Laboratorio N.3

Introduccion a Los Metodos Estadisticos
Estimacion por intervalos y simulacion

Diana Carolina Arias Sinisterra Cod. 1528008
Kevin Steven Garcia Chica Cod. 1533173
Cesar Andres Saavedra Vanegas Cod. 1628466

Universidad Del Valle
Facultad De Ingenieria
Estadistica
Octubre
2017

Índice

1. Situación 1

Se generaron 5000 muestras aleatorias de tamaño n=10, de una poblacion normal con parametros $\mu = 5$ y $\sigma = 1$ y para cada una de las muestras, se encontro un intervalo de confianza para la media y la varianza respectivamente, con una confianza del 95 %.
Para la media:
Nos arrojo que el porcentaje de intervalos que atrapan la verdadera media es: 95.22 %
Y la longitud promedio o esperada de cada intervalo fue hallada como: $\sum_{i=1}^{5000} \frac{(LS_i - LI_i)}{5000}$ lo que nos arrojo un resultado de 1.2395.

El porcentaje de intervalos que atrapan la verdadera media tiene mucho sentido, ya que, al trabajar con una confianza del 95 %, estamos diciendo que el 95 % de las veces que se repita el experimento (en este caso que se obtenga una muestra distinta de la misma poblacion), el verdadero valor de la media caera en el intervalo obtenido. Como obtuvimos un porcentaje del 95.22 %, nos indica que el 95.22 % de las veces que se repitio el experimento (remuestrear y obtener el IC para la media), la verdadera media cayo en el intervalo encontrado.

Con respecto a la longitud promedio, esta nos arroja la amplitud que tiene aproximadamente cada intervalo. En este caso, todos los intervalos tienen exactamente la misma longitud, ya que utilizamos la estimacion cuando σ es conocida, y en este tipo de estimacion, la parte que se le suma y se le resta a la media muestral para hallar el intervalo de confianza, no depende de los datos. esa parte es $\pm Z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$. Entonces,en este caso, el 1.2395 que nos arrojo la longitud promedio,nos dice que cada uno de los 5000 intervalos tiene una longitud de 1.2395.

Para la varianza:

1.1. Punto a.

PARA LA MEDIA:
Utilizamos la siguiente tabla para ver mas facilmente la comparacion por tamaños de muestra, de la proporcion de intervalos que atrapan la verdadera media y de la longitud promedio de los intervalos.

	n=10	n=30	n=50	n=100
Porcentaje de cubrimiento	0.9522	0.9474	0.9508	0.9464
Longitud promedio del intervalo	1.2395	0.7156	0.5543	0.3919

La siguiente figura nos muestra los 100 primeros intervalos con cada tamaño de muestra (10,20,50,100),

ya que si elaborabamos la grafica con los 5000 intervalos estimados, no nos daba una vision clara ni comparable de lo que ocurre.

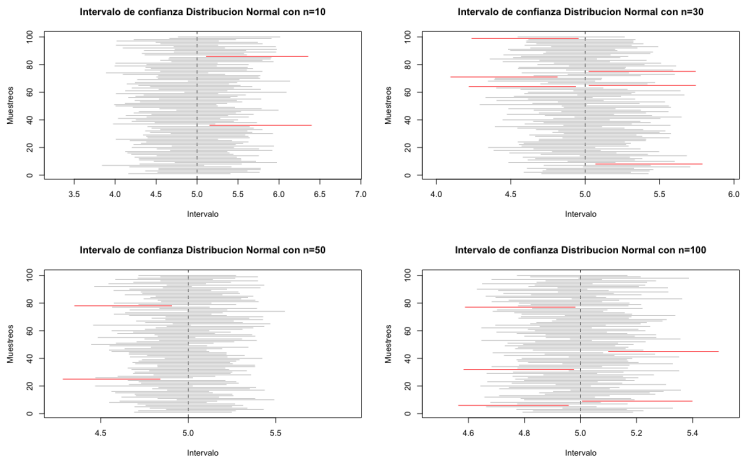

Figura 1: Grafica comparativa de los 100 primeros intervalos por cada tamaño de muestra

Representacion grafica del porcentaje de cubrimiento esperado de los intervalos de la media:

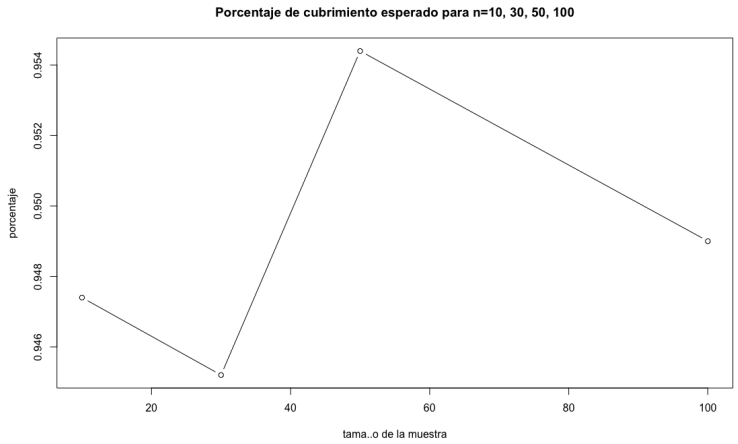

Figura 2: Grafica del porcentaje de cubrimiento de los intervalos para μ por cada tamaño de muestra

En esta imagen podemos ver que todos los porcentajes de cubrimiento estan cerca del 95 %, lo cual es logico, ya que estamos trabajando con una confianza del 95 %, lo que nos dice que el 95 % de las veces que se repita el proceso (remuestrear y hallar el intervalo de confianza) la verdadera media o la media real, va a caer dentro del intervalo estimado. Ademas podemos ver que el porcentaje de cubrimiento disminuye o aumenta sin depender de el tamaño de muestra n(10,30,50,100), es

decir, no se ve un patron claro de dependencia entre el tamaño de las muestras y el porcentaje de cubrimiento de los intervalos estimados.

Representacion grafica de la longitud esperada por intervalo para la media:

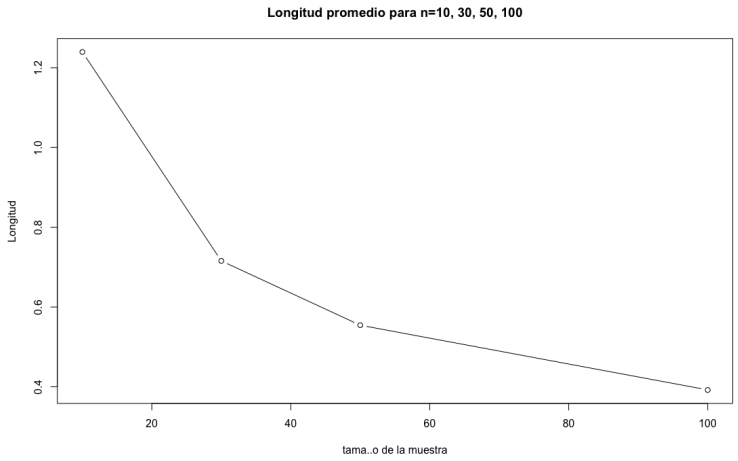

Figura 3: Grafica de la longitud esperada de cada intervalo para μ por cada tamaño de muestra

En la anterior imagen observamos que la longitud mas alta es la de n=10, y no sobrepasa de 1.5, y la mas baja es la de n=100 y es de aproximadamente 0.4. Esto nos muestra que hay una correlacion lineal negativa entre el tamaño de muestra y la longitud del intervalo, es decir, mientras mayor es el tamaño de muestra n, menor es la longitud del intervalo. Lo anterior tiene mucho sentido, ya que la parte que se le suma y se le resta a la media de cada muestra, para hallar una estimacion por intervalos para la media real μ es, $Z_{(1-\alpha)} \cdot \frac{\sigma}{\sqrt{n}}$ y vemos que esta expresion es mas pequeña cuando el n aumenta, haciendo menor, la amplitud de cada intervalo.

PARA LA VARIANZA:

Utilizamos la siguiente tabla para ver mas facilmente la comparacion por tamaños de muestra, de la proporcion de intervalos que atrapan la verdadera varianza y de la longitud promedio de los intervalos.

	n=10	n=30	n=50	n=100
Porcentaje de cubrimiento	0.9524	0.9454	0.9496	0.9502
Longitud promedio del intervalo	2.8446	1.1712	0.8543	0.5795

Representacion grafica del porcentaje de cubrimiento esperado de los intervalos de la varianza:

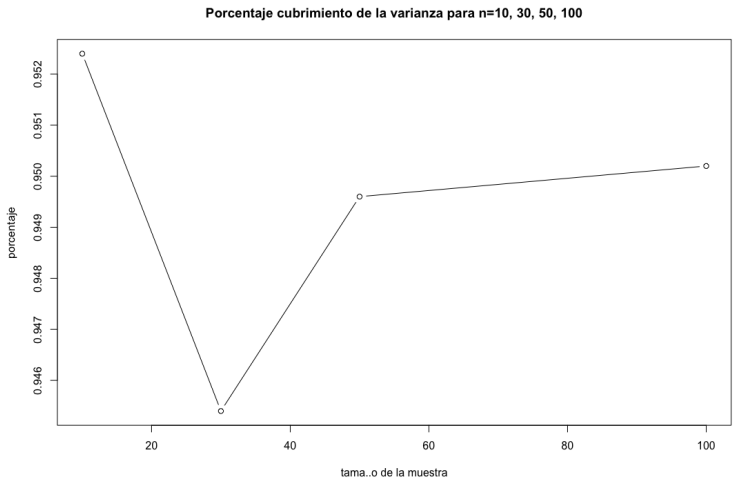

Figura 4: Grafica del porcentaje de cubrimiento de los intervalos para σ^2 por cada tamaño de muestra

En esta imagen podemos ver que todos los porcentajes de cubrimiento de los intervalos para la varianza estan cerca del 95 %, lo cual es logico, ya que como se explico en el punto anterior, estamos trabajando con una confianza del 95 %, lo que nos dice que el 95 % de las veces que se repita el proceso (remuestrear y hallar el intervalo de confianza) la verdadera varianza o la varianza real, va a caer dentro del intervalo estimado. Ademas podemos ver que el porcentaje de cubrimiento disminuye o aumenta sin depender de el tamaño de muestra n(10,30,50,100), es decir, no se ve un patron claro de dependencia entre el tamaño de las muestras y el porcentaje de cubrimiento de los intervalos estimados para la varianza.

Representacion grafica de la longitud esperada por intervalo para la varianza:

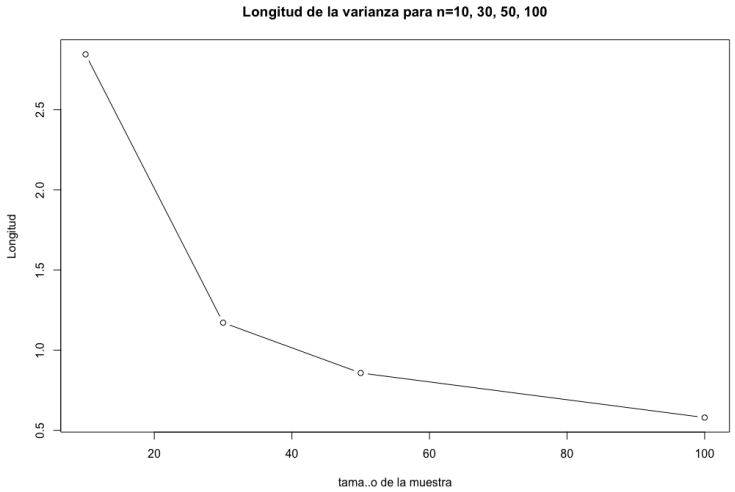

Figura 5: Grafica de la longitud esperada de cada intervalo para σ^2 por cada tamaño de muestra

En esta imagen, al igual que en la de los intervalos para la media. Vemos que la longitud es mayor para los tamaños de muestra de n=10 y es menor para los tamaños de muestra de n=100, lo que nos muestra que entre estas dos variables (tamaño de muestra y longitud del intervalo) existe una correlacion lineal negativa, es decir, cuando aumenta el tamaño de muestra, disminuye la longitud de los intervalos estimados.

1.2. Punto b.

Para este punto se simulo una poblacion exponencial con parametro $\lambda = \frac{1}{5}$ y de alli, se extrayeron las respectivas muestras de los distintos tamaños.
PARA LA MEDIA:
Representacion grafica del porcentaje de cubrimiento esperado de los intervalos para la media de una poblacion exponencial:

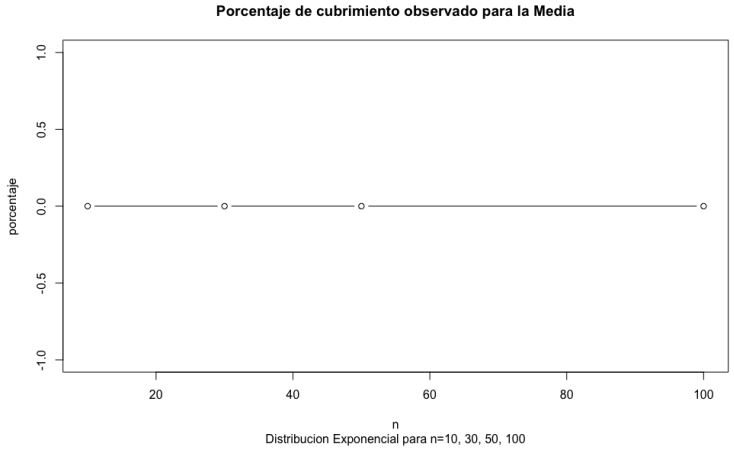

Figura 6: Grafica del porcentaje de cubrimiento de los intervalos para μ por cada tamaño de muestra, para la población exponencial

En esta imagen podemos ver que todos los porcentajes de cubrimiento son del 0 %, al contrario que en la poblacion normal, en la cual todos los intervalos estan cerca del 95 %. Esto se debe a que la estimacion por intervalos que estamos aplicando, solo se utiliza cuando la poblacion es normal. Es decir, estas formulas no sirven para estimar μ de una poblacion no normal. Como se ve en la imagen, nos arroja intervalos que no tienen ningun sentido con la poblacion que estamos trabajando (en este caso, la exponencial) y que no atrapan el verdadero valor de μ.

PARA LA VARIANZA:

Representacion grafica del porcentaje de cubrimiento esperado de los intervalos para la varianza de una poblacion exponencial:

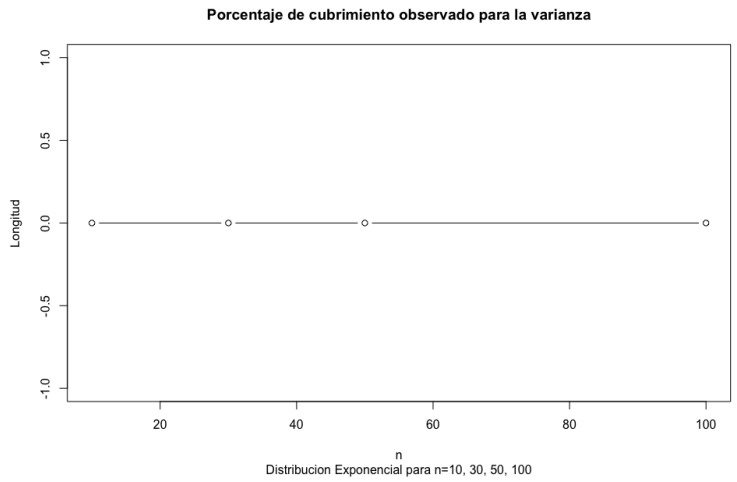

Figura 7: Grafica del porcentaje de cubrimiento de los intervalos para σ^2 por cada tamaño de muestra, para la población exponencial

En esta imagen podemos ver que todos los porcentajes de cubrimiento son del 0 %, al contrario que en la poblacion normal, en la cual todos los intervalos estan cerca del 95 %. Esto se debe a que la estimacion por intervalos que estamos aplicando, solo se utiliza cuando la poblacion es normal. Es decir, estas formulas no sirven para estimar σ^2 de una poblacion no normal. Como se ve en la imagen, nos arroja intervalos que no tienen ningun sentido con la poblacion que estamos trabajando (en este caso, la exponencial) y que no atrapan el verdadero valor de σ^2.

2. Situación 2

Para darnos cuenta que en realidad la mayoria de personas aprueba el proyecto de fluoracion del agua, debemos encontrar un intervalo de confianza de la proporcion de personas que estan a favor, y ver si este esta por encima del 0.5.
Entonces. Tomando los datos del enunciado tenemos:
$n = 200$, $\hat{P} = \frac{110}{200} = 0.55$ (Proporcicion de personas a favor), $\alpha = 0.01$ entonces $1 - \alpha = 0.99$
Ahora, para encontrar un intervalo de confianza para la proporcion, aplicamos la siguiente formula:

$$IC(P)_{(1-\alpha)\,\%} = \left[\hat{P} - Z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{P}(1-\hat{P}}{n}}; \hat{P} + Z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{P}(1-\hat{P}}{n}}\right]$$

Reemplazando en la formula, tenemos:

$$IC(P)_{99\,\%} = \left[0.55 - Z_{0.995}\sqrt{\frac{0.55(0.45)}{200}}; 0.55 + Z_{0.995}\sqrt{\frac{0.55(0.45)}{200}}\right]$$

$$IC(P)_{99\,\%} = [0.55 - 2.5758(0.035178); 0.55 + 2.5758(0.035178)]$$

$$IC(P)_{99\,\%} = [0.4594; 0.6406]$$

INTERPRETACION:

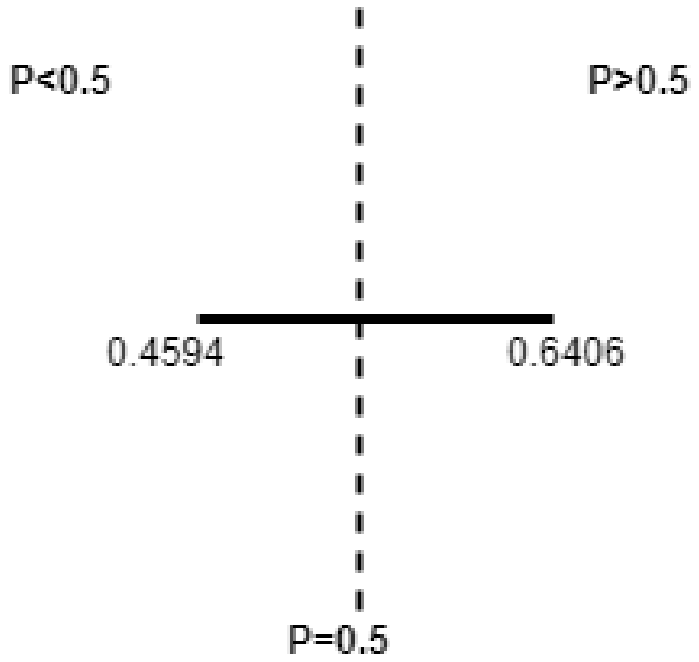

Figura 8: Interpretacion intervalo para proporciones

Como asignamos una confianza del 99 % a nuestro intervalo, decimos que el 99 % de las veces que se

repita el experimento, la proporcion real de personas que estan a favor de que se agregue fluoruro de sodio al agua va a caer en dicho intervalo (entre 0.4594 y 0.6406). Ahora, observando la grafica, podemos ver que en el intervalo esta contenida la probabilidad de 0.5, por lo tanto, la muestra no nos da evidencia para decir que la mayoria de personas aprueba el proyecto de fluoracion.

3. Situación 3

3.1. Punto a.

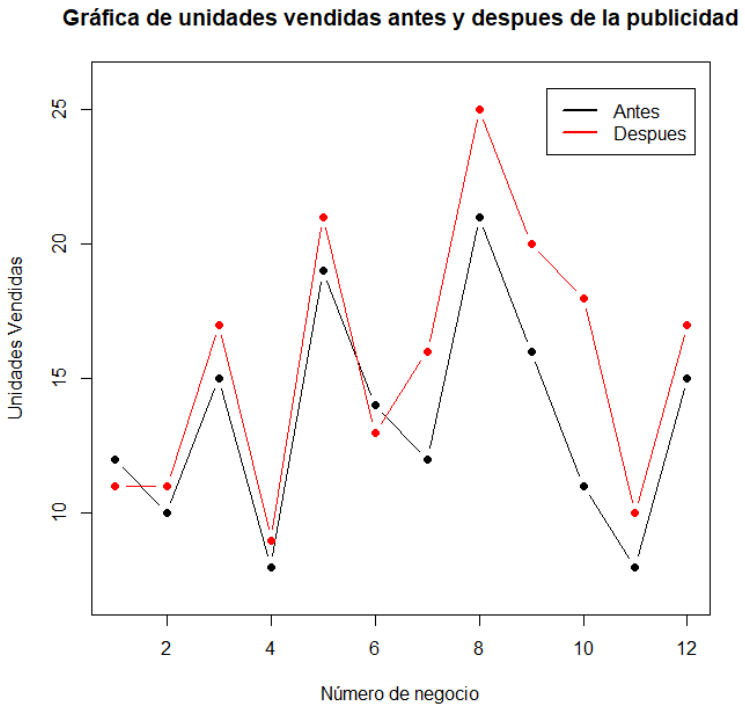

Figura 9: Grafico de puntos y lineas por negocio de unidades vendidas, antes y despues de la publicidad

En el analisis exploratorio de datos que se elaboro, encontramos que la mejor grafica para mostrar la efectividad de la campana publicitaria fue la grafica de puntos y lineas. El eje x son cada una de las sucursales o negocios, y el eje y son las unidades vendidas. La linea negra representa las unidades que se vendian en dichos negocios antes de la campana publicitaria, y la roja representa las unidades vendidas despues de la campana.
Podemos observar que en solo dos de los negocios la linea negra esta por encima de la blanca, es decir, se vendieron mas articulos antes de la campana que despues de ella; en el resto de negocios (10 negocios)la linea roja esta por encima de la negra, es decir,se vendieron mas articulos despues de la campana que antes de ella. Ademas, si miramos detalladamente las diferencias en los negocios en los cuales la linea negra esta por encima de la roja son de tan solo una unidad vendida, en cambio en los negocios en los cuales la linea roja esta por encima de la negra, hay diferencia hasta de 7 unidades vendidas.
Concluimos entonces, que con el analisis exploratorio y con la grafica obtenida, la campana publicitaria si es efectiva.

3.2. Punto b.

Como las muestras estan relacionadas, ya que son tomadas antes y despues de un tratamiento (en este caso la campana publicitaria) a la misma poblacion. Entonces para encontrar este intervalo debemos usar la estimacion para la diferencia de medias para muestras relacionadas.
La formula que tenemos para este tipo de estimacion es:

$IC(\mu_D)_{(1-\alpha)\,\%} = \bar{d} \pm t_{(1-\frac{\alpha}{2};n-1)} \cdot \frac{Sd}{\sqrt{n}}$

En este tipo de intervalo de confianza, todo se basa en la diferencia entre cada uno de los valores del antes y despes del tratamiento, por tanto, para mayor comodida encontramos las diferencias y las añadimos a la tabla:

ANTES	12	10	15	8	19	14	12	21	16	11	8	15
DESPUES	11	11	17	9	21	13	16	25	20	18	10	17
DIFERENCIA	1	-1	-2	-1	-2	1	-4	-4	-4	-7	-2	-2

Encontramos que: $\bar{d} = \frac{\sum_{i=1}^{12} d_i}{12} = \frac{-27}{12} = -2.25$ y $Sd = \sqrt{\frac{\sum_{i=1}^{12}(d_i - \bar{d})^2}{11}} = 2.2613$

Ahora, reemplazando en la formula, nos queda:
$IC(\mu_D)_{95\,\%} = [-2.25 \pm t_{(0.975;11)} \cdot \frac{2.2613}{\sqrt{12}}]$
$IC(\mu_D)_{95\,\%} = [-2.25 \pm 2.2009 \cdot \frac{2.2613}{\sqrt{12}}]$
$IC(\mu_D)_{95\,\%} = [-3.6880; -0.8119]$

En conclusion, un intervalo de confianza para la diferencia de medias de unidades vendidas durante un mes antes y un mes despues de la campaña es (-3.6880 ; -0.8119).

3.3. Punto c.

Para interpretar el intervalo obtenido y darnos cuenta si la campaña publicitaria es efectiva o no, realizamos la siguiente imagen:

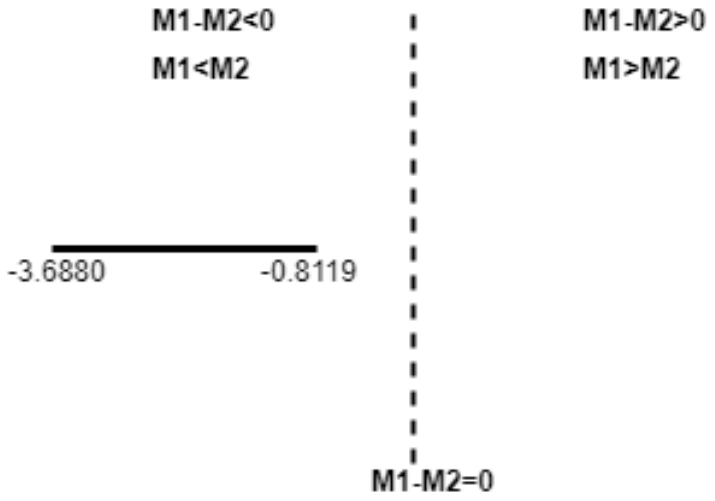

Figura 10: Interpretacion intervalo para diferencia de medias en muestras relacionadas

Como asignamos una confianza del 95 % a nuestro intervalo, quiere decir que el 95 % de las veces que se repita el experimento, la diferencia real de las medias va a caer entre (-3.6880 y -0.8119). Al ver la imagen, observamos que el intervalo obtenido no contiene al cero y ademas esta debajo de el, lo que nos indica que $\mu_2 > \mu_1$, y esto nos dice que el promedio de ventas despues de la campaña publicitaria es mayor que el promedio de ventas antes de la campaña.

En conclusion, podemos decir que la campaña publicitaria es efectiva ya que aumenta el promedio de ventas.

4. Situación 4

4.1. Punto a.

4.2. Punto b.

4.3. Punto c.

5. Situación 5

Como las muestras de las básculas estan relacionadas, ya que una de ellas tiene sus medidas certificadas mientras que la otra se deben verificar. Entonces para encontrar un intervalo de confianza debemos usar la estimación para la diferencia de medias para muestras relacionadas.
La formula que tenemos para este tipo de estimación es:

$$IC(\mu_1 - \mu_2)_{(1-\alpha)\,\%} = (\bar{x}_1 - \bar{x}_2) \pm t_{(n_1+n_2-2;1-\frac{\alpha}{2})} Sp\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

En este tipo de intervalo de confianza, se basa en la diferencia entre cada uno de los valores de las bascula en la cual se quiere verificar su estado de calibracion con respecto a la báscula que se encuentra certificada, se tienen los siguientes valores:

BÁSCULA 1	11.23	14.36	8.33	10.5	23.42	9.15	13.47	6.47	12.4	19.38
BÁSCULA 2	11.27	14.41	8.35	10.52	23.41	9.17	13.52	6.46	12.45	19.35

Encontramos que:

$\bar{x}_1 = 12.871$ y $\bar{x}_2 = 12.891$

$S^2x_1 = 26.687$ y $S^2x_2 = 26.593$

$Sx_1 = 5.166$ y $Sx_2 = 5.156$

$1 - \alpha = 0.98$, $\alpha = 0.02$

Ahora, reemplazando en la formula para hallar el intervalo de confianza, nos queda:
$IC(\mu_1 - \mu_2)_{(1-\alpha)\,\%} = (12.871 - 12.891) \pm t_{(18;0.99)}(5.16)\sqrt{\frac{1}{10} + \frac{1}{10}}$

$IC(\mu_D)_{98\,\%} = [-0.02 \pm 2.552(5.16)\frac{1}{\sqrt{5}}]$
$IC(\mu_D)_{98\,\%} = [-5.91; 5.87]$

Con un nivel de confianza del 98 % se puede decir que la bascula 1 con respecto a la bascula 2 obtiene un buen rendieminto

6. Situación 6

6.1. Punto a.

Para darnos cuenta si las varianzas se puden considerar como iguales o diferentes, debemos encontrar un intervalo de confianza para la razon de las varianzas.
Extrayendo la informacion del enunciado, tenemos:
$1 - \alpha = 0.98$ entonces $\alpha = 0.02$
Ahora, la formula para obtener un intervalo de confianza para la razon de varianzas es:

$$IC(\frac{\sigma_1^2}{\sigma_2^2})_{(1-\alpha)\,\%} = \left[\frac{S_1^2}{S_2^2} \cdot F_{(\frac{\alpha}{2},n_2-1,n_1-1)}; \frac{S_1^2}{S_2^2} \cdot F_{(1-\frac{\alpha}{2},n_2-1,n_1-1)}\right]$$

Hallamos S_1^2 y S_2^2:
$S_1^2 = \frac{\sum_{i=1}^{10}(x_{1i}-\bar{x_1})^2}{9} = 76875.9550$ y $S_2^2 = \frac{\sum_{i=1}^{8}(x_{2i}-\bar{x_2})^2}{7} = 1044021.331$
Reemplazando todo en la formula del intervalo, nos queda:

$$IC(\frac{\sigma_1^2}{\sigma_2^2})_{98\,\%} = \left[\frac{76875.9550}{1044021.331} \cdot F_{(0.01,7,9)}; \frac{76875.9550}{1044021.331} \cdot F_{(0.99,7,9)}\right]$$

$$IC(\frac{\sigma_1^2}{\sigma_2^2})_{98\,\%} = [0.0765 \cdot 0.14884; 0.0765 \cdot 5.61287]$$

$$IC(\frac{\sigma_1^2}{\sigma_2^2})_{98\,\%} = [0.01138; 0.42938]$$
Para interpretar mas facilmente el intervalo obtenido, hicimos la siguiente grafica:

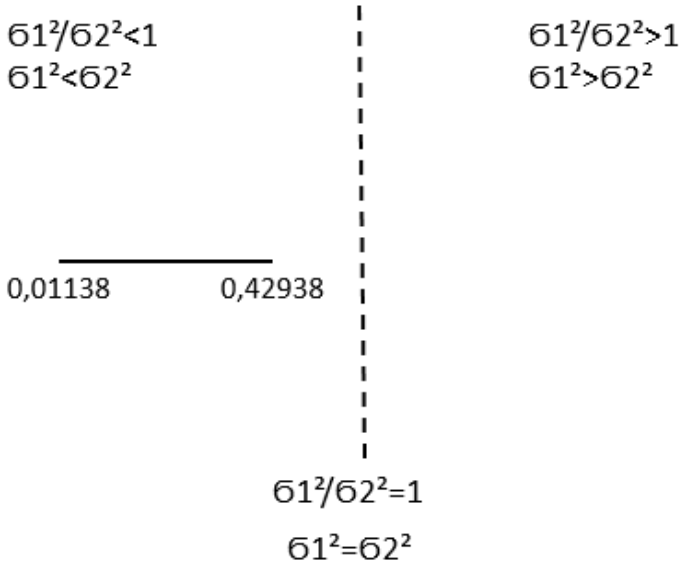

Figura 11: Interpretacion intervalo para la razon de varianzas

Como asignamos una confianza del 98 % quiere decir, que el 98 % de las veces que se repita el

experimento con las mismas condiciones, la razon de varianzas poblacionales va a estar entre (0.01138 y 0.42938). Si nos fijamos en la imagen, vemos que el intervalo no contiene al 1, y cae por debajo de el. Entonces concluimos que no se pueden considerar las varianzas poblacionales como iguales, y se tiene fuerte evidencia de que $\sigma_1^2 < \sigma_2^2$.

6.2. Punto b.

Para recomendar o no el uso del revestimiento como mecanismo complementario, encontraremos un intervalo de confianza para diferencia de medias, y asi darnos cuenta si la resistencia promedio de las tuberias aumento, disminuyo, o se mantuvo igual.
Como informacion tenemos: $\bar{x_1} = 2902.8$, $\bar{x_2} = 2783.125$, $S_1^2 = 76875.9550$ y $S_2^2 = 1044021.331$

Debemos aplicar la estimacion para diferencia de medias con σ_1^2 y σ_2^2 desconocidas pero diferentes, ya que en el punto anterior, mostramos que con una confianza del 98 % las varianzas poblacionales no van a ser iguales.
Sustituyendo los valores en la formula, tenemos:

$$IC(\mu_1 - \mu_2)_{98\,\%} = (2902.8 - 2783.125) \pm t_{(0.99;16)} \cdot Sp\sqrt{\tfrac{1}{10} + \tfrac{1}{8}}$$

Debemos hallar Sp:
$$Sp = \sqrt{\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}} = \sqrt{\frac{9\cdot 76875{,}9550+7\cdot 1044021{,}331}{10+8-2}} = \sqrt{500002.057} = 707.1082$$

Ahora, reemplazando todo, nos queda:

$$IC(\mu_1 - \mu_2)_{98\,\%} = [119.675 \pm 2.58349 \cdot 707.1082 \cdot 0.47434]$$

$$IC(\mu_1 - \mu_2)_{98\,\%} = [-746.8526; 986.2026]$$

INTERPRETACION:

Para interpretar mas facilmente el intervalo obtenido, elaboramos la siguiente grafica:

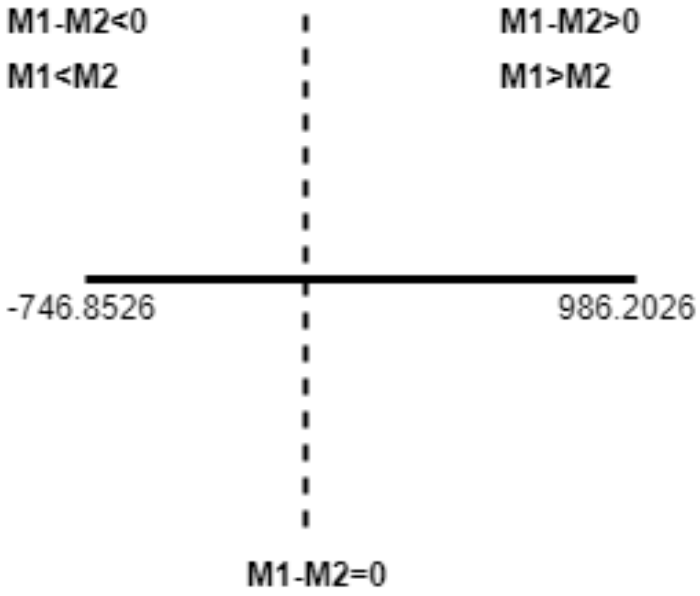

Figura 12: Interpretacion intervalo para diferencia de medias en muestras independientes

Como el intervalo encontrado tiene una confianza del 98 %, decimos que el 98 % de las veces que se repita el experimento en las mismas condiciones, la diferencia de medias de resistencia reales va a caer entre -746.8526 y 986.2026. Ademas, si detallamos la imagen, podemos ver que el intervalo obtenido contiene al cero, esto quiere decir que $\mu_1 = \mu_2$ o son muy cercanos. Por lo que concluimos que el uso del revestimiendo no eleva la resistencia promedio de las tuberias, y no es recomendable el uso de este como mecanismo complementario.

7. Situación 7

Para comparar los tres estimadores propuestos para el CV, se genero una poblacion para cada distribucion (normal, gamma y uniforme) y se obtuvieron 5000 muestras aleatorias para cada uno de los distintos tamaños de muestra (5,10,20,30,...,100). Las poblaciones sobre las cuales se obtuvieron las muestras fueron definidas de la siguiente manera:
1. $Normal(\mu = 200, \sigma = \sqrt{400})$
2. $Gamma(\alpha = 0.5, \beta = 100)$
3. $Uniforme(a = 200, b = 600)$

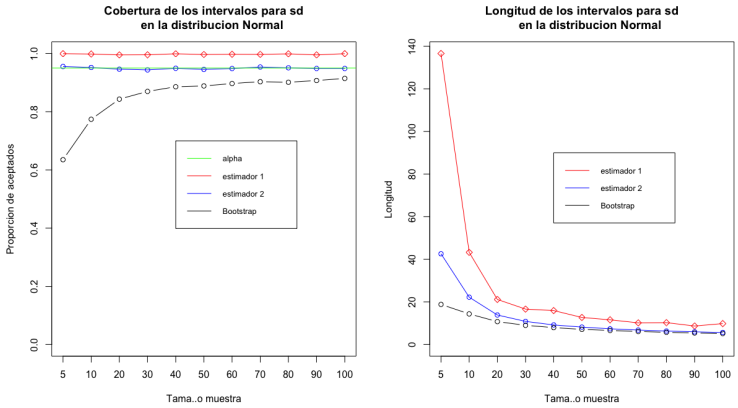

Figura 13: Proporcion de cobertura y longitud promedio de los intervalos para σ de la distribucion normal

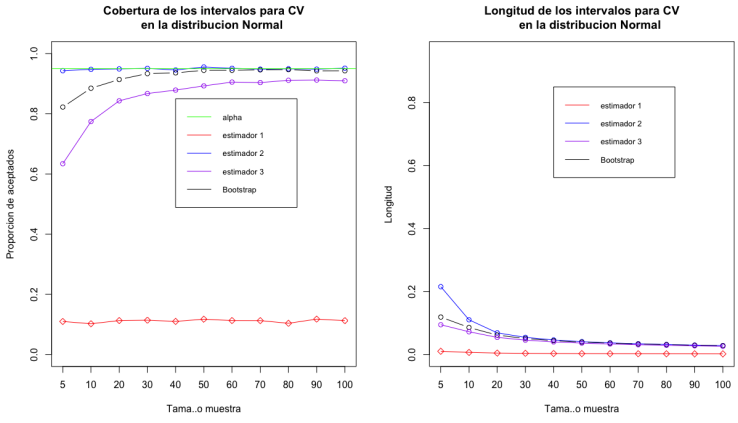

Figura 14: Proporcion de cobertura y longitud promedio de los intervalos para cv de la distribucion normal

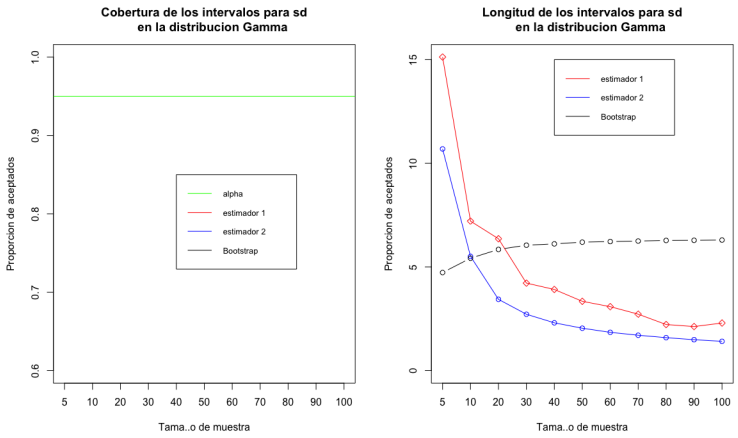

Figura 15: Proporcion de cobertura y longitud promedio de los intervalos para σ de la distribucion gamma

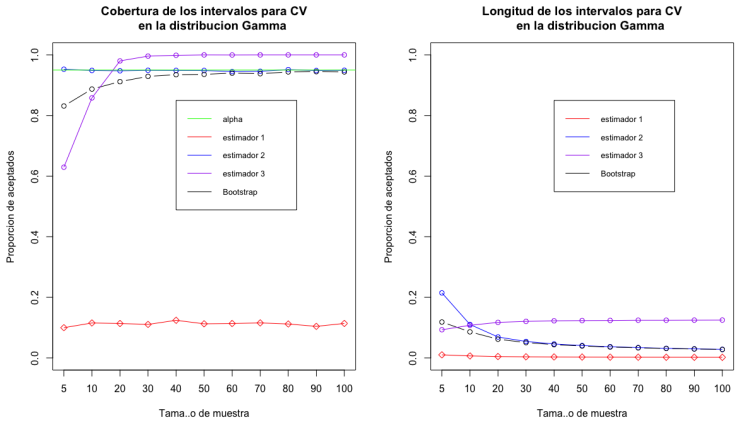

Figura 16: Proporcion de cobertura y longitud promedio de los intervalos para cv de la distribucion gamma

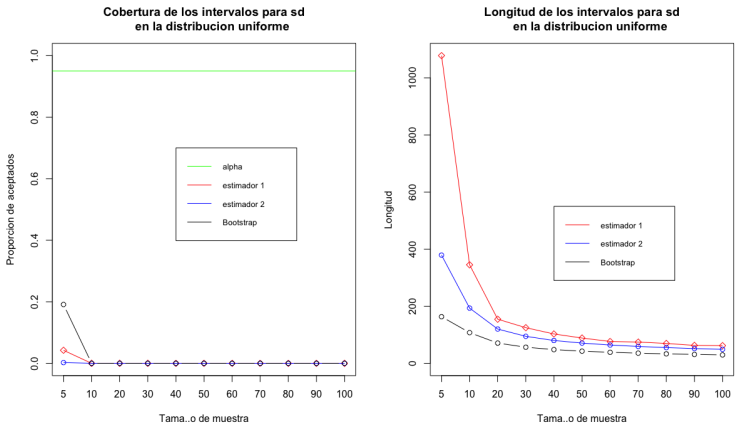

Figura 17: Proporcion de cobertura y longitud promedio de los intervalos para σ de la distribucion uniforme

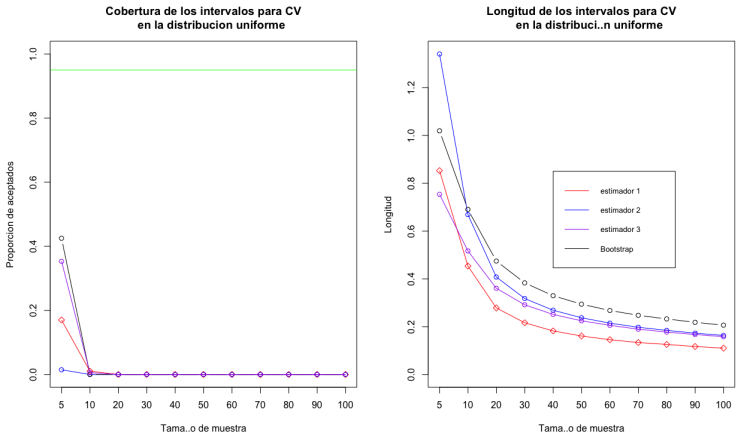

Figura 18: Proporcion de cobertura y longitud promedio de los intervalos para cv de la distribucion uniforme