	0.3043478	0.127777778	0.36111111
5:	MOLINAA	6	2	4	0.3333333	0.033333333	0.39444444
6:	MOLINAB	2	1	1	0.5000000	0.011111111	0.40555556
7:	SAN BORJAA	16	1	15	0.0625000	0.088888889	0.49444444
8:	SAN BORJAB	7	2	5	0.2857143	0.038888889	0.53333333
9:	SAN BORJAC	1	1	0	1.0000000	0.005555556	0.53888889
10:	SAN ISIDROA	20	3	17	0.1500000	0.111111111	0.65000000
11:	SAN ISIDROB	2	2	0	1.0000000	0.011111111	0.66111111
12:	SURCOA	21	3	18	0.1428571	0.116666667	0.77777778
13:	SURCOB	20	3	17	0.1500000	0.111111111	0.88888889
14:	SURCOC	20	3	17	0.1500000	0.111111111	1.00000000

Los números aleatorios generados son:

runif(4) => 0.11016850 0.95160297 0.02318828 0.43512783

Procedemos a seleccionar el elemento segundo pues es el primer elemento superior a 0.11016850 , con lo que nuestra data quedará en :

	DIS_SECTOR	TAM_SECTOR	BIM_SI_SECTOR	BIM_NO_SECTOR	PROP_SECTOR	phi	phi2
1:	BARRANCOA	17	3	14	0.1764706	0.096590909	0.09659091
2:	MIRAFLORESB	21	7	14	0.3333333	0.119318182	0.21590909
3:	MIRAFLORESC	23	7	16	0.3043478	0.130681818	0.34659091
4:	MOLINAA	6	2	4	0.3333333	0.034090909	0.38068182
5:	MOLINAB	2	1	1	0.5000000	0.011363636	0.39204545
6:	SAN BORJAA	16	1	15	0.0625000	0.090909091	0.48295455
7:	SAN BORJAB	7	2	5	0.2857143	0.039772727	0.52272727
8:	SAN BORJAC	1	1	0	1.0000000	0.005681818	0.52840909
9:	SAN ISIDROA	20	3	17	0.1500000	0.113636364	0.64204545
10:	SAN ISIDROB	2	2	0	1.0000000	0.011363636	0.65340909
11:	SURCOA	21	3	18	0.1428571	0.119318182	0.77272727
12:	SURCOB	20	3	17	0.1500000	0.113636364	0.88636364
13:	SURCOC	20	3	17	0.1500000	0.113636364	1.00000000

Recalculadas las nuevas probabilidades, tomamos el 13.

	DIS_SECTOR	TAM_SECTOR	BIM_SI_SECTOR	BIM_NO_SECTOR	PROP_SECTOR	phi	phi2
1:	BARRANCOA	17	3	14	0.1764706	0.108974359	0.1089744
2:	MIRAFLORESB	21	7	14	0.3333333	0.134615385	0.2435897
3:	MIRAFLORESC	23	7	16	0.3043478	0.147435897	0.3910256
4:	MOLINAA	6	2	4	0.3333333	0.038461538	0.4294872
5:	MOLINAB	2	1	1	0.5000000	0.012820513	0.4423077
6:	SAN BORJAA	16	1	15	0.0625000	0.102564103	0.5448718
7:	SAN BORJAB	7	2	5	0.2857143	0.044871795	0.5897436
8:	SAN BORJAC	1	1	0	1.0000000	0.006410256	0.5961538
9:	SAN ISIDROA	20	3	17	0.1500000	0.128205128	0.7243590
10:	SAN ISIDROB	2	2	0	1.0000000	0.012820513	0.7371795
11:	SURCOA	21	3	18	0.1428571	0.134615385	0.8717949
12:	SURCOB	20	3	17	0.1500000	0.128205128	1.0000000

Para luego tomar el primer elemento.

	DIS_SECTOR	TAM_SECTOR	BIM_SI_SECTOR	BIM_NO_SECTOR	PROP_SECTOR	phi	phi2
1:	MIRAFLORESB	21	7	14	0.3333333	0.151079137	0.1510791
2:	MIRAFLORESC	23	7	16	0.3043478	0.165467626	0.3165468
3:	MOLINAA	6	2	4	0.3333333	0.043165468	0.3597122
4:	MOLINAB	2	1	1	0.5000000	0.014388489	0.3741007
5:	SAN BORJAA	16	1	15	0.0625000	0.115107914	0.4892086
6:	SAN BORJAB	7	2	5	0.2857143	0.050359712	0.5395683
7:	SAN BORJAC	1	1	0	1.0000000	0.007194245	0.5467626
8:	SAN ISIDROA	20	3	17	0.1500000	0.143884892	0.6906475
9:	SAN ISIDROB	2	2	0	1.0000000	0.014388489	0.7050360
10:	SURCOA	21	3	18	0.1428571	0.151079137	0.8561151
11:	SURCOB	20	3	17	0.1500000	0.143884892	1.0000000

Y por ultimo total el 5 elemento.

Con lo que nuestra muestra está dada por :

	ID_unit	DIS_SECTOR	TAM_SECTOR	BIM_SI_SECTOR	BIM_NO_SECTOR	PROP_SECTOR
1	1	BARRANCOA	17	3	14	0.1764706
2	2	MIRAFLORESA	4	1	3	0.2500000
5	5	MOLINAA	6	2	4	0.3333333
13	13	SURCOB	20	3	17	0.1500000

Ahora utilizamos la función para calcular las probabilidades de primer y segundo orden.

```
library(combinat)
pisppt <- function(X,n) {
  N = length(X)
  XT = sum(X)
  m = combn(X,n) # Requiere del paquete combinat
  m = apply(m,2,permn)
  m = matrix(unlist(m),ncol=n,byrow=TRUE)
  nm = dim(m)[1]
  p=0
  for (j in 1:nm) {
    p[j] = prod(m[j,])/(XT*prod(XT-cumsum(m[j,1:n-1])))
  }
  pi1=0
  pi2=matrix(0,N,N)
  for (i in 1:(N-1)){
    aux1 = (m==X[i])
    index = which(apply(1*aux1,1,sum)==1)
    pi1[i] = sum(p[index])
    for (j in (i+1):N){
      aux2 = (m==X[j])
      aux2 = 1*aux2[index,]
      pi2[i,j] = sum(p[index[which(apply(aux2,1,sum)==1)]])
    }}
  pi1[N] = n-sum(pi1)
  pi2 = pi2+t(pi2)
  list(pi1,pi2)
}
```

```
index = rep(0, 14)
index[2] = 1
index[13] = 1
index[1] = 1
index[5] = 1

ppt_obras_sample = getdata(ppt_obras, index)
probs = pisppt(as.numeric( TAM_SECTOR ),4)
ppt_obras_pik = getdata(probs[[1]], as.logical(index))[, "data"]
```

La estimación del Total de la proporción y por lo tanto la división entre el número de distrito sectores para obtener la proporción.

```
bim_result = HTestimator(ppt_obras_sample[,"PROP_SECTOR"], ppt_obras_pik) / 14
bim_result
```

```
> bim_result
          [,1]
[1,] 0.3862333
>
```

Calculamos la varianza del estimador HT para estimar el error utilizando las probabilidades de segundo orden.

```
pik_2 =  probs[[2]][as.logical(index), as.logical(index)]
diag(pik_2) = ppt_obras_pik
se =  sqrt( varHT(ppt_obras_sample[,"PROP_SECTOR"] , pik_2) ) / 14
se

alpha = 0.05
z <- qnorm ( 1 - alpha / 2 )
c(bim_result - z * se, bim_result + se)
```

```
> pik_2 =  probs[[2]][as.logical(index), as.logical(index)]
> diag(pik_2) = ppt_obras_pik
> se =  sqrt( varHT(ppt_obras_sample[,"PROP_SECTOR"] , pik_2) ) / 14
> se
[1] 0.1461025
>
> alpha = 0.05
> z <- qnorm ( 1 - alpha / 2 )
> c(bim_result - z * se, bim_result + se)
[1] 0.09987759 0.53233587
```

Como podemos apreciar la estimación del valor de proporción para el uso de BIM SI es del 38% y un margen de error de 14%, Con lo que podemos calcular el IC que nos demuestra que va desde el 9% hasta el 53%. La precisión del estimador es más baja que la del muestreo por conglomerados bietápico pues el error en ese caso fue del 3%.