Inferencia Estadística: Proyecto Parte B

6 de julio de 2022

Universidad Nacional de Colombia

Profesor: Mario Enrique Arrieta Prieto

Ander Steven Cristancho Sánchez John Anderson Guarin Lopez
Juan David Carrascal Ibañez

Problema 1:

A. Comparación de varios intervalos de confianza para una proporción en una muestra aleatoria Bernoulli.

Solución:

1. Para llegar al intervalo de confianza del la posibilidad 1, se siguió el siguiente procedimiento.

   Tenemos que:

$$p\left(-z_{1-\frac{\alpha}{2}} \leq \frac{\sqrt{n}(\hat{p_n}-p)}{\sqrt{p(1-p)}} \leq z_{1-\frac{\alpha}{2}}\right) \approx 1-\alpha$$

   De aquí, haremos lo siguiente:

$$\begin{aligned}
p\left(-z_{1-\frac{\alpha}{2}} \leq \frac{\sqrt{n}(\hat{p_n}-p)}{\sqrt{p(1-p)}} \leq z_{1-\frac{\alpha}{2}}\right) &= p\left(\left|\frac{\sqrt{n}(\hat{p_n}-p)}{\sqrt{p(1-p)}}\right| \leq z_{1-\frac{\alpha}{2}}\right) \\
&= p\left(\left|\frac{\sqrt{n}(\hat{p_n}-p)}{\sqrt{p(1-p)}}\right|^2 \leq z^2_{1-\frac{\alpha}{2}}\right) \\
&= p\left(n(\hat{p_n}-p)^2 \leq z^2_{1-\frac{\alpha}{2}}(p(1-p))\right) \\
&= p\left(n\hat{p_n}^2 - 2n\hat{p_n}p + np^2 \leq z^2_{1-\frac{\alpha}{2}}p - z^2_{1-\frac{\alpha}{2}}p^2\right) \\
&= p\left(n\hat{p_n}^2 - 2n\hat{p_n}p + np^2 - z^2_{1-\frac{\alpha}{2}}p + z^2_{1-\frac{\alpha}{2}}p^2 \leq 0\right) \\
&= p\left(p^2(n+z^2_{1-\frac{\alpha}{2}}) - p(2n\hat{p_n} + z^2_{1-\frac{\alpha}{2}}) + \hat{p_n}^2 \leq 0\right) \\
&\approx 1-\alpha
\end{aligned}$$

   Ahora, resolvemos la desigualdad para p, usando la fórmula cuadrática, con lo que obtenemos que:

$$
\begin{aligned}
p &= \frac{(2n\hat{p_n} + z^2_{1-\frac{\alpha}{2}}) \pm \sqrt{(2n\hat{p_n} + z^2_{1-\frac{\alpha}{2}})^2 - 4(n + z^2_{1-\frac{\alpha}{2}})(\hat{p_n}^2)}}{2(n + z^2_{1-\frac{\alpha}{2}})} \\
&= \frac{(2n\hat{p_n} + z^2_{1-\frac{\alpha}{2}}) \pm \sqrt{4n^2\hat{p_n}^2 + 4n\hat{p_n}z^2_{1-\frac{\alpha}{2}} + z^4_{1-\frac{\alpha}{2}} - 4n^2\hat{p_n}^2 - 4n\hat{p_n}^2 z^2_{1-\frac{\alpha}{2}}}}{2(n + z^2_{1-\frac{\alpha}{2}})} \\
&= \frac{(2n\hat{p_n} + z^2_{1-\frac{\alpha}{2}}) \pm \sqrt{4n\hat{p_n}z^2_{1-\frac{\alpha}{2}} + z^4_{1-\frac{\alpha}{2}} - 4n\hat{p_n}^2 z^2_{1-\frac{\alpha}{2}}}}{2(n + z^2_{1-\frac{\alpha}{2}})} \\
&= \frac{(2n\hat{p_n} + z^2_{1-\frac{\alpha}{2}}) \pm \sqrt{z^2_{1-\frac{\alpha}{2}}(4n\hat{p_n} + z^2_{1-\frac{\alpha}{2}} - 4n\hat{p_n}^2)}}{2(n + z^2_{1-\frac{\alpha}{2}})} \\
&= \frac{(2n\hat{p_n} + z^2_{1-\frac{\alpha}{2}}) \pm z_{1-\frac{\alpha}{2}}\sqrt{4n\hat{p_n} + z^2_{1-\frac{\alpha}{2}} - 4n\hat{p_n}^2}}{2(n + z^2_{1-\frac{\alpha}{2}})}
\end{aligned}
$$

Ahora, como el coeficiente de p^2 es positivo y queremos los valores que hacen que esa inecuación sea menor o igual a 0, la desigualdad se cumple para cualquier p entre el intervalo conformado por las dos soluciones, a las que llamaremos $L(\hat{p_n})$ siendo el límite inferior del intervalo, y $U(\hat{p_n})$ siendo el límite superior del intervalo. Por tanto, $(L(\hat{p_n}), U(\hat{p_n}))$ es un intervalo de confianza aproximado de nivel $1 - \alpha$ para p, el cual es dado por:

$$ICA_{100(1-\alpha)\,\%}(p) = (L(\hat{p_n}), U(\hat{p_n}))$$

Siendo $L(\hat{p_n})$:

$$L(\hat{p_n}) = \frac{(2n\hat{p_n} + z^2_{1-\frac{\alpha}{2}}) - z_{1-\frac{\alpha}{2}}\sqrt{4n\hat{p_n} + z^2_{1-\frac{\alpha}{2}} - 4n\hat{p_n}^2}}{2(n + z^2_{1-\frac{\alpha}{2}})}$$

Y siendo $U(\hat{p_n})$:

$$U(\hat{p_n}) = \frac{(2n\hat{p_n} + z^2_{1-\frac{\alpha}{2}}) + z_{1-\frac{\alpha}{2}}\sqrt{4n\hat{p_n} + z^2_{1-\frac{\alpha}{2}} - 4n\hat{p_n}^2}}{2(n + z^2_{1-\frac{\alpha}{2}})}$$

2. **Implementación del algoritmo** Presentaremos el algoritmo completo en el informe pues consideramos que fue de vital importancia en el desarrollo del proyecto. En el código correspondiente a este punto descrimbimos los pasos más releventes.

```
alpha <- 0.05
a<-seq(from = 0.05,to = 0.95,by = (0.05))
b<-c(5,10,50,100,200,500,1000)
z_0.975 <- qnorm(0.975)
```

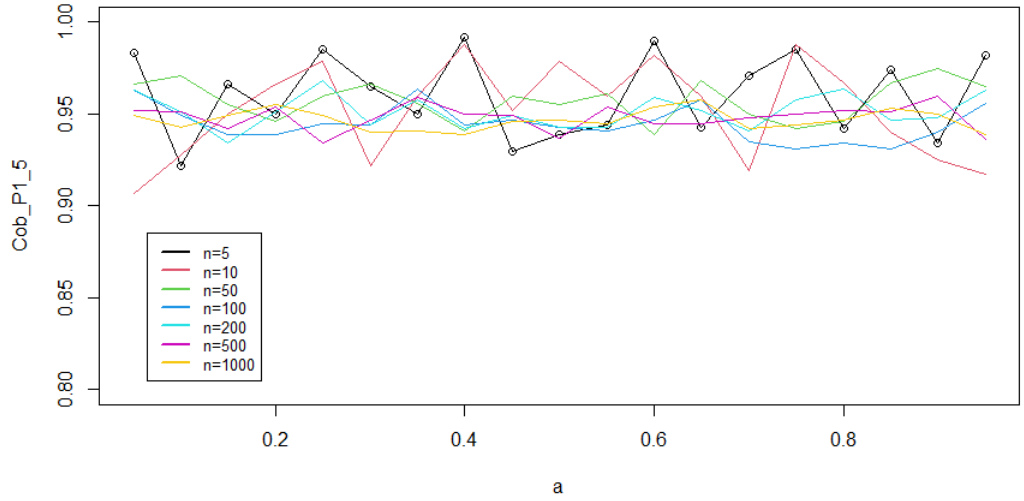
```
5
6  p_n <- function(x){
7    mean(x)
8  }
9  se2 <- function(x){
10   sqrt((p_n(x)*(1-p_n(x)))/n)
11 }
12 error2 <- function(x){
13   z_0.975*(se2(x))
14 }
15
16 l1 <- function(x){
17   ((2*n*p_n(x)+(z_0.975)^2)-(z_0.975)*sqrt(4*n*p_n(x)+(z_0.975)^2-4*n*(p_n(x))
       ^2))/(2*(n+(z_0.975)^2))
18 }
19 u1 <- function(x){
20   ((2*n*p_n(x)+(z_0.975)^2)+(z_0.975)*sqrt(4*n*p_n(x)+(z_0.975)^2-4*n*(p_n(x))
       ^2))/(2*(n+(z_0.975)^2))
21 }
22
23
24 l2 <- function(x){
25   p_n(x)-error2(x)
26 }
27 u2 <- function(x){
28   p_n(x)+error2(x)
29 }
30
31 l3 <- function(x){
32   (sin(asin(sqrt(p_n(x)))-(z_0.975)/(2*sqrt(n))))^2
33 }
34 u3 <- function(x){
35   (sin(asin(sqrt(p_n(x)))+(z_0.975)/(2*sqrt(n))))^2
36 }
37
38 long1 <- function (x){
39   u1(x)-l1(x)
40 }
41 long2 <- function (x){
42   u2(x)-l2(x)
43 }
44 long3 <- function (x){
45   u3(x)-l3(x)
46 }
47
48 for (n in b) {
49   Cobertura1<-paste("Cob","P1",n,sep="_")
50   assign(Cobertura1,numeric())
51   Cobertura2<-paste("Cob","P2",n,sep="_")
52   assign(Cobertura2,numeric())
53   Cobertura3<-paste("Cob","P3",n,sep="_")
54   assign(Cobertura3,numeric())
55
```

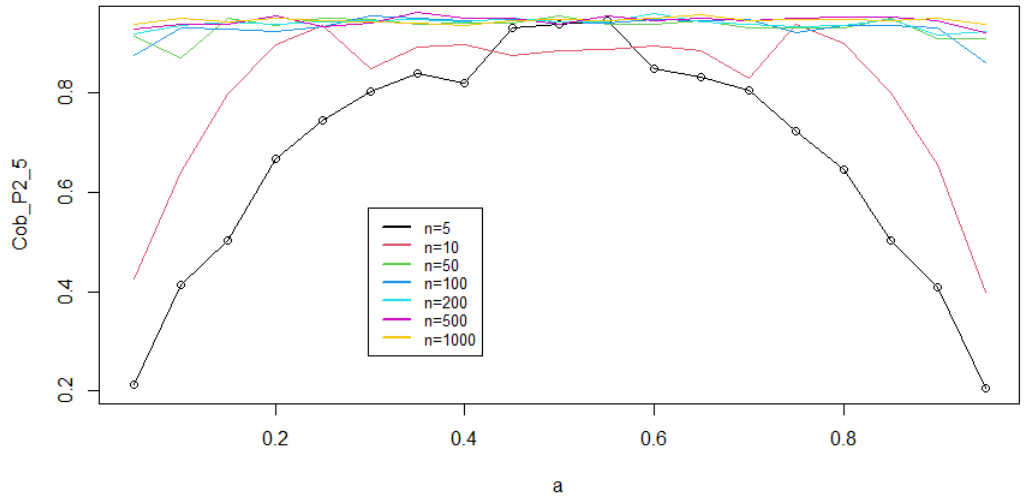
Problema 1 continúa en la siguiente página...

```
  Longitud1<-paste("Long","P1",n,sep="_")
  assign(Longitud1,numeric())
  Longitud2<-paste("Long","P2",n,sep="_")
  assign(Longitud2,numeric())
  Longitud3<-paste("Long","P3",n,sep="_")
  assign(Longitud3,numeric())
}

set.seed(129)
for(p in a){
  for(n in b){

    P1<-paste("P1",n,p,sep="_")
    assign(P1,numeric())

    P2<-paste("P2",n,p,sep="_")
    assign(P2,numeric())

    P3<-paste("P3",n,p,sep="_")
    assign(P3,numeric())


    for(i in 1:1000){

      x<-rbinom(n,1,p)


      x_1 <- c(eval(parse(text=P1)),0)
      x_2 <- c(eval(parse(text = P1)),long1(x))
      y_1 <- c(eval(parse(text=P2)),0)
      y_2 <- c(eval(parse(text = P2)),long2(x))
      z_1 <- c(eval(parse(text=P3)),0)
      z_2 <- c(eval(parse(text=P3)),long3(x))

      if(l1(x)>=p | u1(x)<=p){
        assign(P1,x_1)
      }else{
        assign(P1,x_2)
      }
      if(l2(x)>=p | u2(x)<=p){
        assign(P2,y_1)
      }else{
        assign(P2,y_2)
      }
      if(l3(x)>=p | u3(x)<=p){
        assign(P3,z_1)
      }else{
        assign(P3,z_2)
      }
    }


    c_1 <-(sum(eval(parse(text=P1))!=0)/1000)
```

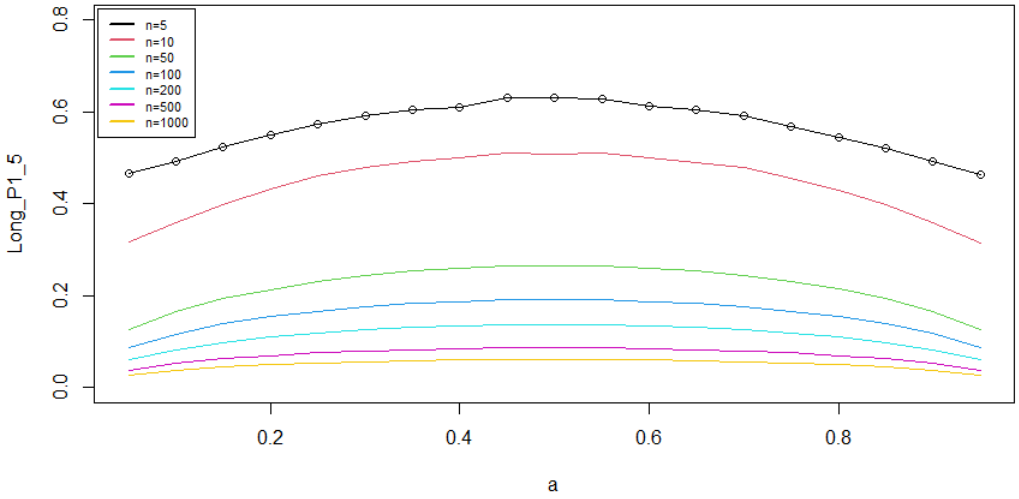
```
    c_2 <-(sum(eval(parse(text=P2))!=0)/1000)
    c_3 <-(sum(eval(parse(text=P3))!=0)/1000)
    c_4 <-c(eval(parse(text=paste("Cob","P1",n,sep="_"))),c_1)
    c_5 <-c(eval(parse(text=paste("Cob","P2",n,sep="_"))),c_2)
    c_6 <-c(eval(parse(text=paste("Cob","P3",n,sep="_"))),c_3)

    l_1 <-sum(eval(parse(text=P1)))/sum(eval(parse(text=P1))!=0)
    l_2 <-sum(eval(parse(text=P2)))/sum(eval(parse(text=P2))!=0)
    l_3 <-sum(eval(parse(text=P3)))/sum(eval(parse(text=P3))!=0)
    l_4 <-c(eval(parse(text=paste("Long","P1",n,sep="_"))),l_1)
    l_5 <-c(eval(parse(text=paste("Long","P2",n,sep="_"))),l_2)
    l_6 <-c(eval(parse(text=paste("Long","P3",n,sep="_"))),l_3)


    CP_I1<-paste("Cob","Prom",P1,sep="_")
    assign(CP_I1,c_1)
    CP_I2<-paste("Cob","Prom",P2,sep="_")
    assign(CP_I2,c_2)
    CP_I3<-paste("Cob","Prom",P3,sep="_")
    assign(CP_I3,c_3)

    LL_I1<-paste("Long","Prom",P1,sep="_")
    assign(LL_I1,l_1)
    LL_I2<-paste("Long","Prom",P2,sep="_")
    assign(LL_I2,l_2)
    LL_I3<-paste("Long","Prom",P3,sep="_")
    assign(LL_I3,l_3)

    assign(paste("Cob","P1",n,sep="_"),c_4)
    assign(paste("Cob","P2",n,sep="_"),c_5)
    assign(paste("Cob","P3",n,sep="_"),c_6)


    assign(paste("Long","P1",n,sep="_"),l_4)
    assign(paste("Long","P2",n,sep="_"),l_5)
    assign(paste("Long","P3",n,sep="_"),l_6)
  }
}
```

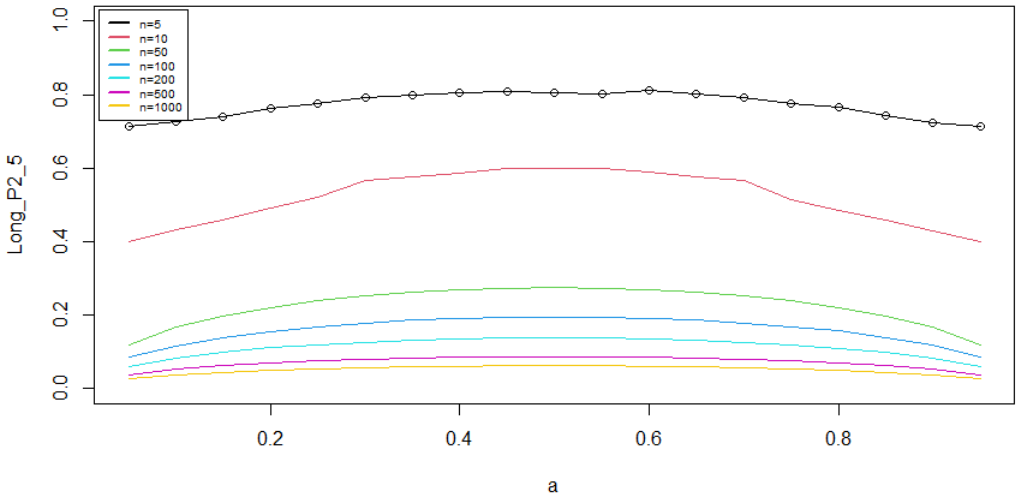
3. Los plots correspondientes las coberturas promedio, para cada tamaño de muestra y cada p, son los siguientes:



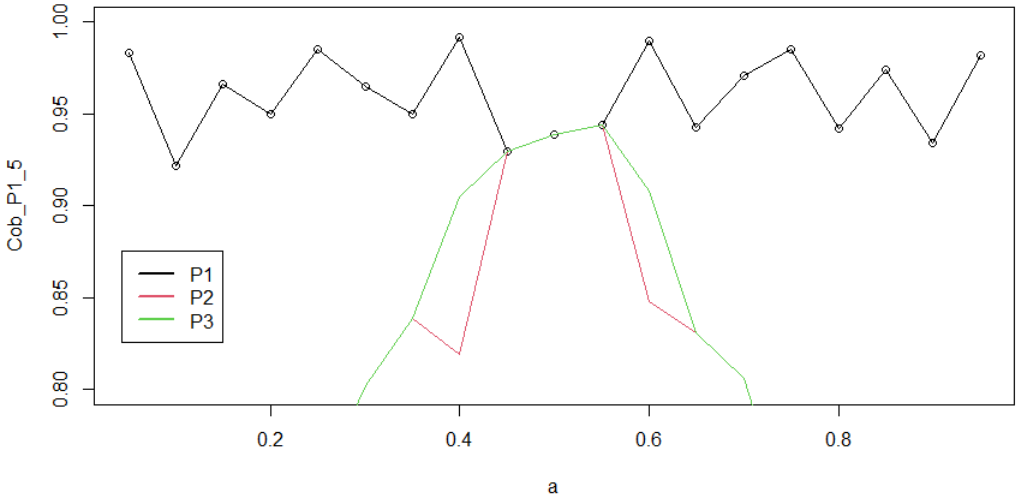


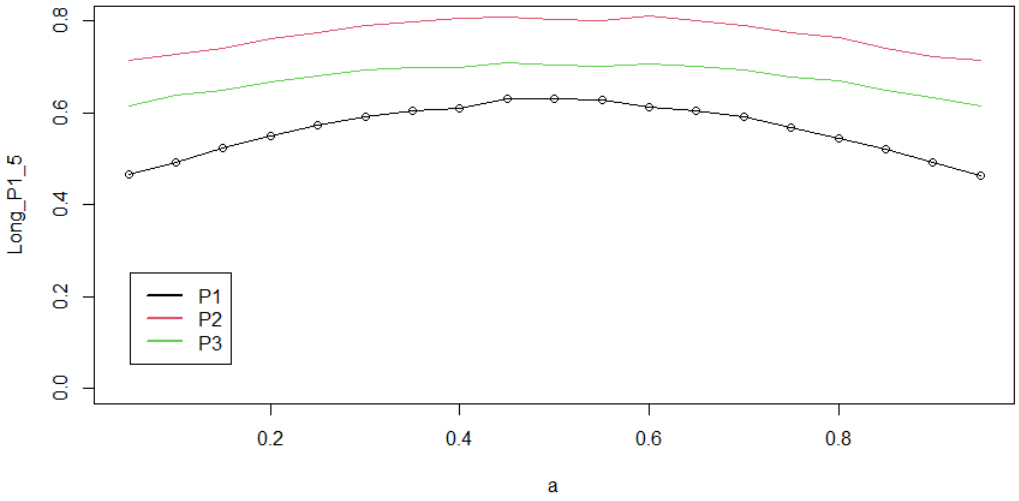
Y los plots correspondientes a las longitudes promedio, para cada tamaño de muestra y cada p, son los siguientes:



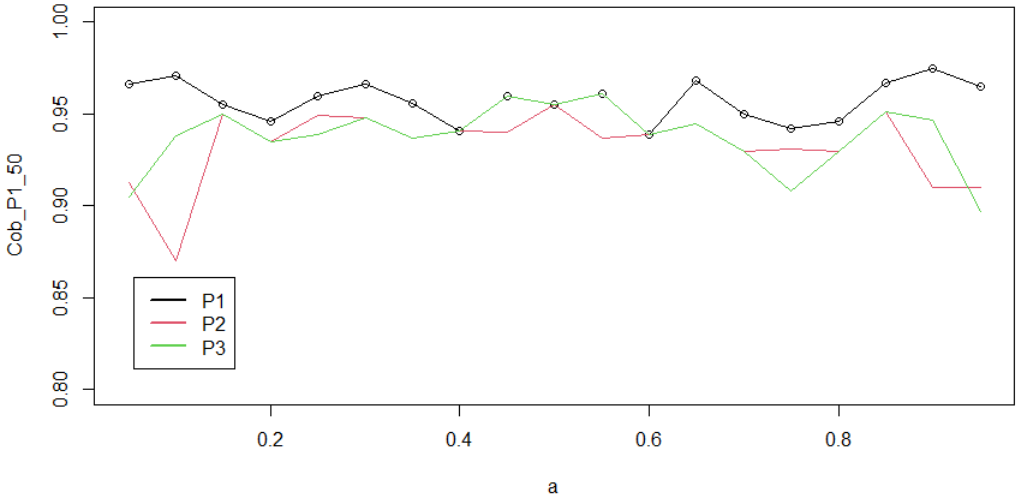


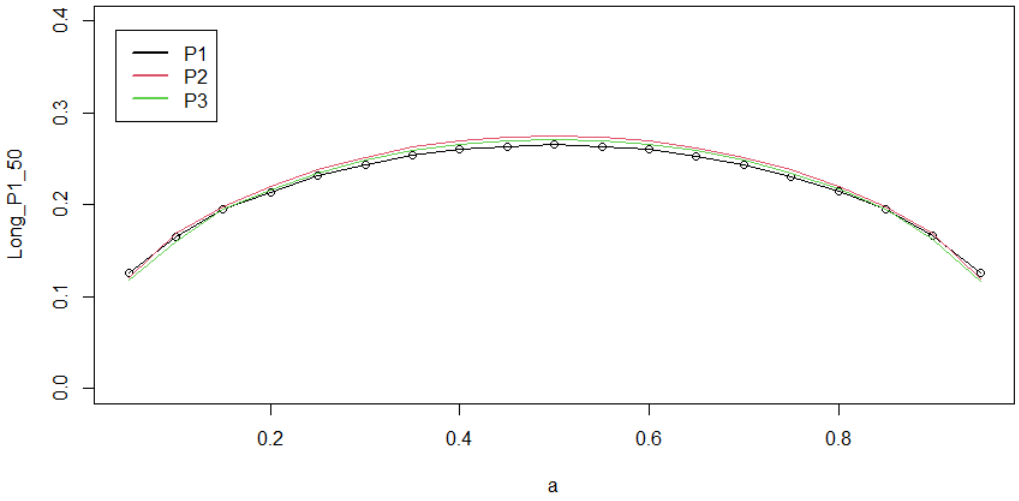
4. Los plots correspondientes a la cobertura y longitud promedio, cuando el tamaño de muestra es $n = 5$, para cada posibilidad, son los siguientes:



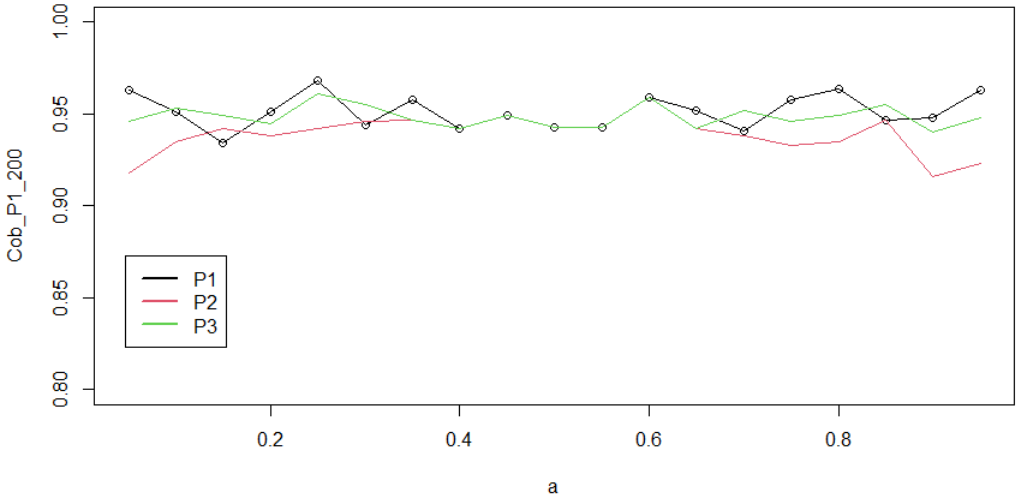


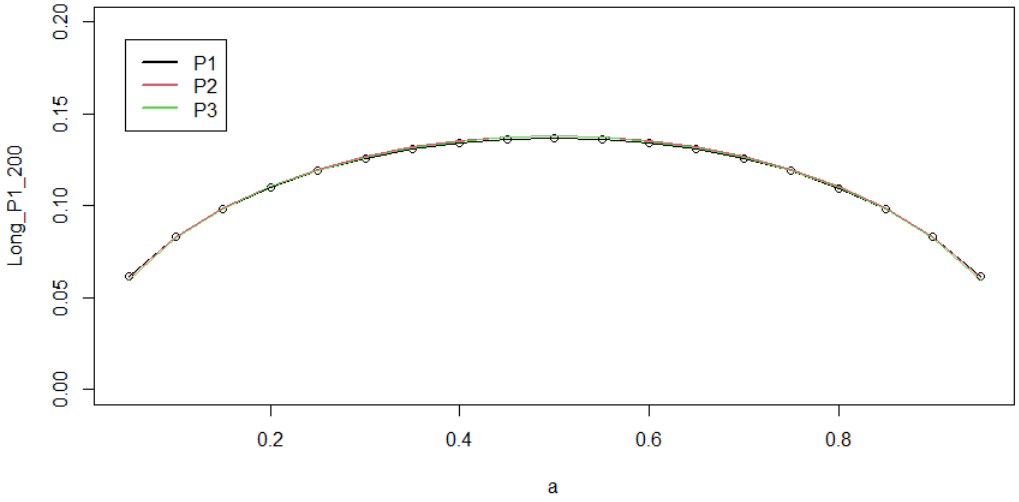
Los plots correspondientes a la cobertura y longitud promedio, cuando el tamaño de muestra es $n = 50$, para cada posibilidad, son los siguientes:



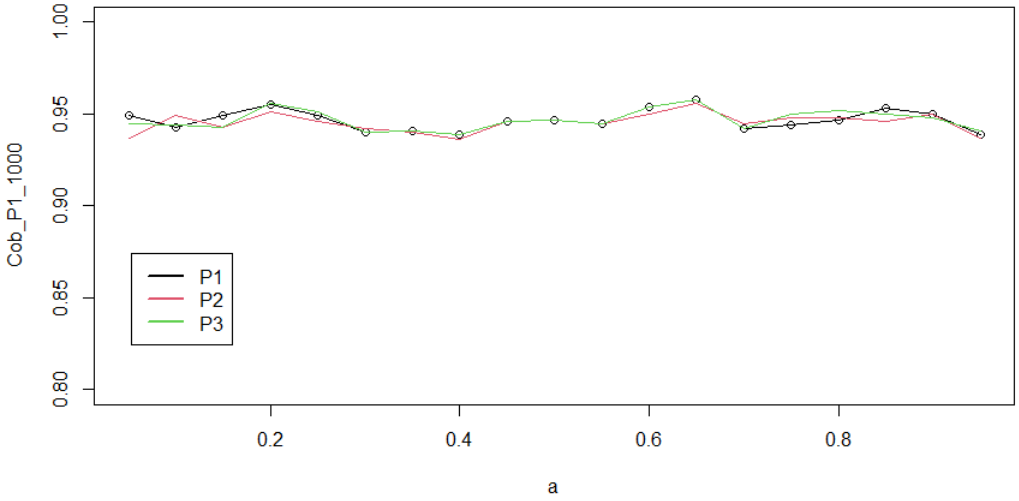


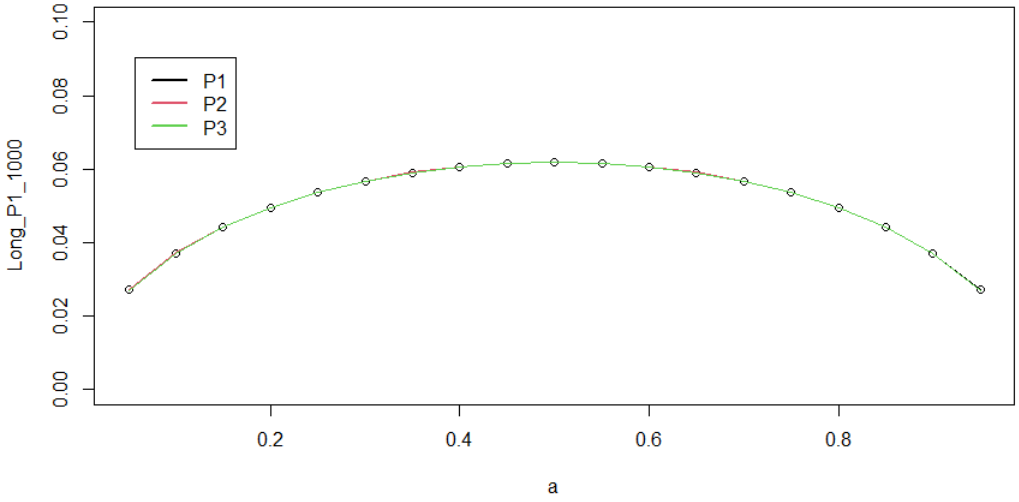
Los plots correspondientes a la cobertura y longitud promedio, cuando el tamaño de muestra es $n = 200$, para cada posibilidad, son los siguientes:





Y por último, los plots correspondientes a la cobertura y longitud promedio, cuando el tamaño de muestra es $n = 1000$, para cada posibilidad, son los siguientes:





5. Podemos concluír que cuando aumentamos el tamaño de muestra, la longitud de los intervalos promedio se reduce mucho, y la cobertura promedio de los intervalos no sólo aumenta sino que además parece estabilizarse.

   Además de esto, cuando el parámetro p está muy cerca de los extremos del intervalo $(0, 1)$, la cobertura y la longitud se reduce. Esto se debe a que, cuando el parámetro estaba cerca de los extremos, en las simulaciones obteníamos muestras con valores también cerca a los extremos,

y esto afectaba en que, al hallar los intervalos de confianza para cada muestra, se pasaran del soporte del parámetro y además, dieran valores muy parecidos pues la probabilidad de éxito era o muy baja o muy alta, por lo que la longitud del intervalo de confianza era menor pues la proporción muestral no tenía datos muy atípicos. Esto repercute en que, al ser intervalos más pequeños, su cobertura fuera un poco menor.

También podemos concluir que la posibilidad 1 trabaja mucho mejor que la posibilidad 2 y 3, esto debido a que, aunque las tres posibilidades se basan en el teorema del límite central y por tanto requieren una muestra grande para funcionar muy bien, la posibilidad 1 es más confiable pues por ejemplo, a diferencia de la posibilidad 2, se basa únicamente en el teorema del límite central, mientras que la posibilidad 2 también se basa en una estimación consistente del parámetro, para que las cuentas no sean tan engorrosas, sacrificando un poco de precisión con muestras más pequeñas, como se puede ver en los plots del punto 3 y 4.

Pero, como las tres posibilidades están basadas en el teorema del límite central, con muestras muy grandes parecen trabajar muy parecido, y ser igual de confiables. Esto se puede notar en los plots del punto 4, con las longitudes y coberturas de las tres posibilidades cuando $n = 1000$.

Problema 2:

B. Análisis inferencial sobre datos reales

A continuación se presentan las rutinas usadas para generar las estimaciones bootstrap por intervalo:

```
#Boostrap_1
#Funcion generadora de estimadores
x_i=NULL
var_1<-function(n){
  for (i in 1:42){
    x_i<-append(x_i,bootstrap[i,n])
  }
  return(var(x_i))
}
#Funcion que almacena estos estimadores
y_i=NULL
for (i in 1:1000){
  y_i=append(y_i,var_1(i))
}
#Boostrap_2
#Funcion generadora de estimadores
x_i=NULL
var_2<-function(n){
  for (i in 1:42){
    x_i<-append(x_i,bootstrap[i,n])
  }
  return(var(x_i)-v_o)
}
#Funcion que almacena estos estimadores
y_i=NULL
for (i in 1:1000){
```

```
  y_i=append(y_i,var_2(i))
}
#En ambos casos para el punto 2
#Limite inferior
quantile(y_i,.0275)
#Limite superior
quantile(y_i,.975)
```

Solución:

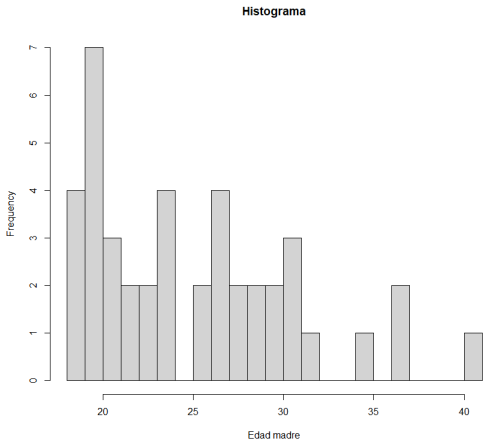
1. Calcule estimaciones puntuales de los siguientes parámetros en la población. No es necesario que use estimadores UMVUE, simplemente justifique qué estimador usaron en cada caso y qué propiedades deseables tiene.

a. ¿Qué porcentaje de nacidos tienen una madre que no supera los 20 años?

Solución:

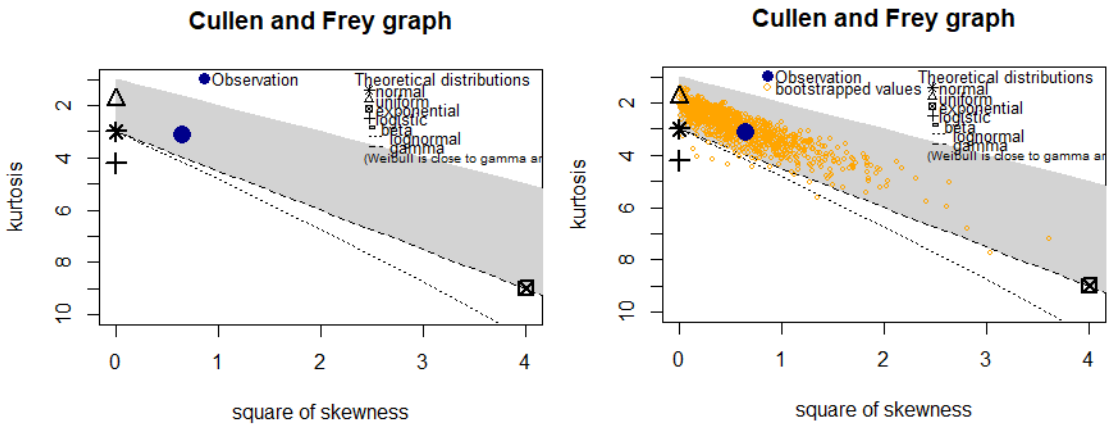
Definimos la variable aleatoria X :'Edad de la madre' y pocederemos a encontrar su distribución. Estableceremos el data set a trabajar y haremos uso de las siguientes librerías:

```
data=read.xlsx("C:/Users/kaido/Downloads/dataset_trabajo1.xlsx",startRow=1,colNames=
    TRUE)
#Ver tabla
View(data)
#Fijar semilla
set.seed(233)
#Librerias
library(openxlsx)
library(tidyverse)
library(MASS)
library(survival)
library(readr)
library(fitdistrplus)
library(datasets)
library(dplyr)
```



Haremos uso del diagrama de Cullen y Frey para intentar determinar cuál es la mejor distribución a ser usada:

```
# Paso (1) Plot Cullen y Frey
descdist( data = data$mage , discrete = FALSE)
descdist(data = data$mage, discrete = FALSE, boot=1000)
```



El uso de la muestra bootstrap ver hacia que distribución parecen inclinarse los datos cuando se hace remuestreo.

Ahora bien, si intentamos ajustar los datos a una distribucion beta como nos indica el diagrama anterior mediante la implementación:

```
fitdist(data$mage,"beta")
max(x)
```

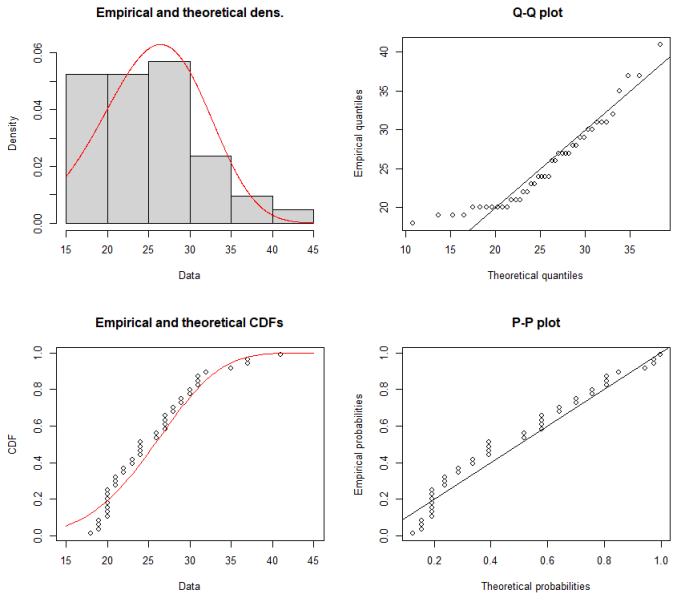
Obtenemos la siguiente salida:

```
> fitdist(data$mage,"beta")
Error in computing default starting values.
Error in manageparam(start.arg = start, fix.arg = fix.arg, obs = data,  :
  Error in start.arg.default(obs, distname) :
  values must be in [0-1] to fit a beta distribution
> max(x)
[1] 41
```

Es decir, esto falla precisamente porque hay datos que no pertencen al soporte de la distribución β. Por otra parte, el diagrama nos indica que los modelos que mejor se ajustan después de β son: normal, Weibull y gamma.

```
normal_ = fitdist(data$mage, "norm")
weibull_ = fitdist(data$mage, "Weibull")
gamma_ = fitdist(data$mage, "gamma")
```

De estos tres el que mejor parece ajustarse es el modelo Weibull



Realizaremos la prueba de bondad de ajuste mediante el test de Kolmogorov- Smirnov y obtendremos los parámetros de la distribución deseada:

```
> prueba<-gofstat(weibull_)
> prueba$kstest
 1-mle-weibull
"not rejected"
> prueba$chisqpvalue
[1] 0.3968121 #p-value> 0.05-> no se rechaza la hipotesis nula
> print(weibull_)
Fitting of the distribution ' weibull ' by maximum likelihood
```

```
Parameters:
      estimate Std. Error
shape  4.65238  0.5202477
scale 27.83948  0.9810004
```

De esta manera,con un nivel de significancia del 95 % aceptamos que nuestro modelo se aproxima a una distribución Weibull , $W(4{,}65, 27{,}84)$. Con esto podemos calcular la estadística de interes.

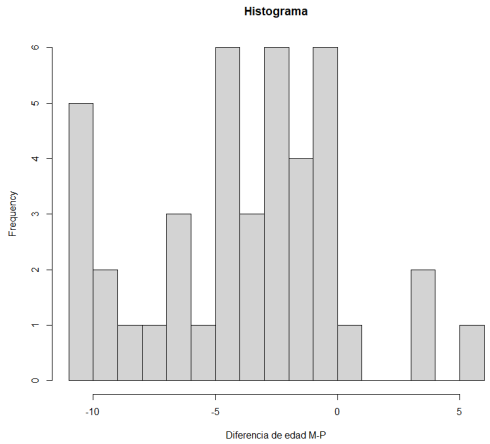
```
> pweibull(20, 4.65, scale = 27.83)
[1] 0.1936042
```

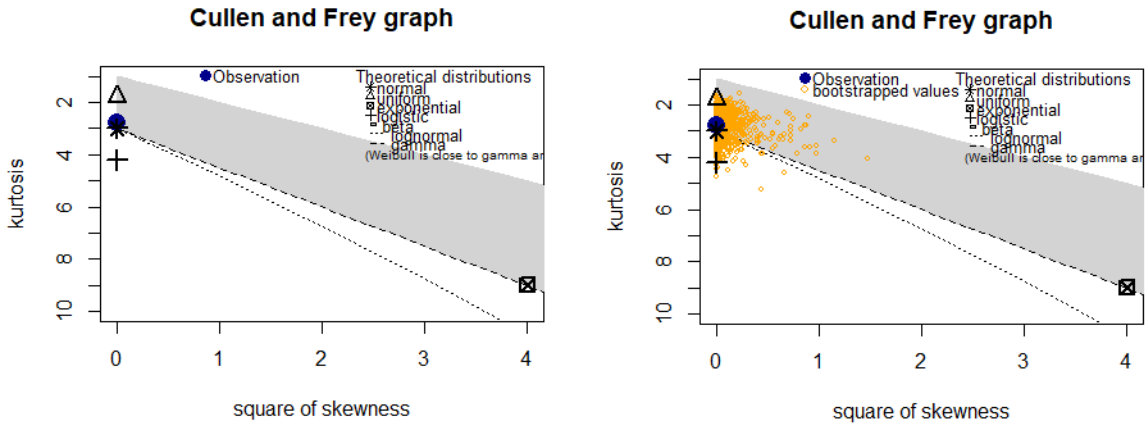
Así, afirmamos que el 19.36 % de nacidos tienen una madre que no supera los 20 años.

b. ¿Qué porcentaje de recién nacidos tienen un padre más joven que la madre?

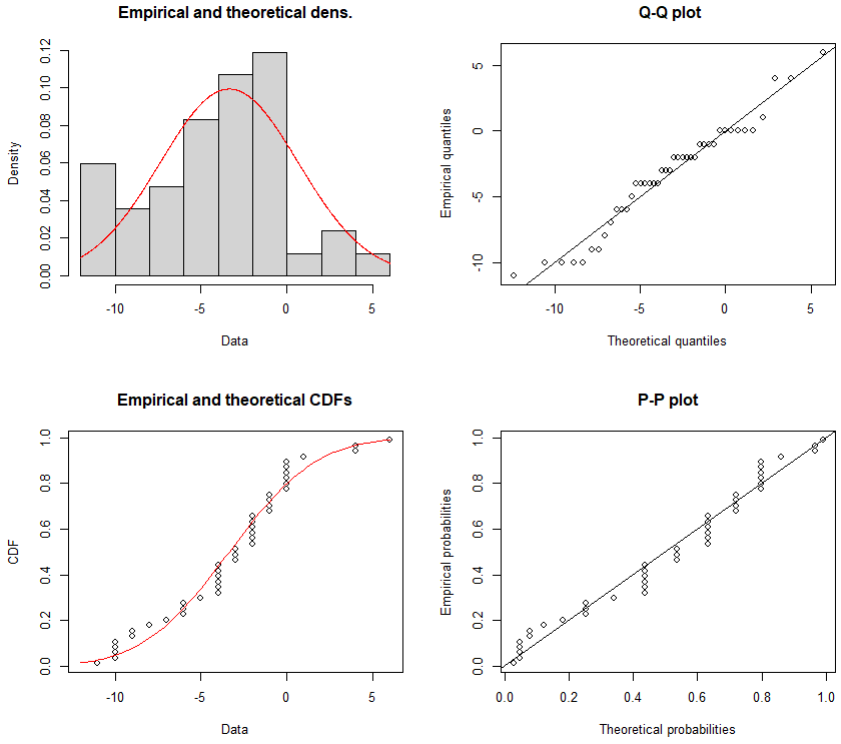
Solución:

Sean X:'Edad de la madre' y Y:'Edad del padre', estamos interesados en estudiar la variable aleatoria $Z = X - Y$. Implementando las rutinas anteriores a esta nueva variable aleatoria obtenemos los siguientes resultados:





Estos diagramas nos indican que la distribución de estos datos parece ser normal. En efecto, la curva parece ajustarse bien como lo indica el siguiente gráfico:



Haciendo las respectivas pruebas de bondad

```
> prueba<-gofstat(normal_)
> prueba$kstest
    1-mle-norm
"not rejected"
> prueba$chisqpvalue
[1] 0.03969815
> print(normal_)
Fitting of the distribution ' norm ' by maximum likelihood
Parameters:
      estimate Std. Error
mean -3.357143  0.6188841
sd    4.010827  0.4376170
```

De esta manera,con un nivel de significancia del 95 % aceptamos que nuestro modelo se aproxima a una distribución Normal , $N(-3{,}35, 4{,}01)$. Con esto podemos calcular la estadística de interes, para esto debemos calcular $P(Z \geq 1)$ pues al ser la edad una cantidad discreta no sería correcto usar $P(Z \geq 0)$.

```
1-pnorm(1,mean=-3.35,sd=4.01)
[1] 0.1390078
```

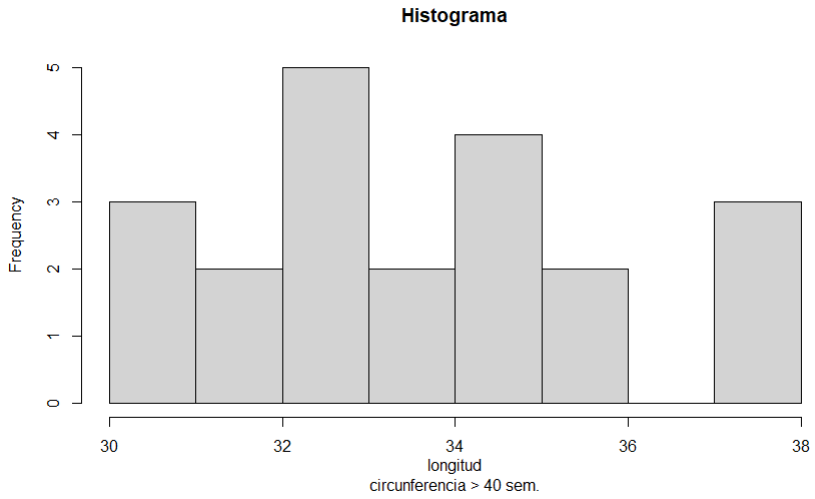
Finalmente, concluimos que el 13,9 % de recién nacidos tienen un padre más joven que la madre.

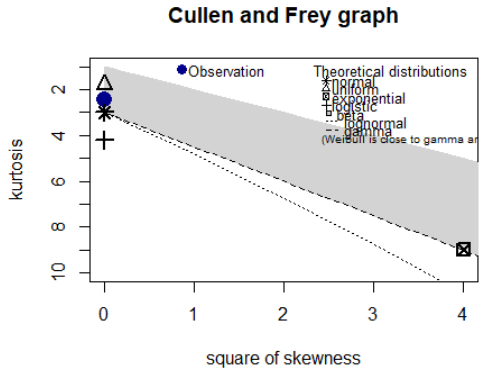
c. ¿Cuál es el promedio y la desviación estándar de la longitud de la circunferencia de los bebés que no superaron las 40 semanas de gestación?

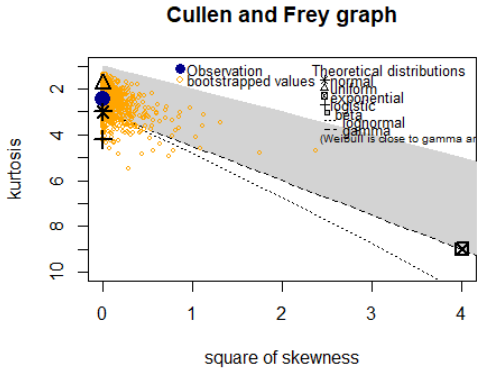
Solución:

Inicialmente pensamos generar la distribución de estos datos a partir de la definición de distribución condicional, sin embargo esto requería muchos supuestos y debiamos realizar una integral doble. Preferimos tomar una submuestra hecha con los datos de los bebés que no superaron las 40 semanas de gestación y repetir el procedimiento el procedimiento de los puntos anteriores. Esta muestra fue generada con:

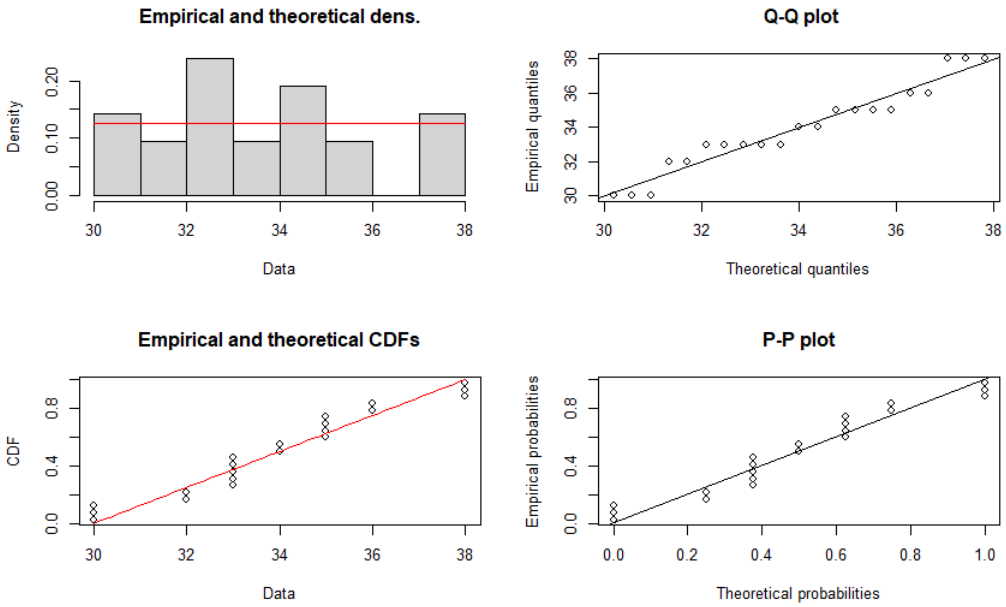
```
submuestra_c<-data %>%filter(Gestation <40)
x=submuestra_c$Headcirc
```







Tenemos evidencia, de que los datos se ajustan a una distribución uniforme. En efecto,



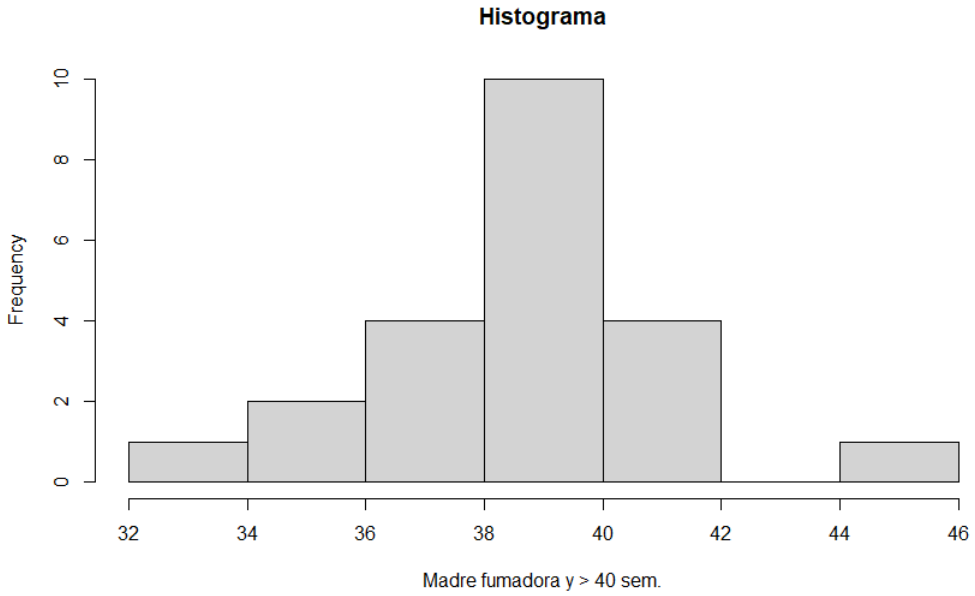
Con los siguientes resultados observamos que falla el calculo del test ks sin embargo para la prueba de chi de Pearson es posible aceptar estos resultados si tenemos una significancia $\alpha = 0{,}93$.

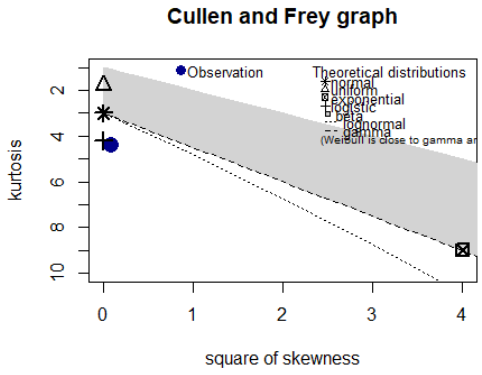
```
> prueba<-gofstat(unif_)
> prueba$kstest
    1-mle-unif
"not computed"
> prueba$chisqpvalue
[1] 0.06854764
> print(unif_)
Fitting of the distribution ' unif ' by maximum likelihood
Parameters:
    estimate Std. Error
min       30         NA
max       38         NA
> 1/2*(30+38)
[1] 34
> 1/12*(38-30)^2
[1] 5.333333
```

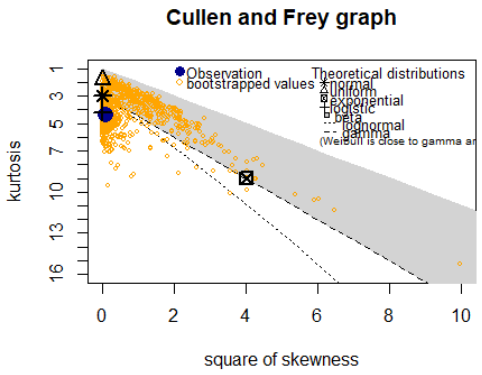
De manera que, las estimaciones de la media y la desviación estándar son 34 y 5.33, respectivamente.

d. ¿Cuál es la proporción de madres fumadoras de los bebés que no superaron las 40 semanas de gestación?

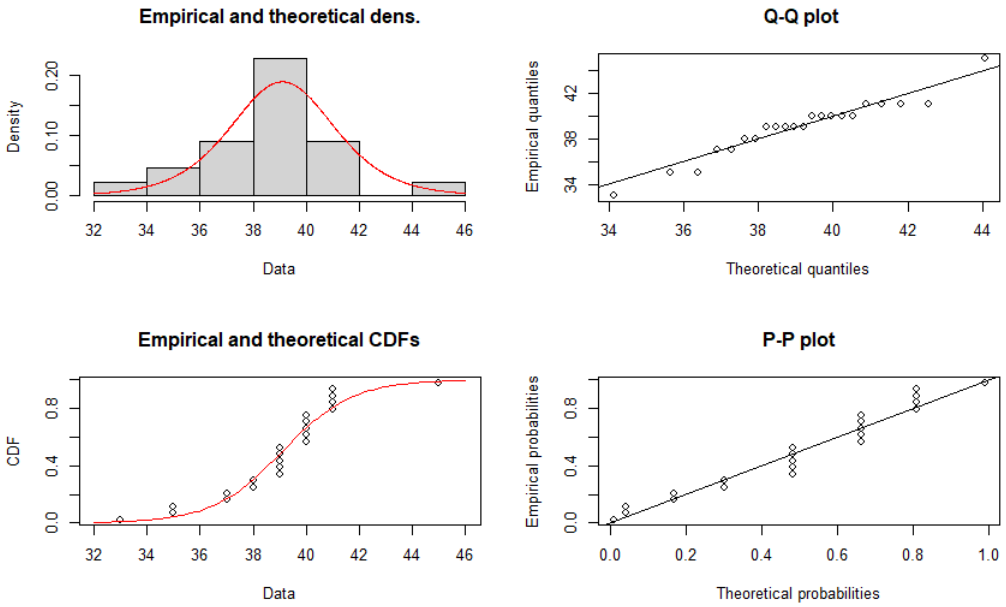
Solución:







Los diagramas nos dan bastante evidencia de que estamos trabajando con una distribución logística.




```
#Pruebas de bondad de ajuste
> logis_=fitdist(x,"logis")
> plot(logis_)
> prueba<-gofstat(logis_)
> prueba$kstest
   1-mle-logis
"not computed"
> prueba$chisqpvalue
[1] 0.23136
> print(logis_)
Fitting of the distribution ' logis ' by maximum likelihood
Parameters:
          estimate Std. Error
location 39.095971  0.4807186
scale     1.320257  0.2411096
> plogis(40,location = 39.09, scale = 1.32)
[1] 0.6658321
```

De nuevo falla la prueba K-S pero la prueba del p-value no. Así,con un nivel de significancia del 95 % aceptamos que nuestro modelo se aproxima a una distribución logística , $Logis(39{,}09, 1{,}32)$.Y determinamos que la proporción de la pregunta es 66.58 %

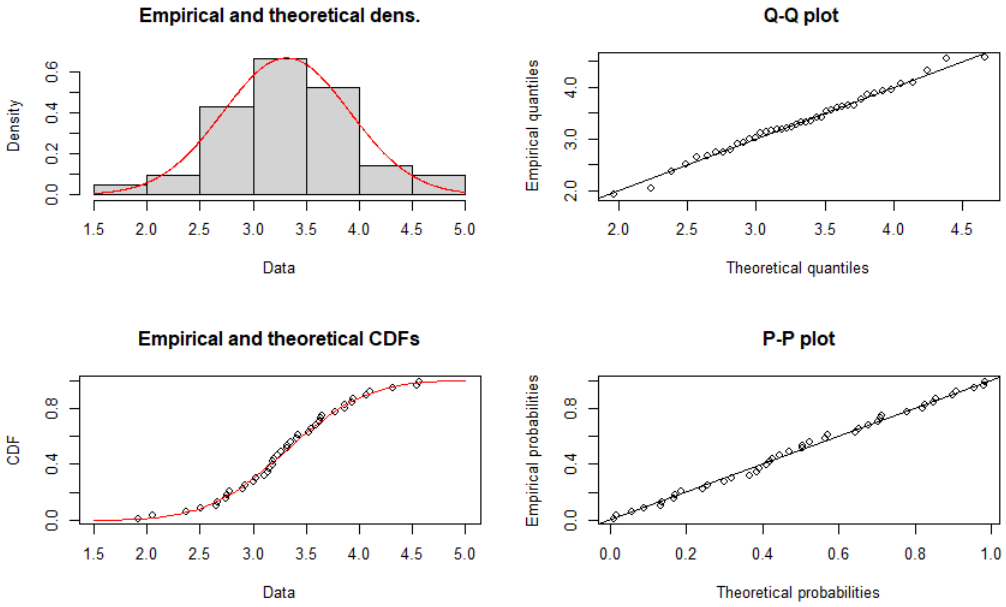
2. Calcule e interprete un intervalo confidencial para los siguientes parámetros con un nivel de confianza del 99 %. Primero, verifique si las suposiciones distribucionales requeridas en cada caso se tienen2. Si no, use ambos métodos de intervalos de confianza Bootstrap y compare los resultados.

a. Peso promedio del bebé.

Solución:
Vamos a verificar que si estos datos vinen de una distribución normal. La siguiente rutina presenta las pruebas de bondad de ajuste y la evidencia gráfica que sustenta la veracidad de esto:

```
> #Peso promedio bebes
> x=data$Birthweight
> #Verificamos si los datos provienen de una dist. normal
> normal_ = fitdist(x, "norm")
> plot(normal_)  #Evidencia grafica
> prueba<-gofstat(normal_)
> prueba$kstest
    1-mle-norm
"not rejected"
> prueba$chisqpvalue
[1] 0.7054872
```

Debido a que trabajamos con un nivel de significancia es claro que con este p-value no se rechaza la hipótesis nula de bondad de ajuste.



Con la siguiente rutina calculamos los límites de el intervalo de confianza requerido:

```
> #Calculo intervalo de confianza
> x_barra=mean(x)
> sd=sqrt(var(x))
> x_barra-qt(0.995,df=41)*sd/sqrt(42) #Limite inferior
[1] 3.061153
> x_barra+qt(0.995,df=41)*sd/sqrt(42) #Limite superior
```

```
[1] 3.564561
```

De esta manera concluimos que:

$$IC_{99\,\%}(\mu) = [3{,}06, 3{,}56]$$

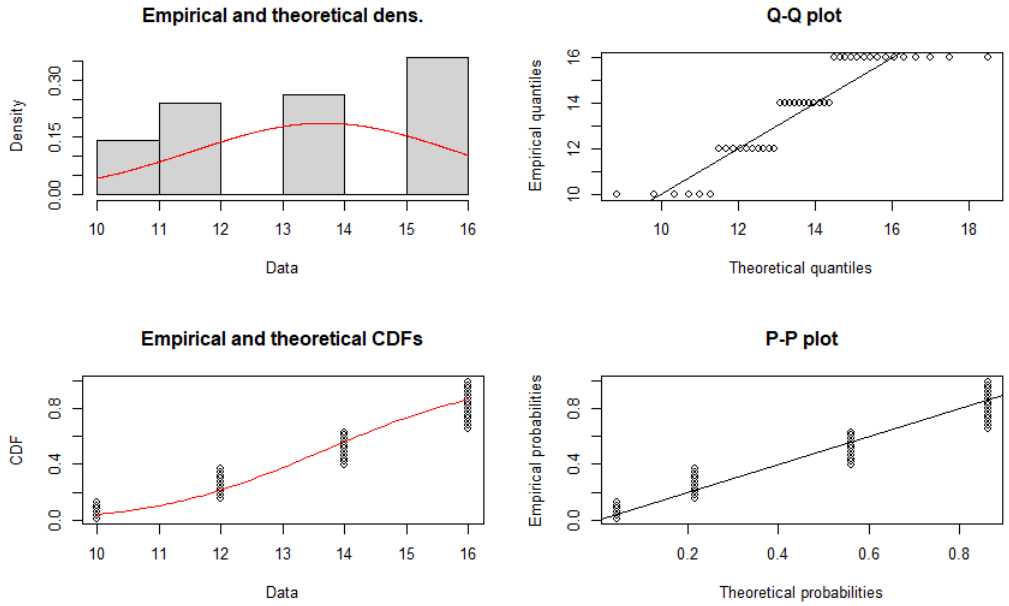
b. Varianza de la cantidad de años de educación del padre.

Solución:

Observemos que los datos no provienen de una distribución normal:

```
> #Verificamos si los datos provienen de una dist. normal
> normal_ = fitdist(x, "norm")
> plot(normal_)  #Evidencia grafica
> prueba<-gofstat(normal_)
> prueba$kstest
1-mle-norm
"rejected"
> prueba$chisqpvalue
[1] 0.0001371704
```

Con este p-value se rechaza la hipótesis nula.



Sin embargo, usando el diagrama de Cullen y Frey tenemos evidencia para sospechar que la distribución uniforme se ajusta a los datos. De esta manera, la siguiente rutina nos permite verificar la prueba de bondad de ajuste y la estimación de dichos parámetros:

```
> #Bondad de ajuste
```

```
> unif_=fitdist(x,"unif")
> plot(unif_)
> prueba<-gofstat(unif_)
> prueba$kstest
1-mle-unif
"rejected"
> prueba$chisqpvalue
[1] 0
```

A pesar de esto obtenemos peores resultados por lo que concluimos que, entre las distribuciones conocidas, la distribución normal es la que ,mejor se ajusta a estos datos. Procedemos a hacer el cálculo de este intervalo de confianza mediante los métodos bootstrap:

```
#Metodo 1
#Limite inferior
> quantile(y_i,.0275)
   2.75%
3.317073
> #Limite superior
> quantile(y_i,.975)
   97.5%
5.865273
#Metodo 2
#Limite inferior
> v_o-quantile(y_i,.975)
  97.5%
3.46806
> #Limite superior
> v_o-quantile(y_i,.0275)
  2.75%
6.01626
```

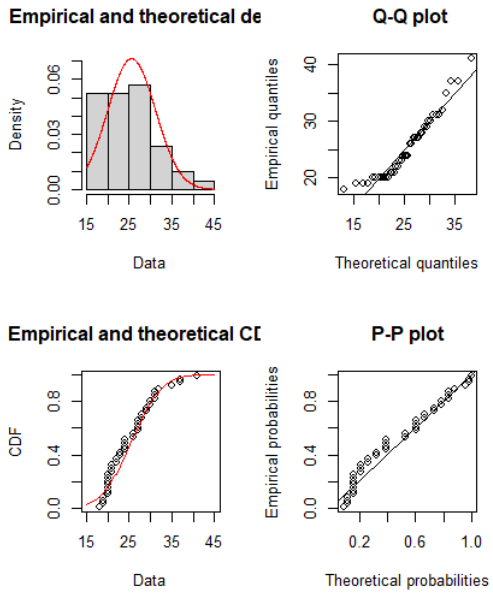
$$IC_{Boot_1}(\sigma^2) = [3{,}31, 5{,}86]$$

$$IC_{Boot_2}(\sigma^2) = [3{,}46, 6{,}01]$$

c. coeficiente de variación de la edad de la madre.

Solución:

Veamos que efectivamente, estos datos provienen de una distribución normal



Para las pruebas de bondad de ajuste obtennemos:

```
> #Verificamos si los datos provienen de una dist. normal
> normal_ = fitdist(x, "norm")
> plot(normal_)  #Evidencia grafica
> prueba<-gofstat(normal_)
> prueba$kstest
    1-mle-norm
"not rejected"
> prueba$chisqpvalue
[1] 0.2328646
```

El p-value calculado nos garantiza que el test no rechaza la hipótesis nula con la significancia dada. Sin embargo, no encontramos una fórmula para este intervalo de confianza, por lo tanto lo estimaremos mediante los métodos bootstrap.

```
#Metodo 1
> #Limite inferior
> quantile(y_i,.0275)
    2.75%
0.1774045
> #Limite superior
> quantile(y_i,.975)
    97.5%
0.2574538
#Metodo 2
> #Limite inferior
> v_o-quantile(y_i,.975)
    97.5%
0.1861368
```

```
> #Limite superior
> v_o-quantile(y_i,.0275)
    2.75%
0.2661862
```

Con esto calculamos los intervalos de confianza dados por:

$$IC_{Boot_1}(c) = [0{,}17, 0{,}25]$$

$$IC_{Boot_2}(c) = [0{,}18, 0{,}26]$$

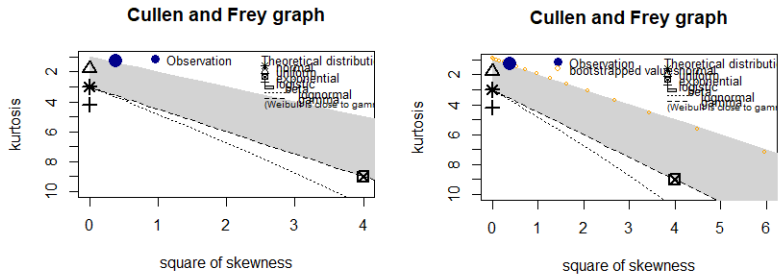
d. proporción de padres fumadores.

Solución:

Dado que el data set no contiene explicitamente construiremos esta columna y la añadiremos al dataframe

```
fnocig<-as.numeric(data$fnocig>0)
data<-cbind(data,fnocig)
x=fnocig
v_0=length(x[x==1])/42 #Estimacion inicial
```

En este caso, ninguna distribución continua se va ajustar. Es interesante ver como esto representa en el diagrama de Callen



Donde el valor observado sale de la frontera de aceptación y las muestras bootstrap parecen converger puntualmente a un punto de acumulación.Como no proviene de una distribución normal pasamos a hacer uso de los métodos bootstrap.

```
#Metodo 1
> #Limite inferior
> quantile(y_i,.0275)
2.75%
  0.5
> #Limite superior
> quantile(y_i,.975)
    97.5%
0.7857143
```

```
#Metodo 2
> #Limite inferior
> v_o-quantile(y_i,.975)
     97.5%
0.07893817
> #Limite superior
> v_o-quantile(y_i,.0275)
    2.75%
0.3646525
```

$$IC_{Boot_1}(c) = [0{,}5, 0{,}78]$$

$$IC_{Boot_2}(c) = [0{,}07, 0{,}36]$$

BONUS [+0.2]: Para las variables de este punto que resultaron no ser normales en la verificación de supuestos, ajuste otra distribución que usted considere y estime los parámetros correspondientes.

Solución: Para visualizar el bono, remitirse a los puntos anteriores; en donde se realizó el ajuste de la distribución y la respectiva estimación de los parámetros.

3. Se considera que el peso de un recién nacido sano oscila alrededor de los 3.4 kg. ¿Considera que los datos muestran evidencia para concluir que, en promedio, los bebés de esta población se encuentran sanos?

Solución:
Queremos determinar si la media del peso de bebés sanos se aproxima a $\hat{p_n}$:

$$\begin{cases} H_0 : \mu = 3{,}4 \\ H_1 : \mu \neq 3{,}4 \end{cases}$$

Para ello, recurrimos a la función t-test de R. Por lo que el código queda de la siguiente manera:

```
############ Punto 3 ##################
######## Importar Base de Datos #########
library(readxl)
data <- read_excel("C:/Users/ander/OneDrive - Universidad Nacional de Colombia/
    Documentos/(2022-01) Cuarto Semestre/Inferencia Estadistica/Proyecto/Proyecto
    parte B/dataset_trabajo1.xlsx")
View(data)
######### Punto 3 #############
t.test(x=data$Birthweight, alternative = 'two.sided', mu = 3.4, conf.level = 0.95)
```

Esto nos arroja como resultado lo siguiente:

```
  One Sample t-test

data:  data$Birthweight
t = -0.93518, df = 41, p-value = 0.3552
alternative hypothesis: true mean is not equal to 3.4
95 percent confidence interval:
 3.124670 3.501044
```

```
sample estimates:
mean of x
 3.312857
```

Por lo tanto, a un nivel de confianza del 95 % podemos concluir que los datos no muestran evidencia suficiente para rechazar Ho. Es decir, que la media del peso de los bebes sanos se aproxima a 3.4 kg, con un nivel de confianza del 95 %. Asimismo, teniendo en cuenta que la media de la población del peso de los bebés sanos se encuentra entre $[3{,}12kg, 3{,}50kg]$ a un 95 % de confianza.

4. ¿Es posible concluir que menos del 40 % de madres gestantes son fumadoras?

Solución:
Queremos determinar si la proporción de madres fumadoras en mayor a 0.4:

$$\begin{cases} H_0 : \hat{\rho_n} \leq 0{,}4 \\ H_1 : \hat{\rho_n} > 0{,}4 \end{cases}$$

Para ello, recurrimos a la función t-test de R. Por lo que el código queda de la siguiente manera:

```
######## Punto 4 ##########
z_critico <- (22/42 - 0.40) / sqrt(0.40 * (1 - 0.40) / 200)
z_critico
```

Esto nos arroja como resultado un valor $z_c = 1{,}58$. Con un $\alpha = 0{,}05$ y la prueba a una cola tenemos un $Z = 1{,}64$, por lo tanto como $z_c < Z$, entonces los datos no evidencian información suficiente para rechazar H_0.

5. ¿Es plausible afirmar que los bebés de madres no fumadoras tienen, en promedio, una longitud mayor a la de los bebés de madres fumadoras?

Solución:
Queremos determinar si la media de la longitud de los bebés de las madres no fumadoras es mayo al de las madres fumadoras. Para ello planteamos la siguiente prueba de hipótesis:

$$\begin{cases} H_0 : \mu_1 - \mu_2 \leq 0 \\ H_1 : \mu_1 - \mu_2 > 0 \end{cases}$$

Teniendo en cuenta que μ_1 representa al promedio de la longitud de los bebés de las madres no fumadoras y μ_2 al promedio de la longitud de los bebés de las madres fumadoras.
Por ello, el test realizado en R, queda de la siguiente manera:

```
####### Punto 5 ###########
x <- data[data$smoker==0, 2]
x
m_1 <- mean(x$Length)
y <- data[data$smoker==1, 2]
y
m_2 <- mean(y$Length)
t.test(x, y, alternative='greater', mu=0, var.equal=TRUE, conf.level=0.95)
```

Al momento de evaluar el test, obtenemos lo siguiente:

```
Two Sample t-test

data:  x and y
t = -0.98185, df = 40, p-value = 0.834
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 -2.418802       Inf
sample estimates:
mean of x mean of y
 50.90909  51.80000
```

En este caso, asumiendo que las varianzas son iguales, obtenemos un $t_c = -2{,}41$, el cual es menor que el $t = -0{,}98$. Por lo tanto, se concluye que los bebés de las madres no fumadoras, en promedio, tienen una longitud mayor que los bebés de las madres fumadoras (Rechazando H_0).

6. ¿Es posible concluir que, en promedio, los padres son mayores en edad a sus correspondientes parejas femeninas?

Dado que la muestras son pareadas trabajemos con el siguiente sistema de hipótesis con cola a izquierda:

Solución:

$$\begin{cases} H_0 : \mu_D \leq 0 \\ H_1 : \mu_D > 0 \end{cases}$$

La siguiente rutina nos permite realizar el test para diferencia media poblacional:

```
#Punto 4
data=read.xlsx("C:/Users/kaido/Downloads/dataset_trabajo1.xlsx",startRow=1,colNames=
    TRUE)
#Ver tabla
View(data)
#Fijar semilla
data
dif_p_m <- as.numeric( data$fage-data$mage)
data <- cbind ( data , dif_p_m )
x = dif_p_m
#Valor critico
c=(mean(x)-0)/(sd(x)/sqrt(42))
c
#Nivel de significancia 95%
qnorm(0.975,mean=0,sd=1)
```

Obtenemos como salida:

```
>#Valor critico
> c=(mean(x)-0)/(sd(x)/sqrt(42))
> c
[1] 5.359544
> #Nivel de significancia 95%
```

```
> qnorm(0.05,mean=0,sd=1)
[1] -1.644854
```

De manera que como, $z_c > z_\alpha$ no rechazamos la hipótesis nula. Esto es, tenemos evidencia para pensar que los padres son mayores en edad a sus parejas femeninas.

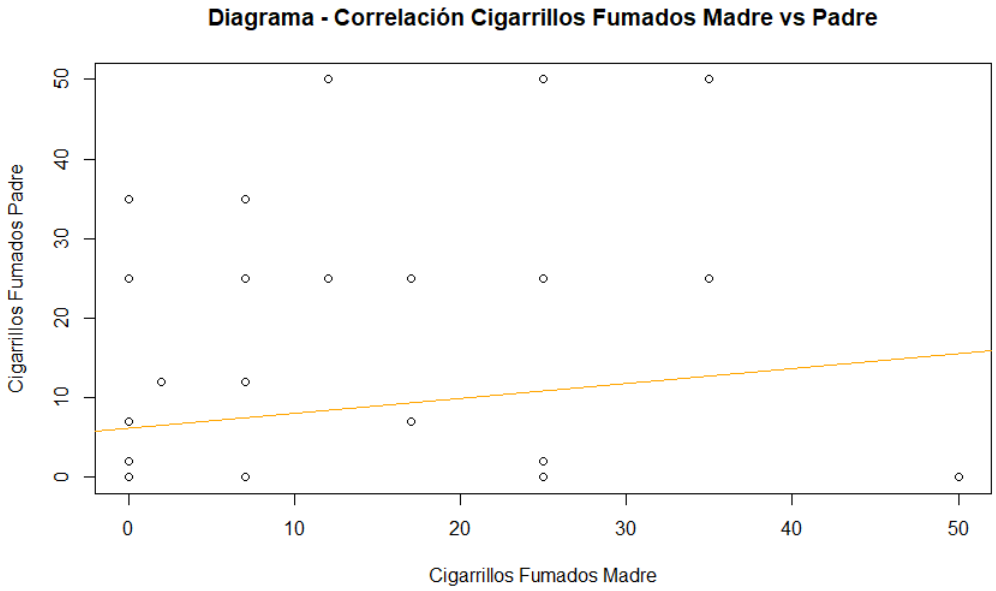
Bonus [+0.5]:

Nota: Para los puntos 7, 8 y 9 se usó como base el siguiente código, con el fin de importar la base de datos e instalar los paquetes necesarios para la resolución del Bonus. El código en cuestión es el siguiente:

```
############## Bonus Parte B ###################
##### Instalacion de Paquetes #######
install.packages("readxl")
install.packages("dplyr")
install.packages("ggplot2")
install.packages("GGally")
install.packages("Hmisc")
install.packages("corrplot")
install.packages("PerformanceAnalytics")
###### Carga de Paquetes ######
library(readxl)
library(dplyr)
library(ggplot2)
library(GGally)
library(Hmisc)
library(corrplot)
library(PerformanceAnalytics)
###### Importar la Base de Datos #######
library(readxl)
datos <- read_excel("C:/Users/ander/OneDrive - Universidad Nacional de Colombia/
    Documentos/(2022-01) Cuarto Semestre/Inferencia Estadistica/Proyecto/Proyecto
    parte B/dataset_trabajo1.xlsx")
View(datos)
attach(datos)
datos
```

Se recomienda, al momento de correr el código, cambiar la ruta de acceso al archivo.
7. ¿Existe una correlación lineal alta entre el consumo de cigarrillo por parte de la madre y del padre? (se considera una correlación alta si supera 0.6). Haga un diagrama de dispersión.

Solución: Teniendo en cuenta las variables a considerar. Tanto el consumo de cigarrillo de las madres, como el condumo de cigarrillo de los padres; se puede evidenciar en el siguiente gráfico de dispersión que no hay una relación lineal entre ambas variables. Al calcular numéricamente la correlación se obtiene como resultado $0,26$. Lo cual confirma que no existe una relación lineal entre las variables mencionadas. Sin embargo, esto no quiere decir que las variables sean independientes; por lo que solo se puede concluir que no existe una relación lineal entre estas variables.

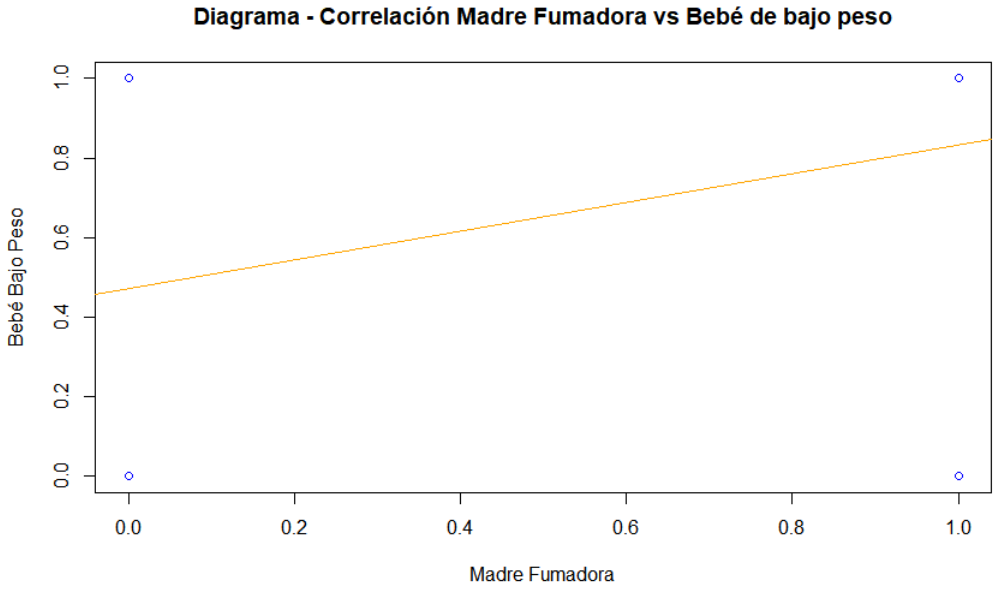


Asimismo, el código para hallar la correlación entre las variables mencionadas es el siguiente:

```
##### Punto 7 #########
correlacion <- round(cor(datos$mnocig,datos$fnocig),2)
correlacion
plot(x = datos$mnocig, y = datos$fnocig, col = "blue",main = "Diagrama - Correlacion
     Cigarrillos Fumados Madre vs Padre",xlab="Cigarrillos Fumados Madre",ylab="
    Cigarrillos Fumados Padre")
abline(lm(datos$mnocig~datos$fnocig),col = "orange")
```

8. ¿Es posible concluir, estadísticamente hablando, que las variables “madre fumadora” y “bebé de bajo peso” guardan cierta relación o son independientes?

Solución: Para responder esta pregunta, hay que tener en cuenta que, dado que las variables en cuestión son variables cualitativas, se debe recurrir al coeficiente de correlación de Spearman. Si queremos ilustrar el diagrama de la correlación, solo nos dará valores de 0 y 1. Tal y como se muestra en la siguiente imagen:



Por ello, para encontrar si las variables "madre fumadoraz Bebé de bajo peso"tienen cierta relación, solo nos debemos basar en el valor dado en el coeficiente de Spearman; el cual da 0,2530. Por lo tanto se conclue que las variables en cuestión no tienen relación alguna. Sin embargo, esto no garantiza la independencia entre las variables.

Nota: Si las variables son independientes, el coeficiente de correlación es igual a 0. Sin embargo, esta afirmación no es recíproca; dado que si el coeficiente da 0, no garantiza que automáticamente las variables sean independientes.

Finalmente, el código para hallar la correlación entre las variables mencionadas es el siguiente:

```
######## Punto 8 ########
correlacion1 <- cor(datos$smoker,datos$lowbwt, method = 'spearman')
correlacion1
plot(x = datos$smoker, y = datos$lowbwt, col = "blue",main = "Diagrama - Correlacion
     Madre Fumadora vs Bebe de bajo peso",xlab="Madre Fumadora",ylab="Bebe Bajo Peso
    ")
abline(lm(datos$smoker~datos$lowbwt),col = "orange")
```

9. ¿Juega la raza un papel en explicar el número de semanas de gestación de las madres embarazadas? Haga un análisis de comparación ANOVA o similar dependiendo del cumplimiento de supuestos.

Solución: A continuación, se muestra la tabla que arroja R al momento de realizar un análisis ANOVA entre las variables Raza y Número de Semanas de Gestación.

	Df	Sum Sq	Mean Sq	F Value	Pr(>F)
datos$Gestation	1	0.04	0.0422	0.038	0.846
Residuals	40	44.43	1.1109		

Tal y como se puede observar, suponiendo un P-Valor de 0,05 se puede interpretar que no existen diferencias estadísticamente significativas entre la raza y las semanas de gestación. Por lo tanto, la raza juega un papel importante en las semanas de gestación de las madres embarazadas.

Finalmente, el código para realizar el análisis ANOVA es el siguiente:

```
####### Punto 9 #########
anova <- aov(datos$race ~ datos$Gestation)
summary(anova)
```