UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE INGENIERÍA

DIVISIÓN DE INGENIERÍA MECÁNICA E INDUSTRIAL

DEPARTAMENTO DE INGENIERÍA INDUSTRIAL

APUNTES DE ESTADÍSTICA APLICADA

PROYECTO PAPIME PE 107916

Trabajo realizado mediante el Programa de Apoyo a Proyectos para la Innovación y Mejoramiento de la Enseñanza (PAPIME) – Proyecto PE 107916 "Empleo del lenguaje R en la enseñanza de Estadística Aplicada"

CIUDAD UNIVERSITARIA 2017

Índice

Capítulo 1 Lo Que Debería Saber 1
- 1.1. Distribución Muestral 3
- 1.2. Media de las Medias 5
- 1.3. Teorema del Límite Central 9
- 1.4. Uso de la Distribución Muestral 11
- 1.5. Relaciones y Diferencias Entre la Distribución de Probabilidad, la Distribución Muestral y una Muestra de Tamaño n 11
- 1.6. Intervalos de Confianza 16
- 1.7. Ejercicios de Aplicación 22
- 1.8. Pruebas de Hipótesis 24

Capítulo 2 Muestreo 29
- 2.1. Muestreo Irrestricto Aleatorio 30
- 2.2 Muestreo Irrestricto Aleatorio para Proporciones 38
- 2.3. Muestreo Estratificado 42

Capítulo 3 Regresión Lineal 49
- 3.1. Introducción 49
- 3.2. Modelo de Regresión Simple 51
- 3.3. Modelo de Regresión Lineal Múltiple 58
- 3.4. Ejemplo Integrador Modelo 1 60
- 3.5. Ejemplo Integrador Modelo 2 67

Capítulo 4 Diseño de Experimentos 73
- 4.1. El Experimento y sus Fines (Conceptos Básicos de Experimentos) 73
- 4.2. Diseño del Experimento 73
- 4.3. Número de Ensayos 74
- 4.4. Análisis de Variaciones 75
- 4.5. Análisis de Resultados 75
- 4.6. Modelos de Análisis de varianza por uno y dos criterios de variación 75
- 4.7. Criterios de Comparaciones Múltiples 85
- 4.8. Modelos de Bloques Completos 88
- 4.9. Modelos de Cuadrados Latinos 90
- 4.10. Modelos de Cuadrados Grecolatinos 93
- 4.11. Análisis de Covarianza 97
- 4.12. Diseño Factorial 2K 99

Capítulo 5 Estadística no Paramétrica 103
- 5.1. Prueba de los Signos 104
- 5.2. Prueba de Rangos con Signos de Wilcoxon 108
- 5.3. Prueba de Kruskal-Wallis 110
- 5.4. Prueba de Rachas 111
- 5.5. Prueba de Kolmogorov-Smirnov 113
- 5.6. Prueba de Correlación de Rangos 117

Capítulo 6 Confiabilidad 119
- 6.1. Confiabilidad, Usos y Aplicaciones 119
- 6.2. Distribuciones del Tiempo de Falla 119
- 6.3. Distribución Exponencial 120
- 6.4. Distribución Gama 121
- 6.5. Distribución Weibull 121
- 6.6. Sistemas en Serie y en Paralelo 121
- 6.7. Modelo Exponencial en Confiabilidad 122

Bibliografías 123

Capítulo 1

Lo Que Debería Saber

Definiciones básicas:

- **Población:** Relación completa de todas las observaciones de interés para el investigador.
- **Parámetro:** Medida descriptiva de la población.
- **Muestra:** Parte representativa de la población que se selecciona para ser estudiada ya que la población es demasiado grande como para analizarla en su totalidad.
- **Estadístico:** Elemento que describe una muestra y sirve como una estimación del parámetro de la población correspondiente.
- **Variable:** Es una característica de la población que se está analizando en un estudio estadístico.
- **Error de muestreo:** Es la diferencia entre el parámetro desconocido de la población y el estadístico de la muestra utilizado para calcular el parámetro.
- **Sesgo muestral:** Es la diferencia que se presenta al favorecer la selección de ciertos elementos de una muestra en lugar de otros.

1 2

[1] El estadístico es a la muestra lo que el parámetro a la población

[2] Existen variables continuas que pueden ser tratadas como discreta, entre ellas está la estatura, el peso, la edad de una persona, que para fines prácticos se tratan como variables discretas por la precisión de los instrumentos de medición empleados para obtener dichos valores simplemente al redondear los valores.

Una variable puede ser cuantitativa o cualitativa, así mismo puede ser continua o discreta. Un ejemplo de una variable cualitativa es el color de ojos, mientras que un ejemplo de una variable cuantitativa puede ser la edad de una persona medida en años. Un ejemplo de una variable discreta es el valor que arroja un dado mientras que un ejemplo de una variable continua es la estatura de una persona. Por otra parte todas las variables cualitativas deben tomar valores cuantitativos para poder ser empleadas en cálculos estadísticos. A la estadística se le puede dividir en dos ramas:

- **Estadística descriptiva:** Es el proceso de recolectar, agrupar y presentar datos de una manera tal, que describa fácil y rápidamente dichos datos.
- **Estadística inferencial:** Involucra el uso de una muestra para sacar alguna inferencia o conclusión sobre la población a la cual pertenece la muestra. Involucra el uso de un estadístico para obtener una conclusión o inferencia sobre el parámetro correspondiente.

1.1. Distribución Muestral

Es muy común que las poblaciones sean demasiado grandes para ser estudiadas en su totalidad, ya que existen poblaciones finitas o infinitas, aún una población finita pude resultar que por motivos de tiempo o recursos (económicos, computacionales, entre otros), se deba emplear una muestra representativa de un tamaño adecuado. Esta muestra se usa para obtener algunas conclusiones de la población. Por ejemplo se puede calcular la media muestral con ayuda del estadístico $\overline{X}$, y utilizarlo como un estimador de la media poblacional μ.
[3]

Es importante recalcar que el valor del estadístico depende de la muestra tomada, en otras palabras, si se tomaran dos muestras aleatorias de la misma población, ambas con el mismo tamaño de muestra, se esperaría que la probabilidad de que la medias muestrales sean iguales entre sí sea casi de cero, lo misma probabilidad ocurriría en caso de que alguna de estas medias fuera igual a la media poblacional, esto se debe a que de cualquier población dada de tamaño N, es posible obtener muchas muestras de tamaño n, con medias diferentes, por lo tanto, es posible obtener una distribución completa de $\overline{X}$ diferentes, de varias muestras posibles.El concepto de distribución muestral lo analizaremos con ayuda del siguiente ejemplo.

Ejemplo **1.1** *Suponga que tenemos una población de $N = 4$ salarios para cuatro estudiantes universitarios. Estos salarios son 1,000, 2,000, 3,000 y 4,000 UM, el ingreso promedio puede calcularse obteniendo $\mu = 2{,}500$ UM. Suponga que calcular la media de cuatro observaciones requiere de mucho esfuerzo. Como alternativa se decide seleccionar una muestra de $n = 2$ observaciones para estimar a la μ "desconocida". En este caso se podría seleccionar una muestra de ${}_4C_2 = 6$ posibles maneras, como se muestra en la siguiente tabla, así como sus respectivas medias muestrales y error de muestreo.*

Muestra	Elementos muestrales [UM]		Medias muestrales $\bar{X}_i$ [UM]	Error de muestreo
1	1,000	2,000	1,500	**1,000**
2	1,000	3,000	2,000	**500**
3	1,000	4,000	2,500	**0**
4	2,000	3,000	2,500	**0**
5	2,000	4,000	3,000	**500**
6	**3,000**	**4,000**	**3,500**	**1,000**

Cuadro 1.1: Medias Muestrales

Recuerde que el error de muestreo es la diferencia entre el parámetro de la población, en este caso 2,500 y el estadístico de la muestra como por ejemplo para la muestra cinco es de 3,000, lo que da un error de 500, en este caso recuerde que un error se trata en valor absoluto, ya que la desviación que se presenta tanto hacia arriba como hacia abajo es igual de importante.

[3] Recuerde que el estadístico se utiliza como estimador del parámetro, por tal motivo al confiar en una muestra para sacar alguna conclusión o inferencia sobre la población, se está hablando de la estadística inferencial

En la realidad no se puede calcular el tamaño exacto del error para una muestra debido a que la media poblacional es desconocida, sin embargo se debe estar consciente de que es probable que ocurra algún error de muestreo.

- **Error de muestreo:** Es la diferencia entre el parámetro poblacional y el estadístico de la muestra utilizado para estimar el parámetro.

Con una población N = 4 se puede enumerar cada media muestral posible junto con su respectiva probabilidad. Tal listado se le denomina una distribución muestral.

Medias muestrales $\bar{X}_i$ [UM]	Número de muestras	$P(\bar{X}_i)$
1,500	1	**1/6**
2,000	1	**1/6**
2,500	2	**1/3**
3,000	1	**1/6**
3,500	**1**	**1/6**

Cuadro 1.2: Probabilidad de las Medias Muestrales

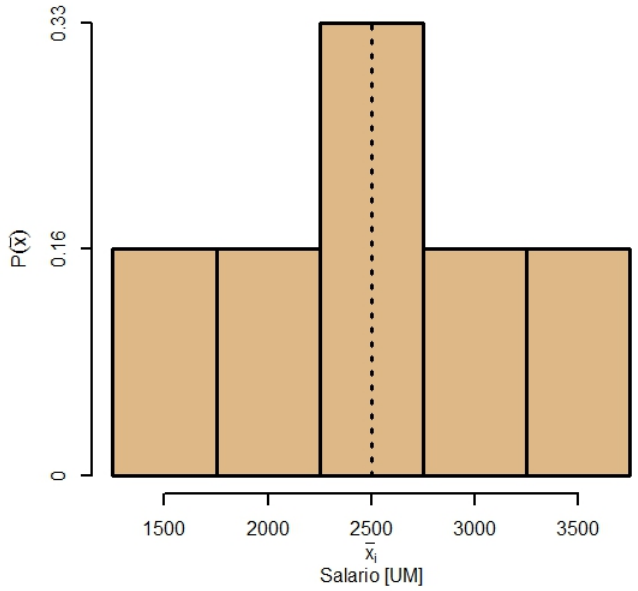


Figura 1.1: Distribución Muestral para el Ejemplo de los Salarios

- **Distribución Muestral:** Es una lista de todos los valores posibles para un estadístico y la probabilidad asociada a cada valor.

En este punto se puede recordar que existen las distribuciones de probabilidad, la cual es una función que asocia a cada evento con su probabilidad, dichas distribuciones pueden ser discretas o continuas. La diferencia entre una distribución de probabilidad y una distribución muestral, radica en que la variable aleatoria de la segunda es la media de la muestra asociada.

1.2. Media de las Medias

Se observa que la distribución muestral de las medias muestrales es simplemente una lista de todas las medias muestrales posibles. Por lo tanto, al igual que cualquier lista de números, tienen una media denominada "la media de las medias muestrales" o "la gran media", que para el caso del ejemplo de los salarios se obtiene de la siguiente manera.

$$\bar{\bar{X}} = \frac{1500 + 2000 + 2500 + 2500 + 3000 + 3500}{6} = 2500$$

$$\bar{X} = \mu = 2500$$

Un punto a recalcar es que la media de las medias o la gran media es igual a la media poblacional, lo cual no es un resultado extraño, ya que la media es una medida de la tendencia central. La expresión para el cálculo de la gran media es la siguiente.

$$\bar{\bar{X}} = \frac{\sum_{i=1}^{k} \bar{x}_i}{K}$$

En donde K son las combinaciones tomadas de m, n a la vez.

$$_mC_n = K$$

Continuación del *Ejemplo* **1.1**:

Con los datos del ejemplo anterior se puede obtener la distribución de probabilidades y la distribución muestral tomando muestras de n = 3.

Observemos la distribución de probabilidad.

N=4 Alumnos

Salarios: $x_1 = 1000$ $x_2 = 2000$ $x_3 = 3000$ $x_4 = 4000$

Distribución de probabilidades.

$$\mu = \frac{\sum_{i=1}^{n} x_i}{N} = \frac{1000 + 2000 + 3000 + 4000}{4} = 2500$$

$$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N} = \frac{(1000 - 2500)^2 + (2000 - 2500)^2 + (3000 - 2500)^2 + (4000 - 2500)^2}{4} = 1250000$$

$$\sigma = \sqrt{\sigma^2} = \sqrt{12500000} = 1118{,}03$$

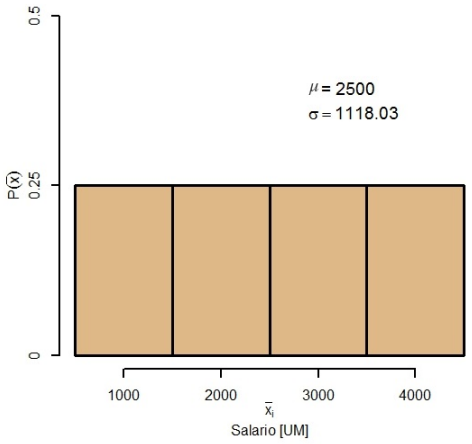


Figura 1.2: Distribución de Probabilidad

Distribución muestral con n=3.

$$_mC_n = k = 4$$

K	$_N C_n$	$\bar{X}_i$	1/k
1	1000, 2000, 3000	2000	1/4
2	1000, 2000, 4000	2333.33	1/4
3	1000, 3000, 4000	2666.67	1/4
4	2000, 3000, 4000	3000	1/4

Cuadro 1.3: Medias Muestrales con n=3

Distribución de probabilidades.

$$\bar{\bar{X}} = \frac{\sum_{i=1}^{k} \bar{x}_i}{K} = \frac{2000 + 2333{,}3 + 2666{,}6 + 3000}{4} = 2500$$

$$\sigma^2{}_{\bar{x}} = \frac{\sum_{i=1}^{k} (\bar{x}_i - \bar{\bar{x}})^2}{k} = \frac{(2000 - 2500)^2 + (2333{,}3 - 2500)^2 + (2666{,}6 - 2500)^2 + (3000 - 2500)^2}{4} = 138886{,}1125$$

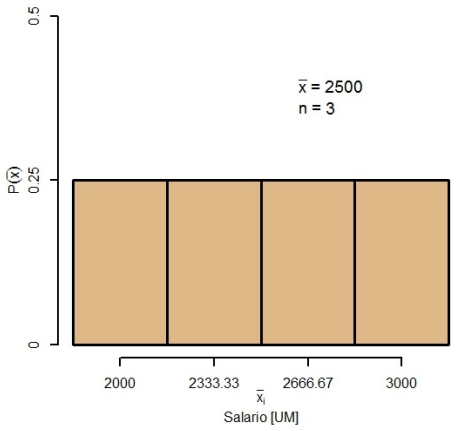


Figura 1.3: Distribución de Medias

[4]

[4]Con esto se concluye que con n = 3 el error estándar de estimación disminuye con respecto a n = 2 ya que 645.49 > 262.81.

Varianza de la distribución muestral de las medias muestrales.

$$\sigma_{\bar{x}}{}^2 = \frac{\sum_{i=1}^{k} (\bar{x}_i - \bar{\bar{x}})^2}{k} = \frac{\sum_{i=1}^{k} (\bar{x}_i - \mu)^2}{k}$$

Error estándar de la distribución muestral de las medias muestrales.

$$\sigma_{\bar{x}_i} = \sqrt{\sigma_{\bar{x}_i}{}^2}$$

Como se habrá observado, la fórmula requiere de mucha aritmética, una aproximación cercana puede obtenerse mediante:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

La fórmula anterior es apropiada sólo si el muestreo se realiza con reemplazo, o si la muestra se toma de una población muy grande (virtualmente infinita). Si el muestreo se realiza sin reemplazo y si el tamaño de la muestra es más del 5 % de la población, n >0.05N, debe aplicarse el factor de corrección para poblaciones finitas (cpf).

Error estándar utilizando el cpf:

$$\sigma_{\bar{x}_i} = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{n-1}}$$

El impacto del tamaño de la muestra en el error estándar.

Indiscutiblemente es probable un estimado más exacto con una muestra más grande. Esto puede verificarse ya que a medida que “n” aumenta, $\sigma_{\bar{x}}$ disminuye.

1.3. Teorema del Límite Central

Es evidente que es posible tener muchas muestras de un tamaño dado de cualquier población, lo que da pie a toda una distribución de medias muestrales. Si la población original está distribuida normalmente, la distribución de las medias muestrales también estará distribuida normalmente.

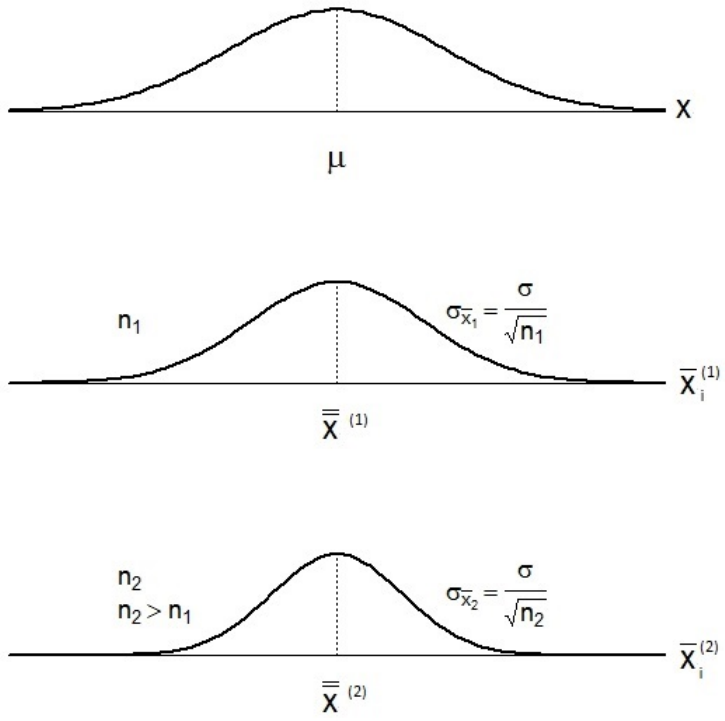


Figura 1.4: Teorema de Límite Central

El la figura superior se muestra la distribución de la observaciones individuales x_i que esta normalmente distribuida. En la gráfica inferior "la distribución de las medias muestrales" que resulta de tomar todas las muestras de tamaño $n_1 = 25$ también esta distribuida normalmente y centrada en la media poblacional. La dispersión de la población original σ= 50 es mayor que la dispersión $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$ de las media muestrales.

¿Cómo será la distribución de las medias muestrales si la población original no esta distribuida normalmente?

Teorema 1.1 *A medida que n aumenta, la distribución de las medias muestrales se aproxima a una distribución normal con una media* $\bar{\bar{x}} = \mu$ *y un error estándar* $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$

[5]

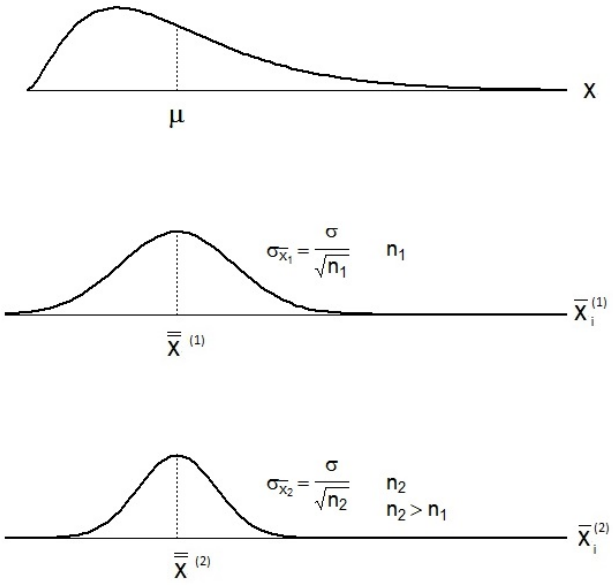


Figura 1.5: Teorema de Límite Central

[5] ***Teorema del Límite Central***

1.4. Uso de la Distribución Muestral

Se sabe que para determinar la probabilidad de seleccionar una observación que estuviera dentro de un rango dado, se puede estandarizar a través de "Z" a una distribución de probabilidad normal.

$$Z = \frac{x_i - \mu}{\sigma}$$

Sin embargo, muchas decisiones dependen de una muestra completa, no sólo de una observación. En este caso la fórmula anterior bebe modificarse.

$$Z = \frac{\bar{x_i} - \mu}{\sigma_{\bar{x}}} = \frac{\bar{x_i} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

1.5. Relaciones y Diferencias Entre la Distribución de Probabilidad, la Distribución Muestral y una Muestra de Tamaño n

Debemos recordar de nuestros cursos de Probabilidad y Estadística que cuando trabajamos con la Distribución de Probabilidad, estamos hablando que conocemos que forma tiene, esto es, si es una Normal, una t de Student, una Xi Cuadrada, una Rectangular, una Pearson IV o V, en el caso de variables aleatorias discretas vemos una Binomial, Bernoully, Poisson, entre otras. Todas ellas se distinguen por los parámetros que las caracterizan (recuerde siempre esto, el parámetro es a la población como el estadístico a la muestra), los más comunes son los primeros y segundos momentos, entre ellos están la media y la varianza, sin embargo existen distribuciones de probabilidad poblacional que requieren de alguno o otros parámetros para quedar bien definidas. En nuestro estudio de las Distribuciones de Probabilidad nos concentraremos solo en los primeros dos momentos, por lo que estudiaremos cómo calcular a la media y la varianza de una población, en este punto cabe aclarar que es mejor usar a la desviación estándar que a la varianza de una población, esto se debe básicamente a que es más sencillo trabajar en unidades con las que estamos familiarizados y que son consistentes con el análisis, por ejemplo suponga que está trabajando con el precio de un producto, la varianza arrojaría pesos al cuadrado ($U.M.^2$), cosa que no es consistente con el análisis y en la realidad no existe una interpretación económica o financiera ante este fenómeno, por lo que es más sencillo trabajar con la desviación estándar, que es simplemente la raíz cuadrada de la varianza y que en términos generales es consistente con las unidades de la variable aleatoria y con el análisis.

Comencemos con identificar la media, la varianza y la desviación estándar de la Distribución de Probabilidad. La media poblacional que comúnmente se denota con la letra griega μ se calcula como sigue:

$$\mu = E[x] = \sum_{i=1}^{N} \{p_i x_i\} = \frac{1}{N} \sum_{i=1}^{N} \{x_i\}, p_i \quad cuando \quad es \quad discreta.$$

Cabe aclara que N puede ser finita o infinita y que sólo se hace la simplificación de dividir entre N cuando se sabe que la probabilidad de que ocurra cada una de las variables aleatorias es la misma.

Ahora pasemos a la expresión matemática de la Varianza que se representa como σ^2:

$$\sigma^2 = Var[x] = E[(x-\mu)^2] = \sum_{i=1}^{N} \{p_i (x_i - \mu)^2\} = \frac{1}{N} \sum_{i=1}^{N} \{(x_i - \mu)^2\} p_i \quad cuando \quad es \quad discreta.$$

Finalmente la desviación estándar es:

$$\sigma = \sqrt{\sigma^2}$$

Continuando con nuestra presentación, ahora toca el turno de la Distribución Muestral, esta se caracteriza porque no importando de donde provengan los datos, se sugiere que se tomen muestras de tamaño mayor o igual a 30, la distribución resultante se aproximará a la normal y a medida que n tienda a N (finita o infinita) la varianza de la Distribución Muestral disminuirá y su raíz cuadrada que ahora denotaremos como error estándar también lo hará. Comencemos con la media de la Distribución Muestral:

$$\bar{\bar{x}} = \mu = E[\bar{x}] = \sum_{i=1}^{k} \{p_i \bar{x_i}\} = \frac{1}{k} \sum_{i=1}^{N} \{\bar{x_i}\}, \quad de \quad forma \quad discreta.$$

Cabe aclarar que ahora no contamos con N elementos (finitos o infinitos) sino más bien con K combinaciones tomadas n a la vez de una población de tamaño N. También es importante señalar que la media de la Distribución Muestral es la misma que la media de la Distribución de Probabilidad.

$${\sigma^2}_{\bar{x}} = Var[\bar{x}] = E[(\bar{x_i} - \mu)^2] = \sum_{i=1}^{k} \{p_i (\bar{x_i} - \mu)^2\} = \frac{1}{k} \sum_{i=1}^{k} \{(\bar{x_i} - \mu)^2\}, \quad de \quad forma \quad discreta.$$

Note que la variable aleatoria no es la misma, en la Distribución Muestral la variable aleatoria es el promedio que se obtiene de cada una de las muestras de tamaño n. Finalmente tenemos al Error Estándar el cual se calcula como la raíz cuadrada de la varianza de la Distribución Muestral.

$$\sigma_{\bar{x}} = \sqrt{{\sigma_{\bar{x}}}^2}$$

Ahora tomemos uno de estas variables aleatorias de la Distribución Muestral y denotémosla como $\bar{x_i}$ ¿Cómo se llegó a ella? Es muy sencillo, una vez que se definió el tamaño de muestra que se tomaría, en este caso cualquier valor mayor o igual que 30, se tenían n elementos que se escogen de la población y se procede a los siguientes cálculos

Media de una muestra:

$$\bar{x_i} = \frac{1}{n} \sum_{i=1}^{n} \{x_i\}$$

Varianza de una muestra :

$$s_i^2 = \frac{1}{n-1} \sum_{i=1}^{n} \{(x_i - \bar{x})^2\}$$

Desviación estándar de una muestra:

$$s_i = \sqrt{s_i^2}$$

Ahora observemos cómo es que se relacionan:

Muestra		Poblacional
$\overline{X}_i$	Estima a	μ
S_i^2	Estima a	σ^2

Cuadro 1.4: Parámetros Poblacionales y Estimadores Muestrales

Por otra parte, se ha demostrado que se puede calcular el error estándar de la Distribución Muestral a partir de la desviación estándar de la Distribución de Probabilidad y es de la siguiente forma:

$$\sigma_{\bar{x}} \cong \frac{\sigma}{\sqrt{n}}$$

Entonces por obvias razones, el error estándar de la Distribución de Muestral también puede ser aproximado con:

$$s_{\bar{x}} = \frac{s_i}{\sqrt{n}}$$

En este momento pareciera no relevante esta última relación, sin embargo, en intervalos de confianza y pruebas de hipótesis es una gran ayuda saber esto.

Terminemos con un ejemplo que involucra algunos de los conceptos vistos en este subtema.

En la siguiente figura se muestra la distribución exponencial y posteriormente se muestran las distribuciones muestrales con n=2, n=6 y n=30.

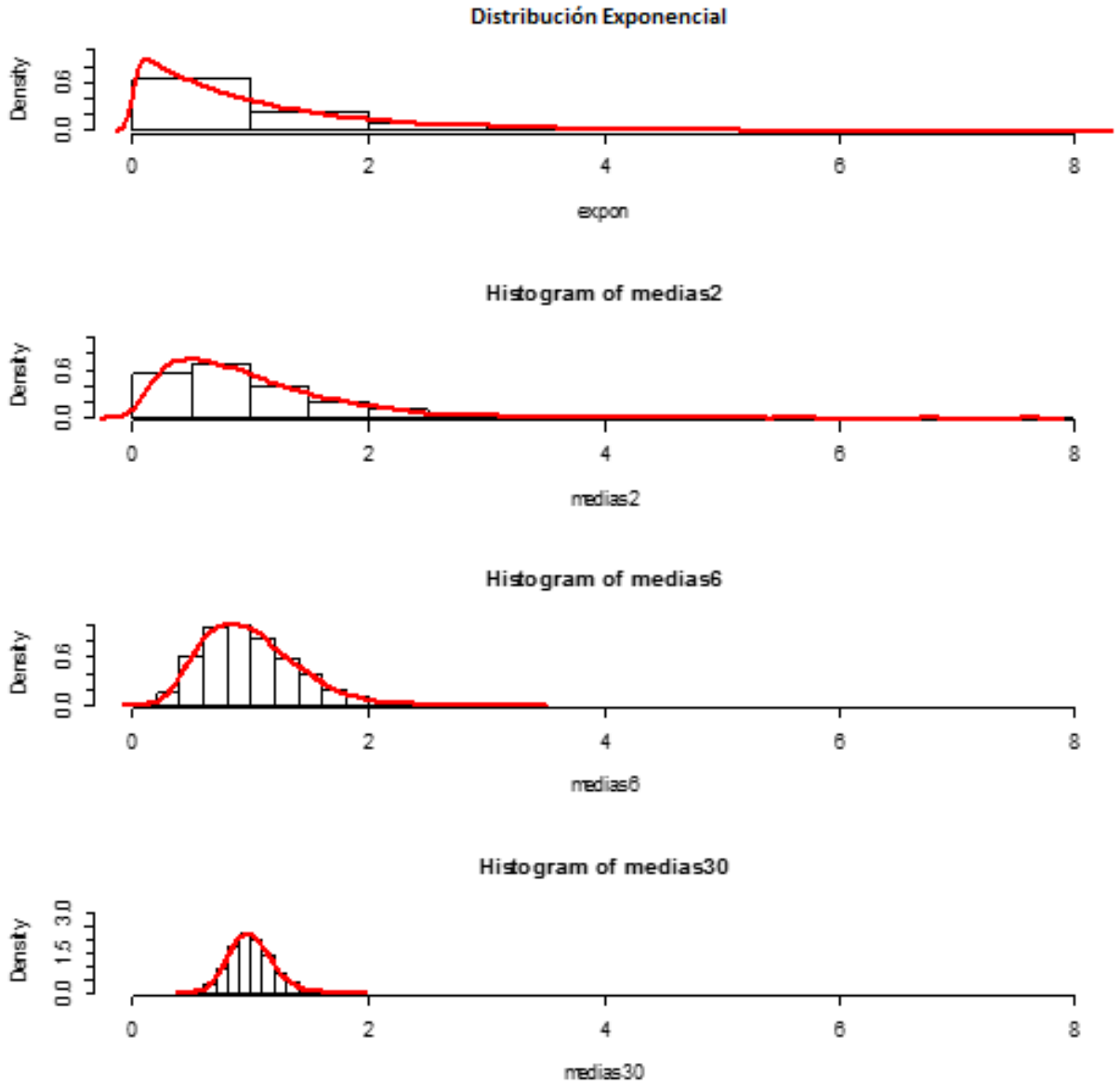


Figura 1.6: Teorema de Límite Central

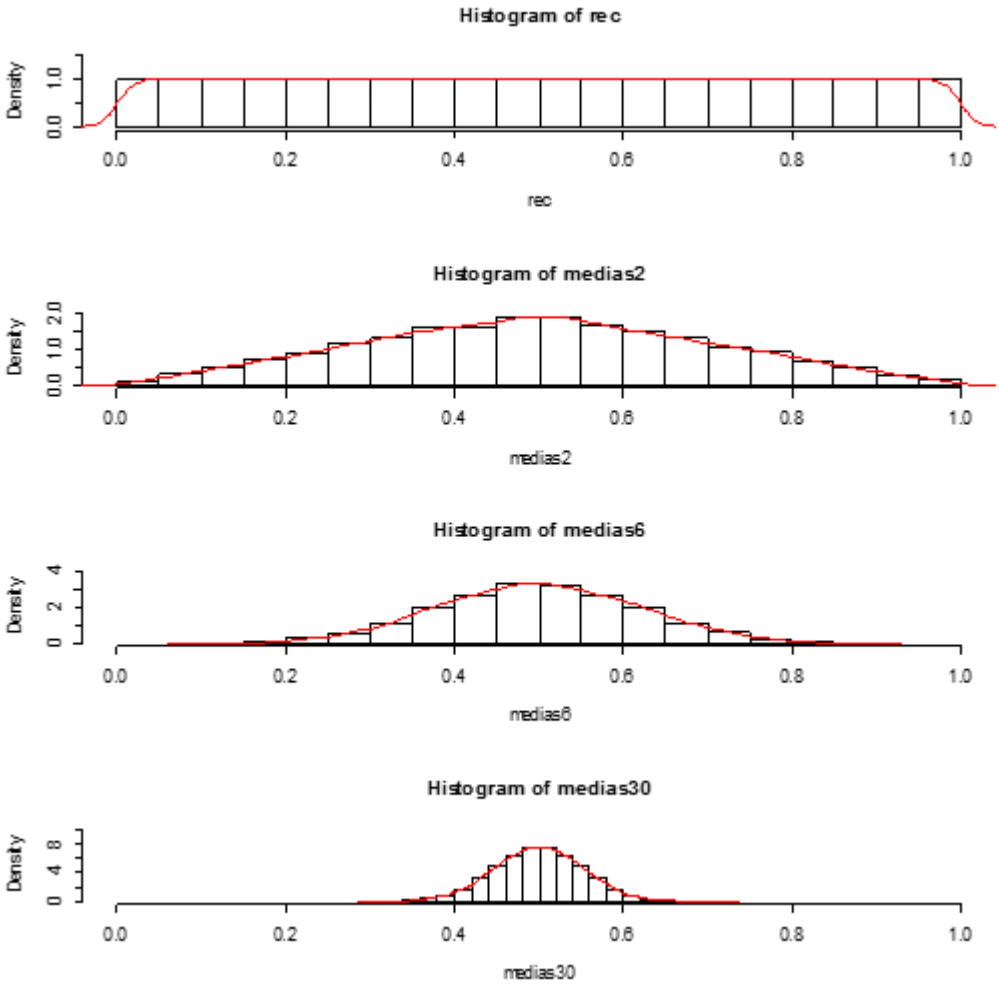


Figura 1.7: Teorema de Límite Central

Al igual que en la figura anterior se aplica el mismo tratamiento a una distribución rectangular.

1.6. Intervalos de Confianza

Recordemos nuevamente que se debe estar consciente de que las poblaciones son generalmente muy grandes como para ser estudiadas en su totalidad, por lo tanto se requiere que se seleccionen muestras, las cuales se pueden utilizar para hacer inferencia sobre las poblaciones.

Hay por lo menos dos tipos de estimadores: un estimador puntual y un estimador por intervalo.

Un estimador puntual utiliza un estadístico para estimar el parámetro en un solo punto o valor.

Un estimador por intervalo especifica el rango dentro del cual puede estar el parámetro desconocido. Tal intervalo con frecuencia va acompañado de una afirmación acerca del nivel de confianza que se da sobre su exactitud. Por lo tanto se llama Intervalo de Confianza (I. C.).

Estimador puntual y por intervalo. Un estimador puntual utiliza un número único o valor para localizar una estimación del parámetro. Un intervalo de confianza denota un rango dentro del cual puede encontrarse el parámetro, así como el nivel de confianza de que el intervalo contiene al parámetro en cuestión.

- **El fundamento de un intervalo de confianza.**
  Un intervalo de confianza tiene un límite inferior de confianza [LIC] y un límite superior de confianza [LSC]. Estos límites se hallan primero calculando la media muestral, $\bar{x}$. Luego se suma una cierta cantidad a $\bar{x}$ para obtener el LSC, y la misma cantidad se resta de $\bar{x}$ para obtener el LIC, como ya sabemos a esta cantidad se le conoce como error "E".
  ¿Cómo se puede construir un intervalo y luego argumentar que se puede tener un 95 % de confianza en que contiene μ, si incluso no se sabe cuál es la media poblacional? Vale la pena recordar que el 95.5 % de todas las medias muestrales caen dentro de dos errores estándar de la media poblacional. Entonces la media poblacional está máximo a dos errores estándar del 95.5 % de todas las medias muestrales. Por tanto, al comenzar con cualquier media muestral, si se pasa de dos errores estándar por encima de dicha media y dos errores estándar por debajo de ella, se puede tener un 95.5 % de confianza en que el intervalo resultante contenga la media poblacional desconocida.
  Sabemos que de toda población se pueden obtener muchas muestras diferentes de un tamaño dado, cada una con su propia media, como se ilustra en la siguiente figura.

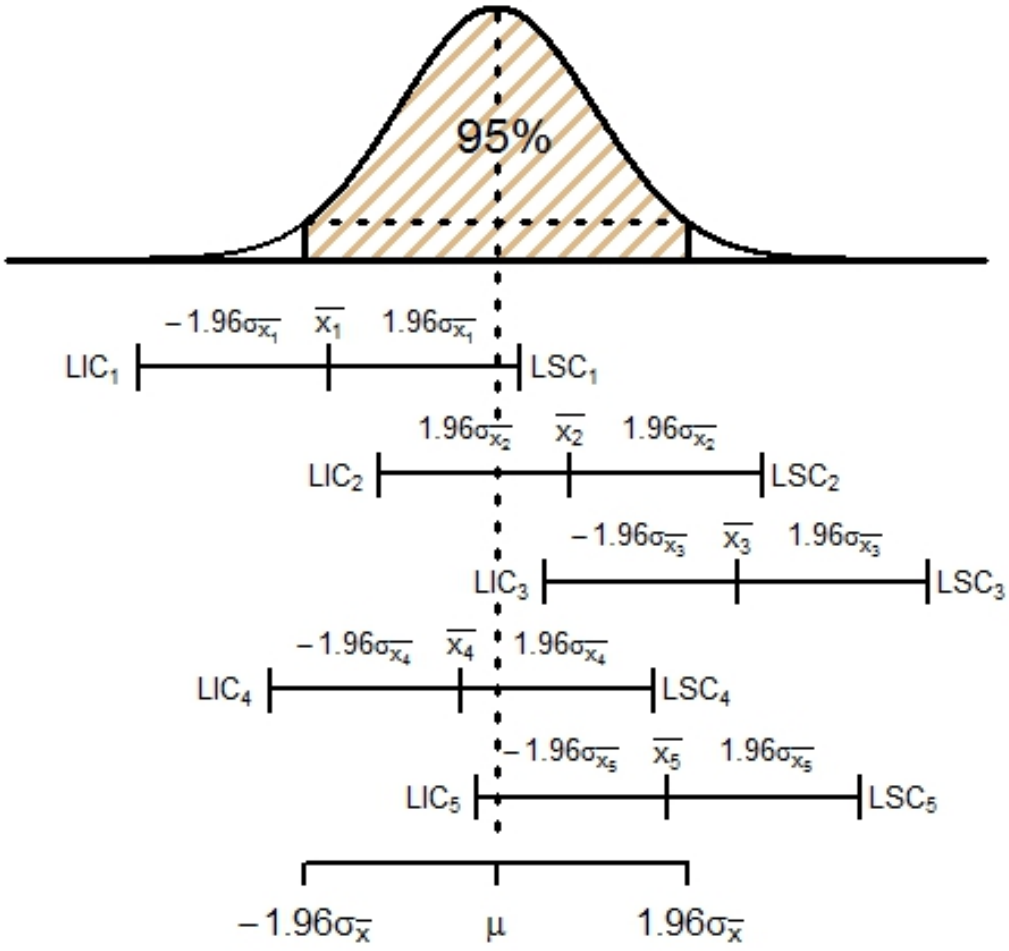


Figura 1.8: Intervalos de Confianza

Si se desea construir un intervalo más convencional del 95 % (en lugar del 95.5 % discutido anteriormente) el valor de z debe ser 1.96 para construir un intervalo de confianza del 95 %, simplemente se especifica un intervalo de 1.96 errores estándar por encima y por debajo de la media muestral. Este valor del 95 % es llamado coeficiente de confianza como se muestra en la figura siguiente.

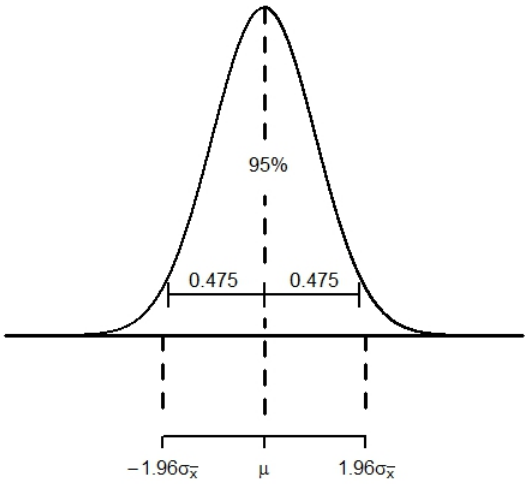


Figura 1.9: Intervalos de Confianza

Coeficiente de confianza: El coeficiente de confianza es el nivel de confianza que se tiene de que el intervalo contenga el valor desconocido del parámetro.

Intervalo de confianza para la media poblacional "Muestras Grandes"
Intervalo de confianza para estimar μ cuando σ es conocido.

I.C. para estimar μ, I.C.[LSC, LIC]

$$\bar{x} \pm z_{\frac{\alpha}{2}} \sigma_{\bar{x}}$$

- **Interpretación de un intervalo de confianza.**

Se puede interpretar lo resultados de un intervalo de confianza de dos formas. La primera, y la más común, establece que se tiene un 95 % de confianza en que la media poblacional real pero desconocida esté entre el LIC y LSC. Aunque el valor real para la media poblacional sigue siendo desconocido, se tiene un 95 % de confianza en que esté entre estos dos valores.

La segunda interpretación reconoce que se pueden desarrollar muchos intervalos de confianza diferentes. Otra muestra probablemente produciría una media muestral diferente debido al error de muestreo.

Con una muestra diferente, el intervalo tendría límites superior e inferior distintos. Por tanto, la segunda interpretación establece que si se construyen todos los ${}_NC_n$ intervalos de confianza, el 95 % de ellos contendrá a la media poblacional real pero desconocida.

Esto por supuesto significa que el 5 % de todos los intervalos estaría errado, no contendría a la media poblacional. Este 5 %, hallado como (1-coeficiente de confianza), es denominado valor alfa o nivel de significancia y es la probabilidad de que cualquier intervalo dado no contenga la media poblacional.

Valor alfa: Es la probabilidad de que un intervalo dado no contenga a la media poblacional desconocida.

Intervalo de confianza cuando σ es desconocida.
Intervalos de confianza para estimar μ cuando σ es desconocida

$$\mu = \bar{x} \pm \quad z_{\frac{\alpha}{2}} s_{\bar{x}} = \bar{x} \pm \quad z_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}}$$

Intervalo de confianza para la media en el caso de muestra pequeñas "La Distribución T".

Cuando debe tomarse una muestra pequeña, la distribución normal puedo no aplicarse. El TLC asegura normalidad en el proceso de muestreo sólo si la muestra es grande. Cuando se utiliza una muestra pequeña, puede ser necesaria una distribución alternativa, la distribución t de Student. Específicamente, la distribución t se utiliza cuando se cumplen tres condiciones:

1. La muestra es pequeña.
2. σ es desconocida.
3. La población de donde se extraen las muestras es normal o casi normal.

Intervalo de confianza para estimar la media poblacional

I.C. para estimar μ, I.C.[LSC, LIC]

$$\mu = \bar{x} \pm \quad t_{\frac{\alpha}{2}} s_{\bar{x}} = \bar{x} \pm \quad t_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}}$$

Control de ancho de un intervalo

- Reducción del nivel de confianza.
- Incremento del tamaño muestra.

Determinación del tamaño apropiado de la muestra.

Tamaño de la muestra para intervalos de la media poblacional.

$$n = \frac{z_{\frac{\alpha}{2}}^2 \sigma^2}{(\bar{x} - \mu)^2}$$

Donde $(\bar{x} - \mu)$ es el error de muestreo "E".

Propiedades de un buen estimador.

- Estimador insesgado. Un estimador es insesgado si el valor esperado de su distribución muestral es igual al parámetro correspondiente.
- Estimador eficiente. Dado todo estimador insesgado, el estimador más eficiente es aquel que tenga la varianza más pequeña.
- Estimador consistente. Un estimador es consistente si a medida que n aumenta el valor de estadístico se aproxima al parámetro.
- Estimador suficiente. Un estimador es suficiente si ningún otro estimador puede proporcionar más información sobre el parámetro.

Ejemplo **1.2** *Calcular el precio promedio del huevo en el Distrito Federal.*

Nivel de Significancia: $\alpha = 5\,\%$
C.C. $= 95\,\%$
Error$= 0.5\$$

La pregunta es: ¿Qué tamaño de muestra usar?

$$n = \frac{z_{\frac{\alpha}{2}}^2 \sigma^2}{e^2}$$

σ^2 no es conocida y se estima con s^2.

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

Como se necesita saber s^2 se hace un muestreo piloto donde $n^p = 30$.

$$s_p^2 = 9$$
$$n = 129$$

Luego se pone un nuevo error y una nueva confiabilidad en caso de que s^2 de la muestra n=139 sea menor. Es correcto en caso de que salga mayor.

Calcular con e= 0.5 y $z_{\frac{\alpha}{2}} = 1{,}96$.

$$n = \frac{(1{,}96)^2(9)^2}{(,5)^2} = 138{,}29 = 139$$

Se recalcula s^2.

$$s^2 = 10$$

Como $s^2 = 10 > s_p^2 = 9$, el error aumenta o la confiabilidad disminuye. En este caso calculamos el error:

$$e = 0{,}52$$

Y los intervalos de confianza será:

$$IC = 21{,}49 \quad - \quad 22{,}52$$

Fijando el error, la confiabilidad será.

$$z_{\frac{\alpha}{2}} = 1,86$$

Por lo tanto:

$$C.C. = 93{,}71\,\%$$

1.7. Ejercicios de Aplicación

Ejemplo **1.3** *Se desea saber el diámetro promedio poblacional de la circunferencia de los CD's que son fabricados por cierta empresa, dicho diámetro debe ser de 120 milímetros. Con el uso de intervalos de confianza se desea estimar el valor poblacional del diámetro, para lo que se le solicita que emplee un 95.5 % de confiabilidad y los siguientes errores en milímetros:*

Número	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Error	1.5	1.4	1.3	1.2	1.1	1	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1

Cuadro 1.5: Tabla 6

1. *Suponga que conoce la varianza poblacional, la cual es igual a 21.333, encuentre los tamaños de la muestra para el error y la confiabilidad especificados.*
2. *Realice una gráfica de los errores y el tamaño de la muestra.*
3. *Encuentre los respectivos intervalos de confianza.*

1. Distintos tamaños de muestra.

Número	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Error	1.5	1.4	1.3	1.2	1.1	1	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1
n	38	44	51	60	71	86	106	134	175	238	342	534	949	2134	8534

Cuadro 1.6: Tabla 7

2. Error vs Tamaño de la muestra.

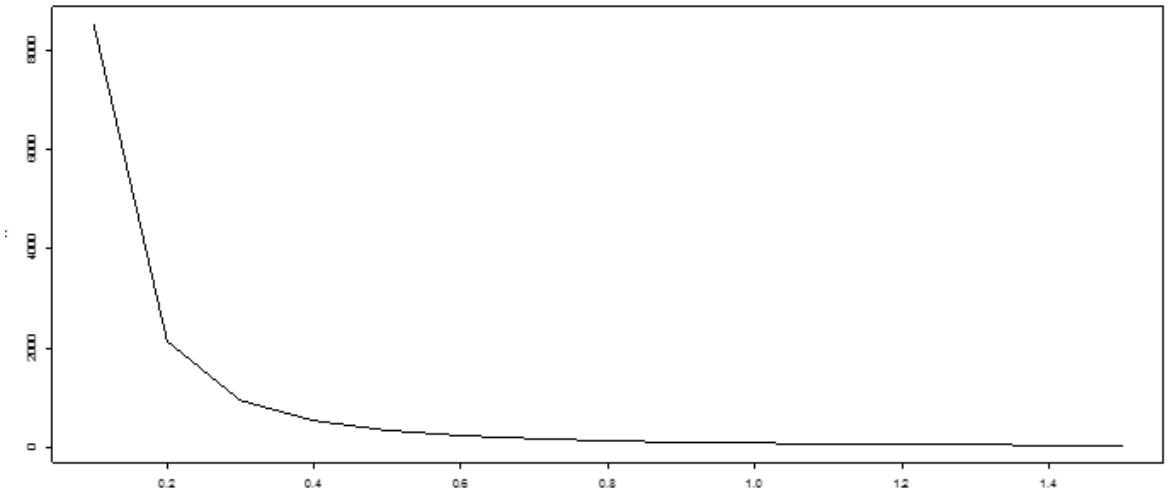


Figura 1.10: Error vs Tamaño de Muestra

3. Intervalos de confianza.

Número	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Error	1.5	1.4	1.3	1.2	1.1	1	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1
n	38	44	51	60	71	86	106	134	175	238	342	534	949	2134	8534
LS	118.6129	118.4081	116.9807	118.7009	117.0334	118.4119	117.3568	117.7029	117.7095	117.7717	117.3711	117.2302	117.4445	117.2774	117.033
LI	115.6129	115.6081	114.3807	116.3009	114.8334	116.4119	115.5568	116.1029	116.3095	116.5717	116.3711	116.4302	116.8445	116.8774	116.833

Cuadro 1.7: Tabla 8

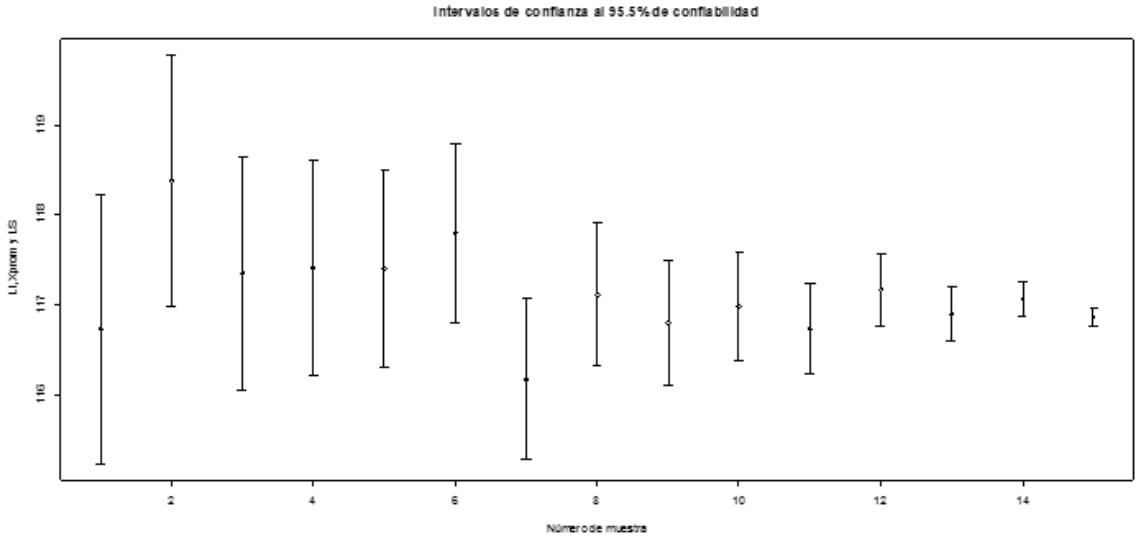


Figura 1.11: Intervalos de Confianza al 95 %

Número	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Error	1.5	1.4	1.3	1.2	1.1	1	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1
n	38	44	51	60	71	86	106	134	175	238	342	534	949	2134	8534
LS	118.6129	118.4081	116.9807	118.7009	117.0334	118.4119	117.3568	117.7029	117.7095	117.7717	117.3711	117.2302	117.4445	117.2774	117.033
LI	115.6129	115.6081	114.3807	116.3009	114.8334	116.4119	115.5568	116.1029	116.3095	116.5717	116.3711	116.4302	116.8445	116.8774	116.833
Xprom	117.1129	117.0081	115.6807	117.5009	115.9334	117.4119	116.4568	116.9029	117.0095	117.1717	116.8711	116.8302	117.1445	117.0774	116.933

Cuadro 1.8: Resumen del Análisis

1.8. Pruebas de Hipótesis

Para realizar una prueba de hipótesis se hacen algunas inferencias o supuestos con sentido acerca de la población. Un embotellador de bebidas debe asumir, o plantear la hipótesis de que el contenido promedio de sus botellas es de 16 onzas (μ=16). Esta hipótesis nula (H_0) se prueba contra la hipótesis alternativa (H_A) que establece lo contrario. En este caso, el contenido promedio no es de 16 onzas ($\mu \neq 16$).

Por tanto se tendría:

$$H_0 : \mu = 16$$
$$H_1 : \mu \neq 16$$

El término nula indica nada o nulo. El término surge de sus primeras aplicaciones por parte de los investigadores agrícolas quienes probaron la efectividad de un nuevo fertilizante para determinar su impacto en la producción de la cosecha. Asumieron que el fertilizante no hacía ninguna diferencia en el rendimiento hasta que éste produjo algún efecto. Por tanto, la hipótesis nula, tradicionalmente contiene alguna referencia de un signo como " $=$ "," $\leq$ "," $\geq$ ".

Con base en los datos muestrales, esta hipótesis nula es rechazada o no rechazada. Nunca se puede "aceptar" la hipótesis nula como verdadera. El no rechazo de la hipótesis solamente significa que la evidencia muestral no es suficientemente fuerte como para llevarla a su rechazo.

Una analogía es que probar una hipótesis es como poner una persona en juicio. El acusado se halla o culpable o no culpable. Un veredicto no culpable simplemente significa que la evidencia no es lo suficientemente fuerte como para encontrar culpable al acusado. No significa que él o ella sean inocentes.

Cuando se realiza una prueba de hipótesis nula se supone que es "inocente" (verdadero) hasta que una preponderancia de la evidencia indique que es "culpable" (falso). Al igual que en un escenario legal, la evidencia de culpable debe establecerse más allá de toda duda razonable. Antes que se rechace la hipótesis nula, la media muestral debe diferir significativamente de la media poblacional planteada como hipótesis. Es decir, que la evidencia debe ser muy convincente y concluyente. Una conclusión con base en un rechazo de la hipótesis nula es más significativa que una que termine en una decisión de no rechazo.

Diferencia estadísticamente significante. Es la diferencia entre el valor de la media poblacional bajo hipótesis y el valor de la media muestral que es lo suficientemente pequeña como para atribuirla a un error de muestreo.

¿Esto significa que la diferencia entre el valor de la media de 16 bajo la hipótesis y el hallado en una muestra con un valor de 16.15 es insuficiente para rechazar la hipótesis nula?

El asunto simplemente es encontrar que tan grande debe ser la diferencia para que sea estadísticamente significativa y conduzca un rechazo de la hipótesis nula. Sabemos que:

σ conocida y n ≥ 30	σ desconocida y n ≥ 30	σ desconocida, n ≤ 30 y la distribución es normal o casi normal
$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$	$Z = \frac{\bar{X} - \mu}{S_{\bar{x}}} = \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}}$	$T = \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}}$

[6]

[6] Si σ es desconocida se utiliza la desviación estándar muestral S.

- Valores críticos de z y zonas de rechazo.

  Los valores de z son los valores críticos que determinan la(s) zona(s) de rechazo que se encuentran en el nivel de significancia, o el valor alfa "α" de la prueba.

  Estos valores permiten establecer una regla de decisión que diga si existe evidencia para rechazar la hipótesis nula o no.

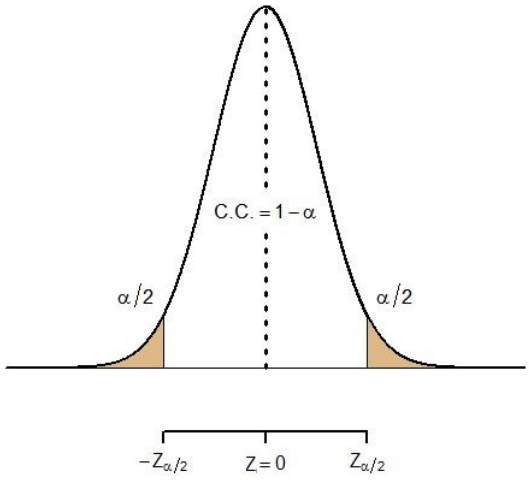


Figura 1.12: Prueba de dos colas

- El nivel de significancia y la probabilidad del error

  Al probar una hipótesis se pueden cometer dos tipos de errores.

  Error tipo I: Rechazar una hipótesis verdadera. La probabilidad de cometer un error de tipo I es igual al nivel de significancia, o valor α en el que se prueba la hipótesis.

  Error tipo 2: Es no rechazar una hipótesis nula que es falsa, la probabilidad de un error tipo II, se representa con la letra β, no se determina fácilmente. No se puede asumir que $\alpha + \beta = 1$.

Pasos involucrados en una prueba de hipótesis.

Existen cuatro pasos involucrados en una prueba:

1. Plantear las hipótesis.
2. Con base en los resultados de la muestra calcular el estadístico de prueba Z.
3. Determinar la regla de decisión con base en los valores críticos de Z.
4. Interpretación y conclusión.

Pruebas de una cola.

Existen ocasiones en las que se está interesado sólo en un extremo u otro. Por ejemplo, un restaurante de comida fresca no se interesa en que tan rápido llegan las langostas provenientes de la costa. Se preocupa sólo si el envío se toma mucho tiempo en llegar. Una tienda minorista sólo se alarmará si los ingresos caen a niveles demasiado bajos. En particular las ventas altas no son problema. En cada uno de estos casos la preocupación se concentra en un extremo u otro y se realiza una prueba de una cola.

$$H_0 : \mu = 16$$
$$H_1 : \mu \neq 16$$

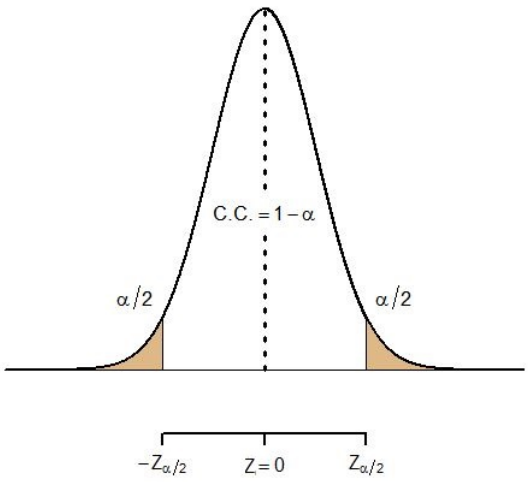


Figura 1.13: Prueba de dos colas

$$H_0: \quad \mu \geq 16$$
$$H_1: \quad \mu < 16$$

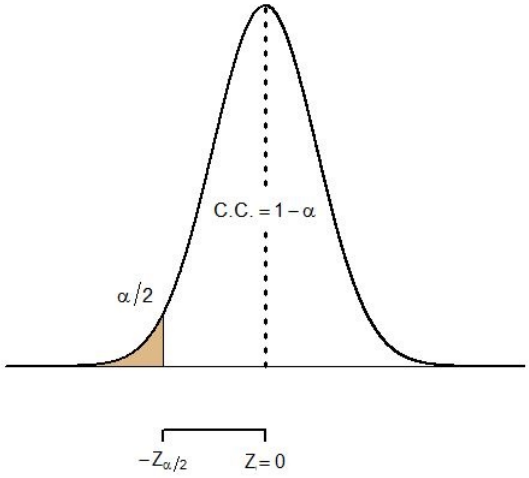


Figura 1.14: Prueba de cola izquierda

$$H_0: \quad \mu \leq 16$$
$$H_1: \quad \mu > 16$$

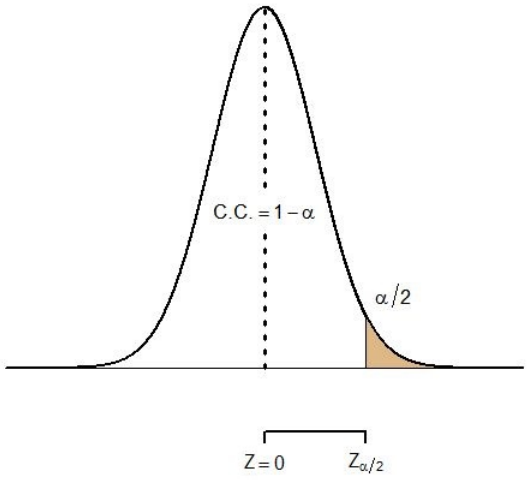


Figura 1.15: Prueba de cola derecha

Valores P: Uso e interpretación.

Como se ha visto para probar una hipótesis se calcula un valor Z y se compara con un valor crítico Z con base en un nivel de significancia seleccionado. Mientras que el valor P de una prueba puede servir como método alternativo para probar hipótesis en realidad es mucho más que eso.

El valor P para una prueba es la probabilidad de obtener resultados muestrales al menos tan extremos como los que se obtuvieron dado que la hipótesis nula es verdadera.

Valor P Es el nivel más bajo de significancia (valor α) al cual se puede rechazar la hipótesis nula. Es el área en la cola que está más allá del valor del estadístico para la muestra.

Capítulo 2

Muestreo

En el capítulo "Lo que debería saber"se trató sobre el tema de las distribuciones muestrales, estimación puntual, estimación por intervalos de confianza y finalmente sobre las pruebas de hipótesis. En este apartado se abordarán temas como el muestreo irrestricto aleatorio y el muestreo estratificado. Dichos temas se basan en gran medida en lo visto en el capítulo "Lo que debería saber".

Antes de iniciar con el primero de ellos (muestreo aleatorio irrestricto), comenzaremos preguntando ¿por qué es importante el muestreo? Como se sabe, existen poblaciones que son infinitas en su dimensión o demasiado grandes para ser censadas. También existen poblaciones finitas de un tamaño adecuado, sin embargo, existen otros factores como lo son el tiempo y dinero que hacen del muestreo una alternativa atractiva para el investigador (cualquier persona con el interés de conocer alguna característica sobre alguna población), buscando minimizar los costos y el tiempo invertido, maximizando la calidad de la información. Cuando nos referimos a maximizar la calidad de la información, en realidad lo que esperamos es que la varianza de la misma sea mínima. Los factores de tiempo y dinero se pueden resumir aún más, alguna vez habrá escuchado la frase “el tiempo es dinero”, por lo que podemos decir que el muestrear permite minimiza las cantidades monetarias necesarias, maximizando la calidad de la información recabada.

Entonces ¿cuándo es importante muestrear? Cuando los recursos con los que disponemos no permitan censar, o cuando la precisión de la información, comparada con las unidades de recursos invertidas, no agreguen más valor al investigador.

Algunas de las ideas vistas en los párrafos anteriores serán abordados con detalle en cada uno de las técnicas de muestreo mencionadas. Comencemos con el muestreo irrestricto aleatorio.

2.1. Muestreo Irrestricto Aleatorio

El muestreo irrestricto aleatorio deriva su nombre, al dar la misma probabilidad de que una muestra sea seleccionada. Supongamos que tenemos una población de elementos, del cual extraemos una muestra lo suficientemente grande, como sabemos, por el Teorema del Límite Central, nos indica que la distribución de las medias muestrales tenderán a formar una campana, que se aproximará asintóticamente a la distribución normal a medida que se aumente el tamaño de la muestra y disminuirá su error estándar, como se plantea en la introducción. Lo que se espera en particular es que la media de la muestra sea un buen estimador de la media poblacional, en otras palabras, se necesita que la media muestral por lo menos sea un estimador insesgado y eficiente, en otras palabras:

$$E(\bar{X}) = \mu \tag{2.1}$$

$$Var(\bar{X}) = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)$$

$$Var(\bar{X}) = \sigma^2_{\bar{X}} \tag{2.2}$$

Como pude apreciar el factor $\frac{N-n}{N-1}$ se le conoce como factor de corrección por población finita. El cual radica su importancia en que si se tiene una población finita pequeña "N" y se toma una muestra "n" lo suficientemente grande, la varianza muestral disminuirá, a continuación se presenta una gráfica con datos hipotéticos para una población de 20,000, tomando muestras de tamaño 30 e incrementándola de diez en diez hasta llegar a 1,000, para ver el impacto que tiene este factor en porcentajes.

Código 1

```
#cpf Es la variable factor de corrección por población finita

N<-20000
n<-seq(30,1000, by=10)
cpf<-(N-n)/(N-1)*100
plot(cpf~n)
```

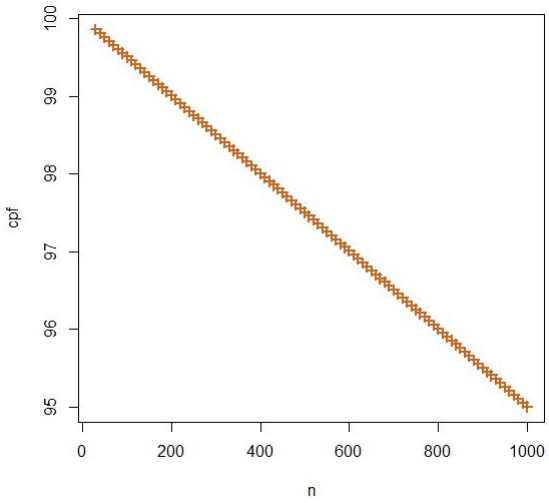


Figura 2.1: pfc vs tamaño de muestra "n"

Como se puede apreciar al tomar un tamaño de muestra de 1,000 elementos, además de que la varianza muestral disminuirá por un tamaño de muestra más grande, también lo hará en aproximadamente un 5 % más, por el cpf que aproximadamente vale 0.95.

Nuevamente recordemos que estamos estimando parámetros poblacionales a partir de estadísticos y que lo que se desea es tener la menor variabilidad de la información recabada, entonces un estimador de la media poblacional y de la varianza muestral son los siguientes:

$$\hat{\mu} = \bar{X} \tag{2.3}$$

$$\hat{\sigma^2}_{\cdot\bar{x}} = \frac{s^2}{n}\left(\frac{N-n}{N-1}\right) \tag{2.4}$$

Donde, como recordará, la varianza muestral estimada la denotamos como:

$$\hat{\sigma^2}_{\cdot\bar{x}} = {s_{\bar{X}}}^2 \tag{2.5}$$

Con la única diferencia dada por el factor $\frac{N-n}{N-1}$, que prácticamente se vuelve uno cuando la población es muy grande o tiende al infinito con “n” constante.

Cuando N es muy grande, $N \equiv N-1$.

$$lim_{N\to\infty}\left(\frac{N-n}{N}\right)$$

$$lim_{N\to\infty}\left(\frac{N}{N}-\frac{n}{N}\right)$$

$$lim_{N\to\infty}(1) - lim_{N\to\infty}\left(\frac{n}{N}\right) = 1 - 0 = 1 \tag{2.6}$$

Por lo tanto, el error estándar estimado será:

$$\hat{\sigma^2}_{\cdot\bar{x}} = {s_{\bar{X}}}^2$$

$$\hat{\sigma^2}_{\cdot\bar{x}} = s_{\bar{X}}$$

$$s_{\bar{X}} = \frac{s}{\sqrt{n}}$$

$$s_{\bar{X}} = \frac{s}{\sqrt{n}}\sqrt{\frac{N-n}{N}} \tag{2.7}$$

Suponga que establecemos un tamaño para el error de la estimación que se plantea hacer, llamémoslo "B", recordando que la distribución de las medias de las muestras se aproximan asintóticamente a una normal podemos emplear un valor de z de alfa medios, tal que se encuentre a z de alfa medios desviaciones estándar aproximadamente.

Demosración 2.1

$$B = z_{\frac{\alpha}{2}}\sqrt{{\sigma_{\bar{X}}}^2}$$

Considere

$$D = \frac{B}{z_{\frac{\alpha}{2}}}$$

Entonces

$$D = \sqrt{{\sigma_{\bar{X}}}^2}$$

Donde

$${\sigma_{\bar{X}}}^2 = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)$$

Por lo tanto

$$D = \sqrt{\frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)}$$

Despejando "n" tenemos

$$D^2 = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)$$

$$D^2 = \frac{\sigma^2}{n}\left(\frac{N}{N-1} - \frac{n}{N-1}\right)$$

$$\frac{D^2}{\sigma^2} = \frac{1}{n}\left(\frac{N}{N-1} - \frac{n}{N-1}\right)$$

$$\frac{D^2}{\sigma^2} = \frac{N}{n(N-1)} - \frac{1}{N-1}$$

$$\frac{D^2}{\sigma^2} + \frac{1}{N-1} = \frac{N}{n(N-1)}$$

$$\frac{(N-1)D^2 + \sigma^2}{\sigma^2(N-1)} = \frac{N}{n(N-1)}$$

$$n = \frac{N\sigma^2(N-1)}{((N-1)D^2 + \sigma^2)(N-1)}$$

$$n = \frac{\sigma^2 N}{(N-1)D^2 + \sigma^2} \tag{2.8}$$

Ejemplo **2.1** *Suponga que se tiene una población de tamaño 1000, de la cual se sabe que su varianza poblacional para el parámetro estudiado es de 20, se desea un nivel de significancia de alfa del 5 % y un límite para el error de 4. ¿De qué tamaño debe ser la muestra?*

Código Ejemplo 3.1

```
B<-4
alfa<-.05
sig<-20
N<-1000

D<-B/qnorm(alfa/2,0,1)
n<-sig^2*N/((N-1)*D^2+sig^2)
```

El tamaño de la muestra deberá ser 218.354, que en nuestro caso será de 219 elementos muestreados aleatoriamente.

De la ecuación 2.8 podemos obtener el tamaño de la muestra, siempre y cuando se conozca la varianza poblacional y el tamaño de la población, es claro que es menos probable que conozcamos la varianza poblacional, ya que estamos intentado estimar los parámetros a través de una muestra, esto en general representa un problema, ya que sí es difícil obtener el parámetro poblacional de la media, el de la varianza debe ser un poco más complicado, ya que este último depende de la media poblacional, nuevamente recurriremos a nuestros estadísticos para hacer estimaciones puntuales, en este caso el de la varianza poblacional, a través de la varianza muestral, quedando la siguiente expresión.

$$n = \frac{s^2 N}{(N-1)D^2 + s^2} \tag{2.9}$$

Es evidente que de la ecuación 2.9 se pueden obtener una familia de tamaños de muestra, ya que el valor de la varianza muestral es en sí una variable aleatoria que cambiará de muestra en muestra, esto nos remonta a la paradoja

propuesta en "lo que debería saber", en donde se observa que para obtener el tamaño de la muestra debemos conocer la varianza estimada, pero para conocer la varianza estimada, necesitamos conocer el tamaño de la muestra, en estos casos lo que se sugiere es tomar una muestra piloto o estimar dicha varianza con estudios previos y emplearla para calcular el tamaño de la muestra.

Ejemplo **2.2** *Suponga que se les asignan a diez personas calcular el tamaño de la muestra de una población que se distribuye de forma normal, suponga también que conoce la media poblacional la cual es de 70 y su varianza poblacional la cual es de 100. El tamaño de la población es de 10,000 elementos y el límite para el error es de 3. ¿Cuáles serían algunos de los posibles tamaños de muestra que obtendrían estas diez personas? Suponga que cada uno de ellos toma un muestreo piloto de 30 elementos de forma aleatoria para estimar la varianza poblacional. Tome un nivel de significancia del 5 %.*

Lo primero que habría que hacer es generar 10,000 números aleatorios que se distribuyan de forma normal con media igual a 70 y varianza de 100, posteriormente tomar diez muestras de tamaño 30 para estimar la varianza poblacional y finalmente emplear cada una de estas varianzas estimadas para calcular el tamaño de la muestra empleando la ecuación 2.9.

Código Ejemplo 3.2

```
set.seed(186) #Se establece una semilla aleatoria
N<-10000
sig<-10 #Desviación estándar poblacional
pob<-rnorm(N,70, sig) #Distribución de la población de 10000 elementos

hist(pob)
```

Código Ejemplo 3.2

```
varm<-c() #Varaianzas muestrales
for (i in 1:10){
varm[i]<-var(sample(pob,30,replace=FALSE))
}

B<-3
alfa<-.05
D<-B/qnorm(alfa/2,0,1)
n<-varm*N/((N)*D^2+varm)
```

Los diez posibles resultados para cada una de las estimaciones de las varianza poblacionales serían:

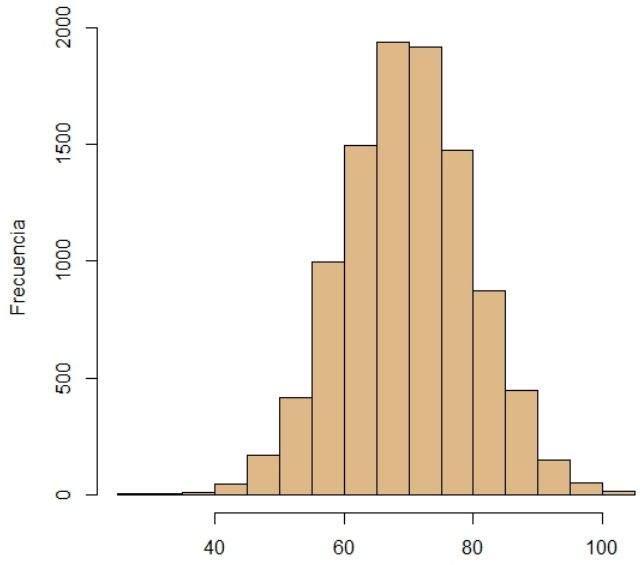


Figura 2.2: Distribución de los 10000 números aleatorios

Código Ejemplo 3.2

```
Persona<-c(1:10)
R2<-data.frame(Persona, varm, n)
R2
```

Como es de esperarse, la varianza de la distribución de las medias de las muestras cambiará para cada una de las diez personas que toma una muestra de tamaño 30 para estimar la varianza poblacional y posteriormente con esta información calcular el tamaño de la muestra para el nivel de error y confiabilidad buscado. Si observa el tamaño de la muestra de la persona ocho, es inclusive más pequeña que el de la muestra tomada al inicio para estimar la varianza de la población. Se puede observar que si aumentamos el tamaño de la muestra para estimar la varianza poblacional, entonces se tendrá un tamaño de muestra y una varianza más parecidas para cada uno de las diez personas, como por ejemplo un tamaño de 100 elementos.

```
> R2
   Persona      varm        n
1        1 124.55502 52.88252
2        2 114.72416 48.72896
3        3 106.72579 45.34706
4        4 113.37940 48.16052
5        5  88.85259 37.78155
6        6  82.77848 35.20784
7        7  84.68015 36.01376
8        8  46.53475 19.82300
9        9  80.44925 34.22055
10      10 118.33318 50.25418
```

Cuadro 2.1: Varianzas de las 10 personas, n=30

Código Ejemplo 3.2

Cambie en el ciclo for el tamaño de muestra 30 por un tamaño de muestra 100.

```
for (i in 1:10){
varm[i]<-var(sample(pob,100,replace=FALSE))
}
```

Ahora corra el código completo otra vez.

```
> R2
   Persona      varm        n
1        1 106.01862 45.04794
2        2  92.83294 39.46738
3        3  80.42294 34.20939
4        4 113.79125 48.33462
5        5 108.89815 46.26582
6        6  90.72720 38.57559
7        7 117.70605 49.98918
8        8  81.45796 34.64813
9        9  84.46534 35.92273
10      10  68.64151 29.21258
```

Cuadro 2.2: Varianzas de las 10 personas, n=100

Como se puede apreciar, la diferencia de la varianza de las muestras de la tabla uno tiene una diferencia entre su valor más grande, respecto al más pequeño de 78, mientras que para el caso de la tabla 2, este valor es de 49, sin

embargo, para calcular estos valores se tomaron 100 elementos de la población para estimar a la varianza, y en ninguno de los casos de la tabla 2 se obtuvo un tamaño de muestra para el error y confiabilidad requeridos mayor a 100.

Lo anterior plantea varias conclusiones, una de ellas, es que cundo no se conoce a la varianza poblacional, se obtendrán familias de tamaños de muestras, porque se deberán estimar la varianza poblacional. Por otro lado, todas aquellas muestras del cuadro 1 y 2 cuya varianza de la muestra sea menor que 100, no cumplirán con el nivel de error o confiabilidad propuesto, mientras que todas aquellas que sean mayores o iguales que 100 si lo harán o los mejorarán.

A continuación, se analizará el cálculo del tamaño de muestra para proporciones.

2.2. Muestreo Irrestricto Aleatorio para Proporciones

En muchas ocasiones lo que interesa es calcular el tamaño de muestra de una población, para conocer su opinión sobre algún tema político o preferencia, como es el caso de las encuestas de opinión sobre algún candidato, lo que le interesa saber a este último, es sí las personas votarán por él o no. Como sabemos este ejemplo muestra una característica del experimento binomial. Ahora bien, si este experimento se tratara igual que en los casos anteriores, supongamos que, si las personas votan por el candidato, entonces les asignamos el valor de uno, y en caso contrario se coloca un cero, entonces para obtener la media se emplea la siguiente expresión:

$$\mu = \sum_{i=1}^{N} \frac{x_i}{N} \tag{2.10}$$

Esto es equivalente a obtener la proporción de personas que votaran por el candidato, la cual expresaremos con la letra "P", por lo tanto:

$$\mu = p \tag{2.11}$$

Lo anterior lo podemos ejemplificar con el siguiente código en "R", suponga que se tiene una población de 1000 votantes de los cuales cierto número de personas votarán por el candidato "A", podemos calcular la media de dichos votantes o la proporción y comprobar que valen lo mismo:

Ejemplo **2.3**

Código Ejemplo 2.3

```
set.seed(186) #Establecemos una semilla
pob<-sample(c(0,1),1000,replace=TRUE) #Población de 1000 elementos
mean(pob) #Media de la población
A<-sum(pob[pob==1]) #Cantidad de personas que votarán por el candidato
A
P<-A/1000 #Proporción de personas que votarán por el candidato A.
```

Al correr el código se puede comprobar que el valor de la media es igual al de la proporción, en este caso de 0.506. De forma similar se puede establecer que la varianza de la proporción es igual a "P*(1-P)" que sería aproximadamente igual a la varianza de la población del ejemplo anterior empleando la fórmula:

$$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \bar{x})^2}{N-1} \tag{2.12}$$

Retomando el código del *Ejemplo 2.3* el siguiente código nos permitirá comprobar esto:

Código Ejemplo 3.3

```
var(pob)
varprop<-P*(1-P)
```

La diferencia es mínima, ya que la varianza empleando (2.12) y la varianza empleando "P*(1-P)" son 0.2502142 y 0.249964, respectivamente, además dicha diferencia desaparecerá cuando "N" tiende a infinito.

Por lo que podemos concluir que, si se desea calcular el tamaño de la muestra para una proporción, conociendo la varianza poblacional y haciendo que "q = 1-p", la expresión empleada sería la siguiente:

$$n = \frac{pqN}{(N-1)D^2 + pq} \tag{2.13}$$

En el caso de no conocer la varianza poblacional, que es el que realmente se presenta, entonces se puede hacer una estimación a través de un muestreo piloto, calcular la proporción estimada y con esta última la varianza con la siguiente expresión:

$$s^2 = \widehat{p}(1-\widehat{p}) \tag{2.14}$$

Por lo que el tamaño de la muestra en la situación anterior sería el siguiente:

$$n = \frac{\widehat{pq}N}{(N-1)D^2 + \widehat{pq}} \tag{2.15}$$

Pero la ecuación 2.15 presenta el mismo problema que su similar, ya que para el nivel de error y confiabilidad requerido, solo lo cumplirán algunos de ellos como en el ejemplo 2.1, sin embargo existe una alternativa a esto y es emplear la varianza más alta posible para proporciones, con lo que se estará por lo menos cumpliendo con el nivel de error y confiabilidad propuesto. Esto se logra haciendo a "p" igual a 0.5, renombremos a "p" con "p*=0.5" por lo que la varianza será:

$$\sigma^{2*} = p^* (1-p^*) = 0{,}25 \tag{2.16}$$

De tal manera, que ahora la expresión para calcular el tamaño de la muestra será:

$$n = \frac{0{,}25N}{(N-1)D^2 + 0{,}25} \tag{2.17}$$

Podemos ejemplificar todo lo anterior de la siguiente manera:

Ejemplo **2.4** *Suponga que existe una población de 10,000 habitantes, de la cual se está interesado en saber qué proporción de personas están a favor de una reforma, suponga también que se conoce la varianza, la cual es de 0.21, por lo tanto, el tamaño de la muestra será (considere un nivel de significancia del 5 % y un error del 10 %):*

Código Ejemplo 3.4

```
N<-10000
B<-.1
alfa<-.05
var<-0.21
D<-B/qnorm(alfa/2,0,1)
n<-var*N/((N-1)*D^2+var)
```

Lo que nos arroja un tamaño de muestra de 81, redondeando hacia arriba.

Ahora suponga que no conocemos la varianza, entonces lo que se recomienda hacer es tomar una muestra piloto, digamos de 30 y estimar a "p", para posteriormente calcular el tamaño de la muestra.

Código Ejemplo 3.4

```
set. seed(186)
pilot<-sample(c(0,1),30,replace=TRUE, prob=c(0.3,0.7)) #Muestra piloto
pest<-sum(pilot[pilot==1])/30 #Proporción estimada
```

La estimación de "p" fue aproximadamente igual a 0.66667, lo que arroja una varianza de 0.2222, que es diferente a la varianza poblacional de 0.21, en este caso es más grande, por lo que se espera que el tamaño de la muestra también lo sea.

Código Ejemplo 3.4

```
N<-10000
B<-.1
alfa<-.05
var<-pest*(1-pest)
D<-B/qnorm(alfa/2,0,1)
n<-var*N/((N-1)*D^2+var)
```

El tamaño de la muestra en este caso es de 85. Como es de esperarse, en este caso se disminuyó el error o se aumentó la confiabilidad, sin embargo pudo haber ocurrido lo contrario, por lo que se procederá a calcular el tamaño de la muestra más grande con el nivel de error y confiabilidad propuestos.

Código Ejemplo 3.4

```
alfa<-.05
var<-0.25
D<-B/qnorm(alfa/2,0,1)
n<-var*N/((N-1)*D^2+var)
```

La cual nos arroja un tamaño de muestra de 96.

La expresión anterior puede ser programada dentro de una función, la cual dependerá de “B”, “alfa”, la “varianza” y “N”, a continuación se presenta el código:

Código Ejemplo 3.4

```
nm<-function(B,alfa,var,N){
D<-abs(B/qnorm(alfa/2,0,1))
n<-var*N/((N-1)*D^2+var)
ceiling(n)
}
```

Después de haber revisado el cálculo del tamaño de la muestra empleando elementos de muestreo aleatorio irrestricto, se procede al tema de elemento de muestreo estratificado.

2.3. Muestreo Estratificado

A diferencia del muestreo aleatorio irrestricto, en el que cada muestra tiene la misma probabilidad de ocurrir, y cada elemento de la muestra, la misma probabilidad de ser seleccionado, el muestreo aleatorio estratificado divide a la población en estudio en estratos, dando una probabilidad diferente a cada estrato para seleccionar elementos que pertenecen a cada uno de ellos, pero se realiza de manera aleatoria. Para resaltar la importancia del muestreo aleatorio estratificado, pongamos el siguiente ejemplo, como usted debe saber, a pesar de los esfuerzos que nuestra Facultad de Ingeniería realiza por incorporar más mujeres en su matrícula, existe una diferencia significativa respecto a la cantidad de hombres inscritos, dicha matrícula está conformada aproximadamente por el 75 % de hombres y 25 % de mujeres, aproximadamente. Suponga que su población objetivo son estos alumnos y alumnas de la Facultad de Ingeniería y conduce un estudio para estimar el promedio alcanzado en los cursos que ha tomado la población, para un nivel de significancia y error dados. Si consideramos que no existen diferencias entre los hombres y mujeres, respecto al promedio que alcanzan durante su carrera, entonces el muestreo aleatorio irrestricto será suficiente para estimar un tamaño de muestra. Ahora suponga que por alguna razón sabe que el promedio alcanzado por las mujeres difiere del promedio alcanzado por los hombres en media y varianza, digamos que las mujeres tienen un promedio mayor y una varianza menor en el mismo, ¿acaso lo duda? Y emplea el muestreo aleatorio irrestricto, en este caso y dada la cantidad de hombres y mujeres que están inscritos en algún curso impartido en la Facultad de Ingeniería, puede provocar que en alguna muestra no se seleccionen a mujeres, lo que provocaría un sesgo en la calidad de la información, haciéndola no fidedigna. En este caso el muestreo estratificado juega un papel fundamental al dejar ciertos números tanto para hombres como para mujeres. Tal vez alguien pueda suponer que no existen diferencias entre el promedios alcanzado por las mujeres y los hombres, lo que si suena bastante lógico, es que el promedio de un alumno cambia a medida que va pasando cada semestre, por ejemplo, no es lo mismo preguntarle su promedio alcanzado hasta el momento a un alumna(o) de primer semestre de la carrera que a alguien de noveno semestre o quizás de quinto semestre, ni hablar del tercer semestre, alguien que curso las ciencias básicas podrá dar fe de esto. En este caso, la mayoría de los alumnos se encuentran concentrados en las materias del bloque de ciencias básicas, lo que hace suponer que, si se toma un muestreo aleatorio irrestricto, es muy probable que en alguna muestra no aparezcan personas del noveno semestre, lo que provocaría nuevamente un sesgo involuntario y como consecuencia una falta de confianza en la calidad de los resultados. En este último caso, se puede emplear un muestreo estratificado, en el que cada estrato se distinguiría por el avance de créditos que el alumno lleve. Si las razones anteriores no son suficientes para emplear el muestreo aleatorio estratificado, considere el hecho de que existen poblaciones, que para acceder a dichos estratos son muy costosas o inclusive peligrosas, por otro lado, la varianza de los estratos puede cambiar entre ellos, lo que implicaría un tamaño de muestra menor para estratos con varianza pequeña, lo que hace al muestreo aleatorio estratificado una opción bastante atractiva para este tipo de situaciones. En otras palabras, se puede producir un límite más pequeño para el error de estimación, el costo puede ser reducido y se pueden obtener estimaciones de los parámetros por estratos (Scheaffer, et al).

A manera de ejemplo de lo anterior, revisemos el siguiente código en R

Ejemplo **2.5**

Código Ejemplo 3.5

```
set.seed(200)
H<-rep(0, 7500) #Hombres de la población, indicados con cero
M<-rep(1, 2500) #mujeres de la población, indicadas con uno
P<-c(H, M) #poblacion con 7500 hombres y 2500 mujeres
muestra<-sample(P, 100, replace=FALSE) #muestra de tamaño 100 de la
poblacion
mean(muestra) #Proporción de mujeres encuestadas
hist(muestra,nclass=2)
```

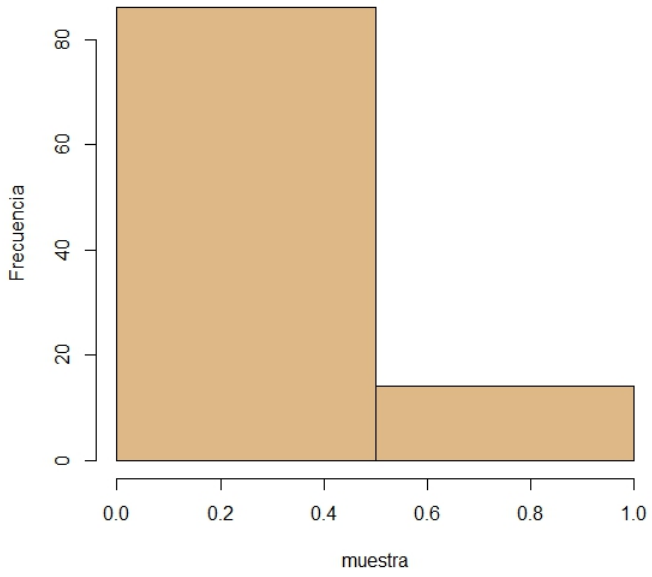


Figura 2.3: Histograma de la muestra {1,0}

En el ejemplo anterior, es claro que, si la proporción de mujeres es del 25 %, en este caso al usar muestreo aleatorio irrestricto, apenas se muestreo el 14 % y es probable que para una muestra se puedan llegar a tener menos.

Para poder empelar el muestreo estratificado, lo primero que hay que hacer es seleccionar y numerar a los estratos de manera adecuada, por ejemplo:

$$\begin{aligned} L &= Número \quad de \quad estratos. \\ N_i &= Número \quad de \quad elementos \quad muestrales \quad en \quad el \quad estrato \quad i. \\ N &= Número \quad de \quad elementos \quad muestrales \quad en \quad la \quad población. \end{aligned} \tag{2.18}$$

Entonces el tamaño de la población estará dado por:

$$N = N_1 + N_2 + ... + N_L \tag{2.19}$$

Ahora bien, para estimar la media poblacional a través de los estratos, se empleará la siguiente ecuación:

$$\bar{Y_{st}} = \frac{1}{N}\left(N_1\bar{Y}_1 + N_2\bar{Y}_2 + ... + N_L\bar{Y}_L\right) = \frac{1}{N}\sum_{i=1}^{L} N_i\bar{Y}_i$$

Donde

$$\bar{Y_{st}} \quad es \quad la \quad estimacion \quad de \quad la \quad media \quad poblacional \quad a \quad traves \quad del \quad muestreo \quad estratificado \tag{2.20}$$

La varianza estimada en este caso sería

$$\hat{V}(\bar{Y_{st}}) = \frac{1}{N^2}\left(N_1^2\hat{V}(\bar{Y}_1) + N_2^2\hat{V}(\bar{Y}_2) + ... + N_L^2\hat{V}(\bar{Y}_L)\right) \tag{2.21}$$

Empleando el tamaño de muestra con población finita como en el caso de muestreo irrestricto aleatorio, cuando no se conoce la varianza de la población del estrato, tenemos

$$\hat{V}(\bar{Y_{st}}) = \frac{1}{N^2}\left(N_1^2\left(\frac{N_1 - n_1}{N_1}\right)\left(\frac{s_1^2}{n_1}\right) + N_2^2\left(\frac{N_2 - n_2}{N_2}\right)\left(\frac{s_2^2}{n_2}\right) + ... + N_L^2\left(\frac{N_L - n_L}{N_L}\right)\left(\frac{s_L^2}{n_L}\right)\right)$$

$$\hat{V}(\bar{Y_{st}}) = \frac{1}{N^2}\sum_{i=1}^{L} N_i^2\left(\frac{N_i - n_i}{N_i}\right)\left(\frac{s_i^2}{n_i}\right) \tag{2.22}$$

Si tomamos el límite para el error de estimación, nuevamente empleando el Teorema del Límite Central

$$Z_{\frac{\alpha}{2}}\sqrt{\hat{V}(\bar{Y_{st}})} = Z_{\frac{\alpha}{2}}\sqrt{\frac{1}{N^2}\sum_{i=1}^{L} N_i^2\left(\frac{N_i - n_i}{N_i}\right)\left(\frac{s_i^2}{n_i}\right)} \tag{2.23}$$

Con la expresión anterior, y los antecedentes vistos en el muestreo irrestricto aleatorio, se procede a seleccionar el tamaño de muestra para estimar la media poblacional. Sabemos que la estimación de la media poblacional en un muestreo irrestricto aleatorio debe estar en "B" unidades dentro de la media poblacional, por lo tanto

$$Z_{\frac{\alpha}{2}}\sqrt{\hat{V}(\bar{Y_{st}})} = B \tag{2.24}$$

Por lo tanto

$$\hat{V}(\bar{Y_{st}}) = \frac{B^2}{{z_{\frac{\alpha}{2}}}^2} \tag{2.25}$$

$$D = \frac{B^2}{{z_{\frac{\alpha}{2}}}^2} \tag{2.26}$$

Una aproximación del tamaño de muestra para estimar a la media poblacional con un límite para el error "B", puede obtenerse a través de la siguiente expresión

$$n = \frac{\sum_{i=1}^{L} \frac{N_i^2 \sigma_i^2}{\omega_i}}{N^2 D + \sum_{i=1}^{L} N_i \sigma_i^2} \tag{2.27}$$

En el caso de no contar con los datos poblacionales de la varianza, de la ecuación 2.27 se pueden sustituir por las varianzas de las muestras, ya sea obtenidas de un muestreo previo o con un muestreo piloto para cada estrato.

Ejemplo **2.6** *Cuál sería el tamaño de la muestra de una población que cuenta con cuatro estratos de los cuales se sabe que las varianzas estimadas en un estudio anterior son 35, 115, 120 y 200 para cada uno de los estratos, además se pide un nivel de significancia del 5 % y un límite para el error de estimación de dos "B=2", así mismo las fracciones asignadas son w1=w2=w3=w4=1/4. Por otro lado, N1=120, N2=60, N3=90 y N4=30.*

Código Ejemplo 3.6

```
B<-2
alfa<-.05
vari<-c(35, 115, 120, 200)
Ni<-c(120,60,90,30)
wi<-c(1/4,1/4,1/4,1/4)
D<-abs(B/qnorm(alfa/2,0,1))
nmi<-sum(Ni^2*vari/wi)/(sum(Ni)^2*D^2+sum(Ni*vari))
```

Por lo tanto se deben tomar 69 muestras y como a cada estrato se le asignó una proporción de wi=1/4, entonces ni=69/4 aproximadamente igual a 18.

La siguiente es una asignación aproximada para minimizar V(yst) para un costo fijo o viceversa, que minimiza el costo para un valor fijo de V(yst).

$$n_i = n(\frac{N_i\sigma_i/\sqrt{C_i}}{\sum_{k=1}^{L} N_k\sigma_k/\sqrt{C_k}}) \tag{2.28}$$

Dodnde

$N_i : Tamao \quad del \quad i-simo \quad estrato.$

$C_i : Consta \quad de \quad una \quad observacin \quad deli-simo \quad estrato.$

$\sigma_i{}^2 : Varianza \quad poblacional \quad del \quad i-simo \quad estrato.$

Sustituyendo la ecuación 2.28 $\frac{n_i}{n}$, en la ecuación 27 por wi.

$$n = \frac{\sum_{i=1}^{L} N_i{}^2\sigma_i{}^2/\varpi_i}{N^2D + \sum_{i=1}^{L} N_i\sigma_i{}^2} ...A)$$

$$\frac{n_i}{n} = \frac{N_i\sigma_i/\sqrt{C_i}}{\sum_{k=1}^{L} N_k\sigma_k/\sqrt{C_k}} ...(B)$$

Haciendo ϖ_i de A igual a $\frac{n_i}{n}$ de B y sustituyendo en A.

$$n = \frac{\frac{\sum_{i=1}^{L} N_i{}^2\sigma_i^2}{\frac{N_i\sigma_i/\sqrt{C_i}}{\sum_{k=1}^{L} N_k\sigma_k/\sqrt{C_k}}}}{N^2D + \sum_{i=1}^{L} N_i\sigma_i{}^2}$$

$$n = \frac{(\sum_{i=1}^{L} N_i{}^2\sigma_i^2)\frac{\sum_{k=1}^{L} N_k\sigma_k/\sqrt{C_k}}{N_i\sigma_i/\sqrt{C_i}}}{N^2D + \sum_{i=1}^{L} N_i\sigma_i{}^2}$$

$$n = \frac{(\sum_{i=1}^{L} \frac{N_i{}^2\sigma^2}{N_i\sigma_i/\sqrt{C_i}})(\sum_{k=1}^{L} N_k\sigma_k/\sqrt{C_k})}{N^2D + \sum_{i=1}^{L} N_i\sigma_i{}^2}$$

$$n = \frac{(\sum_{i=1}^{L} N_i\sigma_i/\sqrt{C_i})(\sum_{k=1}^{L} N_k\sigma_k/\sqrt{C_k})}{N^2D + \sum_{i=1}^{L} N_i\sigma_i{}^2} \tag{2.29}$$

Ejemplo **2.7** *Cuál sería el tamaño de la muestra de una población que cuenta con cuatro estratos de los cuales se sabe que las varianzas estimadas en un estudio anterior son 35, 115, 120 y 200 para cada uno de los estratos, además se pide un nivel de significancia del 5 % y un límite para el error de estimación de dos "B=2", los costos asociados son: c1=5, c2=1, c3=2, c4=4. Finalmente, N1=120, N2=60, N3=90 y N4=30.*

Código Ejemplo 3.7 Estratificado con costos

```
B<-2
alfa<-.05
vari<-c(35, 115, 120, 200)
ci<-c(5, 1, 2, 4)
Ni<-c(120, 60, 90, 30)
D<-(B/qnorm(alfa/2, 0, 1))^2
nmi<-sum(Ni*vari^.5*ci^.5)*sum(Ni*vari^.5/ci^.5)/(sum(Ni)^2*D+sum(Ni*vari))
```

El total de la muestra es de 69 elementos, los cuales se distribuirán de la siguiente manera, empleando la ecuación 2.28.

Código Ejemplo 3.7 Distribución de n en los estratos

```
vari<-c(35, 115, 120, 200)
ci<-c(5,1,2,4)
Ni<-c(120,60,90,30)
niest<-nmi*(Ni*vari^.5/ci^.5)/sum(Ni*vari^.5/ci^.5)
ceiling(niest)
```

Los valores correspondientes a $n_1 = 12$, $n_2 = 24$, $n_3 = 26$ y $n_4 = 8$.

Si los costos para muestreo de los estratos fueran iguales, entonces las ecuaciones serían:

$$n_i = n(\frac{N_i\sigma_i}{\sum_{i=1}^{L} N_i\sigma_i}) \tag{2.30}$$

$$n = \frac{\left(\sum_{i=1}^{L} N_i\sigma_i\right)^2}{N^2D + \sum_{i=1}^{L} N_i{\sigma_i}^2} \tag{2.31}$$

Ejemplo **2.8** *Del ejemplo anterior, suponga que todas las muestras en cada estrato cuestan lo mismo, determine el tamaño total de la muestra y por estrato.*

Código Ejemplo 3.8 Estarificado con costos iguales

```
B<-2
alfa<-.05
vari<-c(35, 115, 120, 200)
Ni<-c(120,60,90,30)
D<-(B/qnorm(alfa/2,0,1))^2
nsc<-sum(Ni*vari^.5)^2/(sum(Ni)^2*D+sum(Ni*vari))
nisc<-nsc*Ni*vari^.5/sum(Ni*vari^.5)
```

Tamaño de muestra para proporciones cuando se emplea muestreo estratificado.

En el caso de proporciones la varianza puede ser sustituido por (pq), $(\hat{p}\hat{q})$, (p^*q^*) en la ecuación 3,27.

$$n = \frac{\sum_{i=1}^{L} {N_i}^2 p_i q_i / \varpi_i}{N^2 D + \sum_{i=1}^{L} N_i^2 p_i q_i} \tag{2.32}$$

En el caso de que se cuente con los costos de muestreo de elementos por estrato, sustituyendo la varianza nuevamente con "$p * q$".

$$n_i = n(\frac{N_i \sqrt{p_i q_i / c_i}}{\sum_{k=1}^{L} N_k \sqrt{p_k q_k / c_k}})$$

Capítulo 3

Regresión Lineal

3.1. Introducción

En estos modelos se relaciona a una variable dependiente llamada "y" con una variable independiente llamada "x". Revise el siguiente modelo de la física clásica, que relaciona a la distancia recorrida de una partícula con el tiempo transcurrido a una velocidad constante.

$$d = d_0 + vt \tag{3.1}$$

Donde :

d= distancia.

d_0= distancia inicial.

v= velocidad.

t=tiempo.

Ejemplo **3.1** *Suponga que* $d_0 = 0$*, v= 1 m/s , entonces para los tiempos t=0, 10, 20, 30, 40 y 50 segundos. ¿Qué valores de "d.ºbtendría?*

Del ejemplo 3.1 podemos concluir que es un modelo lineal, el cual tiene ordenada al origen de cero y pendiente de uno. Ahora suponga que le piden demostrar este modelo en un laboratorio, para lo cual se le provee de un vehículo de control remoto que se acelera hasta alcanzar una velocidad de 1 m/s, así como de un cronómetro, y un flexómetro con el que puede medir y registrar las distancias cada 10, 20, 30, 40 y 50 segundos; el experimento lo realizará 10 veces y registrara sus resultados.

Como es de esperarse, debido a errores de medición, no es posible registrar cada que transcurran 10 segundos la distancia esperada con precisión, estos errores son variables aleatorias, por lo tanto el modelo sería:

$$d = d_0 + vt + \varepsilon_i \tag{3.2}$$

Donde :

ε_i = Errores (variable aleatoria)

Es claro que con equipo sofisticado y controlado las variables como temperatura y humedad se puede hacer tender a cero a la v. a. (ε), y con esto comprobar que la ecuación (3.1) prevalecerá.

Ahora veamos el siguiente caso, suponga que tenemos los datos de la publicidad emitida por una empresa medida en miles [U.M] y las ventas logradas en miles [U] para una empresa dada como se muestra en la tabla 2.

En este caso, tanto las ventas como la inversión en publicidad son registradas con exactitud, por lo que el modulo se puede representar de la siguiente manera:

$$Y_i = \beta_1 + \beta_2 X_i + \varepsilon_i \tag{3.3}$$

Donde:

Y_i= ventas en [mU].
X_i= publicidad en [mUM].
ε_i = Error (v.a).
β_1 = Origen en [mU].
β_2= Pendiente en [mU/mUM].

En la ecuación (3.3) es claro que el término de error no se puede hacer tender a cero, solo con mejorar la precisión al registrar los valores obtenidos a un nivel de publicidad dado, ya que como se aclaró anteriormente, estos fueron tomados con precisión, el término de error (ε) es inherente al modelo y no puede ser eliminado.

Por lo tanto la ecuación (3.2) pertenece a modelos deterministas donde el término de error (ε) puede tender a cero al mejorar la precisión de las mediciones y la ecuación (3.3) pertenece a modelos estocásticos, donde el término de error (ε) no puede desaparecer, porque no depende solamente de la precisión de las mediciones; son estos modelos estocásticos los que nos interesa estudiar.

3.2. Modelo de Regresión Simple

El primer modelo que se tomará en consideración es el modelo de regresión lineal simple, el cual puede ser representada como sigue:

$$Y_i = \beta_1 + \beta_2 X_i + \varepsilon_i \tag{3.4}$$

Donde :
Y_i = Variable dependiente o explicada.
X_i = Variable independiente o explicativa.
β_0 = Ordenada al origen.
β_1 = Pendiente o razón de cambio.
ε_i = Es una v.a.

En esta clase de modelos lo que se pretende es descomponer a "Y_i" en dos partes, una determinista representada por el modelo lineal $\beta_1 + \beta_2 X_i$ y una parte estocástica representada por ε_i.

Para lograr lo anterior se debe partir de una serie de supuestos los cuales se explicarán a través de un ejemplo.

Ejemplo **3.2** *Continuando con nuestro ejemplo de publicidad y ventas, suponga que contamos con los datos poblacionales, entonces se puede obtener los* $E[Y|X_i]$ *el cual se lee como el valor esperado de "Y" dado una "*X_i*".*

El primer supuesto, es que los $E[Y|X_i]$ están perfectamente alineados, entonces se puede representa con una recta.

$$E[Y|X_i] = \beta_1 + \beta_2 X_i \tag{3.5}$$

Si quisiéramos llegar a un valor Y_i, entonces:

$$Y_i = E[Y|X_i] + \varepsilon_i \tag{3.6}$$

Lo que da pie al segundo supuesto $E[\varepsilon|X_i] = 0$, esto debe ser cierto, si se desea que los $E[Y|X_i]$, estén perfectamente alineados.

En la realidad con lo que contaremos es con un subconjunto o una muestra de la población, en el que los supuestos de linealidad y de que $E[\varepsilon|X_i] = 0$ deben ser demostrados. Para tal efecto se deberán estimar β_1 y β_2 haciendo $\sum \hat{\varepsilon}_i$ sea igual a cero.

Esto es:

$$\hat{\varepsilon}_i = Y_i - (\hat{\beta}_1 + \hat{\beta}_2 X_i)$$

Entonces:

$$\sum \hat{\varepsilon_i} = \sum [Y_i - (\hat{\beta_1} + \hat{\beta_2} X_i)]$$

Donde:
$\sum \hat{\varepsilon_i}$ = Es la suma de los errores estimados.

Se puede demostrar que existe una familia de rectas que representamos como:

$$\hat{Y_i} = \hat{\beta_1}^* + \hat{\beta_2}^* X_i \tag{3.7}$$

Donde :
$\hat{Y_i}$ = Variable dependiente o explicada.
X_i = Variable independiente o explicativa.
$\hat{\beta_0}^*$ = Ordenada al origen.
$\hat{\beta_1}^*$ = Pendiente o razón de cambio.

Que cumplirán con que:

$$\sum \hat{\varepsilon_i} = 0$$

Esto, por supuesto genera un problema, ya que existe una infinidad de $\hat{\beta_1}^*$ y $\hat{\beta_2}^*$ que cumplen con que $\sum \hat{\varepsilon_i} = 0$ para cada combinación, por lo que se hace necesaria poner una restricción más fuerte, la cual puede ser la siguiente

$$min \sum \hat{\varepsilon_i}^2 \tag{3.8}$$

Con esta condición se busca obtener una $\hat{\beta_1}$ y una $\hat{\beta_2}$ que cumple con que $\sum \hat{\varepsilon_i} = 0$ y $\sum \hat{\varepsilon_i}^2$ sea mínimo. Nuevamente:

$$Y_i = \hat{Y_i} + \hat{\varepsilon_i} \tag{3.9}$$

Donde:

$$\hat{Y_i} = \hat{\beta_1} + \hat{\beta_2} X_i$$

Entonces:

$$\hat{\varepsilon_i} = Y_i - \hat{Y_i} \tag{3.10}$$

$$\sum \hat{\varepsilon_i}^2 = [Y_i - \hat{Y_i}]^2 \tag{3.11}$$

Sustituyendo (3.8) en (3.11) tenemos:

$$\sum \hat{\varepsilon_i}^2 = mn \sum \left[Y_i - (\hat{\beta}_1 + \hat{\beta}_2 X_i) \right]^2$$

Por lo tanto por:

$$min \sum \hat{\varepsilon_i}^2 = \sum \left[Y_i - \hat{\beta}_1 + \hat{\beta}_2 X_i \right]^2 \tag{3.12}$$

Para encontrar el mínimo debemos derivar parcialmente (3.12) respecto a $\hat{\beta}_1$ y posteriormente con respecto a $\hat{\beta}_2$.

Entonces:

$$\frac{\delta}{\delta \hat{\beta}_1} \sum \left[Y_i - (\hat{\beta}_1 + \hat{\beta}_2 X_i \right])^2 = 0 ...(A)$$

$$\frac{\delta}{\delta \hat{\beta}_2} \sum \left[Y_i - (\hat{\beta}_1 + \hat{\beta}_2 X_i \right])^2 = 0 ...(B)$$

Derivando con respecto a $\hat{\beta}_1$ a (A)

$$2 \sum \left[Y_i - \hat{\beta}_1 + \hat{\beta}_2 X_i \right] (-1) = 0$$

$$\sum \left[Y_i - (\hat{\beta}_1 + \hat{\beta}_2 X_i) \right] (-1) = 0$$

$$- \sum Y_i + \sum \hat{\beta}_1 + \sum \hat{\beta}_2 X_i = 0$$

$$- \sum Y_i + n\hat{\beta}_1 + \hat{\beta}_2 \sum X_i = 0$$

$$n\hat{\beta}_1 = \sum Y_i - \hat{\beta}_2 \sum X_i$$

$$\hat{\beta}_1 = \frac{\sum Y_i}{n} - \hat{\beta}_2 \frac{\sum X_i}{n} \qquad ... [I]$$

$$\beta_1 = \bar{Y} - \hat{\beta}_2 \bar{X} \tag{3.13}$$

Derivando con respecto a $\hat{\beta}_2$ a (B)

$$2 \sum \left[Y_i - (\hat{\beta}_1 + \hat{\beta}_2 x_i) \right] (-X_i) = 0$$

$$\sum \left[Y_i - (\hat{\beta}_1 + \hat{\beta}_2 x_i) \right] (-X_i) = 0$$

$$-\sum Y_i X_i + \sum \hat{\beta}_1 X_i + \sum \hat{\beta}_2 {X_i}^2 = 0$$

$$-\sum X_i Y_i + \hat{\beta}_1 \sum X_i + \hat{\beta}_2 \sum {X_i}^2 = 0 \qquad \ldots [II]$$

Sustituyendo a I en II

$$-\sum X_i Y_i + \left(\frac{\sum Y_i}{n} - \hat{\beta}_2 \frac{\sum X_i}{n}\right) \sum X_i + \hat{\beta}_2 \sum {X_i}^2 = 0$$

$$-\sum X_i Y_i + \frac{\sum Y_i \sum X_i}{n} - \hat{\beta}_2 \frac{\left(\sum X_i\right)^2}{n} + \hat{\beta}_2 \sum {X_i}^2 = 0$$

$$-\sum X_i Y_i + \frac{\sum Y_i \sum X_i}{n} - \hat{\beta}_2 \left[\frac{\left(\sum X_i\right)^2}{n} - \sum {X_i}^2\right] = 0$$

$$\hat{\beta}_2 \left[\sum {X_i}^2 - \frac{\left(\sum X_i\right)^2}{n}\right] = \sum X_i Y_i - \frac{\sum Y_i \sum X_i}{n}$$

$$\hat{\beta}_2 = \frac{\sum X_i Y_i + \frac{\sum Y_i \sum X_i}{n}}{\left[\sum {X_i}^2 - \frac{\left(\sum X_i\right)^2}{n}\right]} \frac{n}{n}$$

$$\hat{\beta}_2 = \frac{n \sum X_i Y_i - \sum Y_i \sum X_i}{n \sum {X_i}^2 - \left(\sum X_i\right)^2} \tag{3.14}$$

Por tanto las ecuaciones (3.13) y (3.14) nos permiten obtener $\hat{\beta}_1$ y $\hat{\beta}_2$ tal que $\sum \hat{\varepsilon}_i = 0$ y$\sum {\hat{\varepsilon}_i}^2$ sea mínima, por tal motivo se le conoce como mínimos cuadrados ordinarios, otra forma de representar a $\hat{\beta}_2$ es a través de:

$$\hat{\beta}_2 = \frac{\sum \left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sum \left(X_i - \bar{X}\right)^2}$$

o

$$\hat{\beta}_2 = \frac{S_{XY}}{S_{XX}}$$

Donde:

$$S_{XY} = \sum \left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)$$

$$S_{XX} \sum \left(X_i - \bar{X}\right)^2$$

Ejemplo **3.3** *Continuando con el ejemplo.*

Comparando los modelos propuestos en el ejemplo 4 con el obtenido empleando (3.13) y (3.14) vemos que esta última cumple con que $\sum \hat{\varepsilon}_i$ es cero y $\sum \hat{\varepsilon}_i{}^2$ es la más pequeña que podemos encontrar.

Una vez encontrada $\hat{\beta}_1$ y $\hat{\beta}_2$ que estima a β_1 y β_2 poblacionales, se procede a tipificar el término de error ε completamente:

$$E[\varepsilon|X_i] = 0$$

$$Var[\varepsilon|X_i] = \sigma^2$$

En donde σ^2 debe ser finita y constante

$$Var[\varepsilon|X_i] = E[[\varepsilon - E[\varepsilon|X_i]]\,|X_i]^2$$

$$Var[\varepsilon|X_i] = \sigma^2$$

$$Cov\,[\varepsilon_i, \varepsilon_j] = 0 \qquad \forall \quad i \neq j.$$

De forma general a manera de resumen podemos listar los siguientes supuestos del modelo de regresión lineal simple:

Supuesto 1: El modelo es lineal.
Supuesto 2: $E[\varepsilon|X_i] = 0$.
Supuesto 3: $Var[\varepsilon|X_i] = \sigma^2$ el modelo es homocedástico, la varianza no cambia.
Supuesto 4: $Cov\,[\varepsilon_i \varepsilon_j] = 0 \quad coni \neq j$.

Básicamente con estos cuatro supuestos hemos tipificado el comportamiento de la ecuación (3.6). Sin embargo , en la realidad solo contaremos con una muestra y por lo tanto de una estimación de β_1, β_2 y de los errores ε_i. Finalmente se supondrá que los errores se distribuyan de forma normal.

Ejemplo **3.4**

$$\varepsilon_i \sim N\left(0, \sigma^2\right)$$

Entonces, lo que se buscará es que los errores sean una variable aleatoria que se distribuyan de forma normal, con media cero y varianza constante. Si lo anterior es correcto, entonces se puede demostrar que las betas estimadas se distribuyen como una gausiana con media en la beta poblacional y error estándar de las betas estimadas.

Entonces podemos establecer pruebas de hipótesis sobre los parámetros beta poblacionales como sigue:

$$H_0: \quad \beta_i = \beta_i^*$$

$$H_A: \quad \beta_i \neq \beta_i$$

A un nivel de significancia dado y empleado el estadístico t como sigue

$$t = \frac{\hat{\beta}_i - \beta_i}{ee(\hat{\beta}_i)}$$

Otros parámetros a considerar son el coeficiente de correlación y el coeficiente de determinación.

Coeficiente de correlación R y coeficiente de determinación R^2.

Podemos decir que la variación total es igual a la suma de la variación explicada por el modelo y la variación explicada por los errores.

$$STC = SRC + SEC.$$

$$STC = Suma \quad total \quad al \quad cuadrado.$$

$$SRC = Suma \quad de \quad la \quad regresion \quad al \quad cuadrado.$$

$$SEC = Suma \quad de \quad los \quad errores \quad al \; quadcuadrado.$$

$$STC = \sum \left(Y_i - \bar{Y}\right)^2$$

$$SRC = \sum \left(\hat{Y}_i - \bar{Y}\right)^2$$

$$SEC = \sum \left(Y_i - \hat{Y}_i\right)^2$$

Por lo tanto:

$$\sum \left(Y_i - \bar{Y}\right)^2 = \sum \left(\hat{Y}_i - \bar{Y}\right)^2 + \sum \left(Y_i - \hat{Y}_i\right)^2 \tag{3.15}$$

Dividiendo ambos lados de (3.15) entre STC:

$$1 = \frac{\sum \left(\hat{Y}_i - \bar{Y}\right)^2}{\sum \left(Y_i - \bar{Y}\right)^2} + \frac{\sum \left(Y_i - \hat{Y}_i\right)^2}{\sum \left(Y_i - \bar{Y}\right)^2}$$

El coeficiente de correlación queda explicado como:

$$R^2 = \frac{\sum \left(\hat{Y}_i - \bar{Y}\right)^2}{\sum \left(Y_i - \bar{Y}\right)^2} \tag{3.16}$$

Otra forma:

$$R^2 = 1 - \frac{\sum \left(Y_i - \hat{Y}_i\right)^2}{\sum \left(Y_i - \bar{Y}\right)^2} \tag{3.17}$$

Ambas expresiones (3.16) y (3.17) indican que tanto de la variación de "Y_i" es explicada por un cambio en el valor de "X_i" y que tanto se debe al error.

Si se obtiene la raíz cuadrada del coeficiente de determinación se obtiene R que es el coeficiente de correlación y mide la fuerza de la relación lineal entre dos variables en este caso de "Y" con "X" o de "X" con "Y". El coeficiente de determinación se encuentra entre 0 y 1, mientras que el coeficiente de correlación se encuentra entre -1 y 1.

A manera de resumen, se presentarán los pasos y las pruebas que se le deben realizar al modelo lineal simple:

1. Se debe aplicar mínimos cuadrados ordinarios para ajustar el modelo lineal y obtener $\hat{\beta}_1$ y $\hat{\beta}_2$.
2. Se debe comprobar que los errores son aleatorios, empleando el estadístico de Durbín Watson.
3. Se debe comprobar que los errores se comportan de forma normal.
4. Se debe comprobar que los errores son homocedásticos.
5. Se hace la prueba global:

$$H_0: \quad \beta_1 = \beta_2 = 0.$$

$$H_A: \quad Alguna \quad \beta_i \neq 0.$$

6. Se hacen pruebas individuales.

$$H_0: \quad \beta_i = 0, .$$

$$H_A: \beta_i \neq 0.$$

7. Se revisa el coeficiente de determinación ajustado. (Esto se verá en el siguiente tema donde se tratan modelos de regresión múltiple).
8. Se pueden obtener los intervalos de confianza por las betas estimados.
9. Se pueden obtener los intervalos de confianza para los valores $\hat{Y}_i$.

Lo anterior será presentado en un ejemplo integrador en un modelo lineal múltiple.

3.3. Modelo de Regresión Lineal Múltiple

Para estos modelos la especificación es:

$$Y_i = \beta_1 + \beta_2 X_{i2} + \beta_3 X_{i3} + ... + \beta_k X_{ik} + \varepsilon_i \qquad para \quad i = 1, 2, ...n$$

$$Y_1 = \beta_1 + \beta_2 X_{12} + \beta_3 X_{13} + ... + \beta_k X_{1k} + \varepsilon_1$$

$$Y_2 = \beta_1 + \beta_2 X_{22} + \beta_3 X_{23} + ... + \beta_k X_{2k} + \varepsilon_2$$

$$Y_3 = \beta_1 + \beta_2 X_{32} + \beta_3 X_{33} + ... + \beta_k X_{3k} + \varepsilon_3$$

$$. \quad . \quad .$$

$$. \quad . \quad .$$

$$. \quad . \quad .$$

$$Y_n = \beta_1 + \beta_2 X_{n2} + \beta_3 X_{n3} + ... + \beta_k X_{nk} + \varepsilon_n$$

De forma matricial esto se puede presentar como:

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ . \\ . \\ . \\ Y_n \end{bmatrix}; \qquad X = \begin{bmatrix} 1 & x_{12} & x_{13}...x_{1k} \\ 1 & x_{22} & x_{23}...x_{2k} \\ 1 & x_{32} & x_{33}...x_{3k} \\ . & . & . \ . \ . \\ . & . & . \ . \ . \\ . & . & . \ . \ . \\ 1 & x_{n2} & x_{n3}...x_{nk} \end{bmatrix}; \quad \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ . \\ . \\ . \\ \beta_k \end{bmatrix}; \qquad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ . \\ . \\ . \\ \varepsilon_k \end{bmatrix}$$

Por lo tanto el modelo quedará como:

$$Y = X\beta + \varepsilon \tag{3.18}$$

Con:

$$\hat{Y} = X\beta$$

Se estimarán las betas con ayuda de los mínimos cuadrados;

$$Y = X\hat{\beta} + \hat{\varepsilon}$$

Sustituyendo:

$$\hat{Y} = X\hat{\beta}$$

$$Y = \hat{Y} + \hat{\varepsilon}$$

$$\hat{\varepsilon} = Y - \hat{Y}$$

Ahora se minimizará:

$$mn_{\hat{\beta}} \quad \left(\hat{\varepsilon}^{'}\hat{\varepsilon}\right)$$

$$mn_{\hat{\beta}} \left[\left(Y - \hat{Y}\right)^{'}\left(Y - \hat{Y}\right)\right]$$

$$mn_{\hat{\beta}} \left[\left(Y - X\hat{\beta}\right)^{'}\left(Y - X\hat{\beta}\right)\right]$$

$$\left(Y - X\hat{\beta}\right)^{'} = Y' - \hat{\beta}'X'$$

$$mn_{\hat{\beta}} \left[\left(Y' - \hat{\beta}'X'\right)\left(Y - X\hat{\beta}\right)\right]$$

$$mn_{\hat{\beta}} \left[Y'Y - Y'X\hat{\beta} - \hat{\beta}'X'Y + X'\hat{\beta}'X'X\hat{\beta}\right]$$

$$\frac{\delta}{\delta\hat{\beta}} \left[Y'Y - (Y'X\hat{\beta})' - \hat{\beta}'X'Y + \hat{\beta}'X'X\hat{\beta}\right]$$

$$\frac{\delta}{\delta\hat{\beta}} \left[Y'Y - \hat{\beta}'X^{'}Y - \hat{\beta}'X'Y + \hat{\beta}'X'X\hat{\beta}\right]$$

$$\frac{\delta}{\delta\hat{\beta}} \left[Y'Y - 2\hat{\beta}'X'Y + \hat{\beta}'X'X\hat{\beta}\right]$$

$$-2X'Y + 2X'X\hat{\beta} = 0$$

$$(X'X)\,\hat{\beta} = X'Y$$

$$\hat{\beta} = (X' \quad X)^{-1}X'Y$$

3.4. Ejemplo Integrador Modelo 1

El siguiente ejemplo está basado en el libro de Introducción a la Econometría un Enfoque Moderno, el cual consiste en encuestar a un total de 38 alumnos a los cuales se les preguntó su edad, sexo, promedio del bachillerato, promedio actual, avance, si trabaja, si trabaja menos de 4 horas o más de cuatro horas, tiempo dedicado a ir a la escuela y regresar a su casa, si cuenta con una Lap Top, si el padre tiene una carrera universitaria, si la madre tiene una carrera universitaria, número de sesiones que falta a la semana, si tiene novia(o), días a la semana que consume alcohol y si realiza alguna actividad deportiva o extra a sus clases.

```
library(lmtest)

## Warning: package 'lmtest' was built under R version 3.3.2

## Loading required package: zoo

##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric

library(tseries)

## Warning: package 'tseries' was built under R version 3.3.2

datos<-read.csv(file.choose(), header=TRUE)

head(datos)

##   EDAD SEXO PROMBACH PROMACT AVANCE TRABAJA TIEMPO TRANSESC LAPTOP
## 1   21    1     9.10    9.30     85       1      1     2.00      1
## 2   21    0     9.40    9.00     85       1      1     1.00      1
## 3   25    0     9.00    7.50     92       0      0     2.00      1
## 4   25    0     8.02    7.60     96       1      1     3.00      1
## 5   21    0     9.30    8.75     80       0      0     0.83      1
## 6   21    0     9.20    8.64     75       0      0     3.00      1
##   PADRELIC MADRELIC INGRES10000 FALTAS NOVIA.O DIASALCOHOL DEPORTE
## 1        1        1           1      0       1           2       1
## 2        0        0           1      0       1           0       1
## 3        0        0           0      0       1           1       1
## 4        1        1           1      0       1           0       0
## 5        1        1           1      0       0           0       1
## 6        0        0           0      0       1           1       1
```

```
tail(datos)
```

```
##    EDAD SEXO PROMBACH PROMACT AVANCE TRABAJA TIEMPO TRANSESC LAPTOP
## 33   22    1     8.99    8.87     86       1    1.0      1.0      1
## 34   26    1     7.80    7.70     78       1    1.0      1.6      1
## 35   18    1     9.00    8.62     64       0    0.6      1.0      0
## 36   29    0     8.10    7.80     73       1    1.0      4.0      1
## 37   21    1     8.10    9.00     85       1    1.0      2.0      1
## 38   25    1     8.96    8.76     80       1    1.0      4.0      0
##    PADRELIC MADRELIC INGRES10000 FALTAS NOVIA.O DIASALCOHOL DEPORTE
## 33        1        1           1      0       1           0       0
## 34        0        0           0      0       0           0       0
## 35        0        0           0      0       0           0       1
## 36        0        0           1      2       0           2       0
## 37        1        1           1      0       0           2       1
## 38        0        0           0      0       1           3       1
```

```
attach(datos)
```

```
names(datos)
```

```
##  [1] "EDAD"        "SEXO"        "PROMBACH"    "PROMACT"     "AVANCE"
##  [6] "TRABAJA"     "TIEMPO"      "TRANSESC"    "LAPTOP"
"PADRELIC"
## [11] "MADRELIC"    "INGRES10000" "FALTAS"      "NOVIA.O"
"DIASALCOHOL"
## [16] "DEPORTE"
```

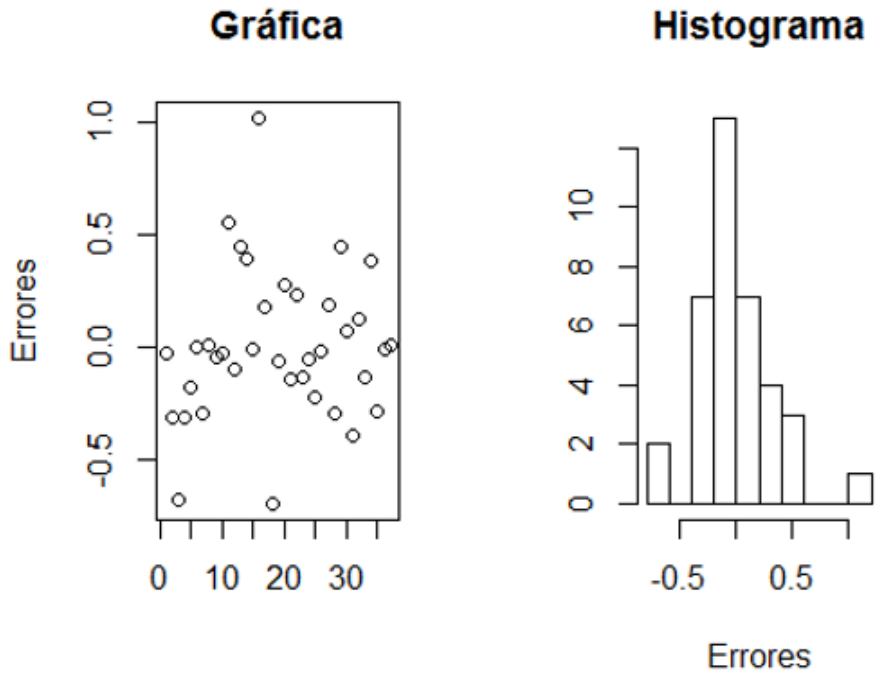
```
mla<-
lm(PROMACT~EDAD+SEXO+PROMBACH+AVANCE+TRABAJA+TIEMPO+TRANSESC+LAPTOP+PADRE
LIC+MADRELIC+INGRES10000+FALTAS+NOVIA.O+DIASALCOHOL+DEPORTE)
```

Se presentan cada uno de los coeficientes de las variables empleadas:

```
mla

##
## Call:
## lm(formula = PROMACT ~ EDAD + SEXO + PROMBACH + AVANCE + TRABAJA +
##     TIEMPO + TRANSESC + LAPTOP + PADRELIC + MADRELIC + INGRES10000 +
##     FALTAS + NOVIA.O + DIASALCOHOL + DEPORTE)
##
## Coefficients:
## (Intercept)         EDAD         SEXO     PROMBACH       AVANCE
##   3.872e+00   -6.431e-02    2.089e-01    6.177e-01   -1.414e-05
##     TRABAJA       TIEMPO     TRANSESC       LAPTOP     PADRELIC
##   8.020e-01   -6.176e-01    8.262e-02    2.876e-01   -1.991e-02
##    MADRELIC  INGRES10000       FALTAS      NOVIA.O  DIASALCOHOL
##  -4.023e-01    6.930e-01   -3.743e-01   -3.001e-01    1.627e-01
##     DEPORTE
##   3.675e-02
```

A continuación se extraen los errores del modelo ajustado y se grafican:

```
erra<-residuals(mla)

par(mfrow=c(1,2))
plot(erra, main="Gráfica", ylab="Errores", xlab="")
hist(erra, main="Histograma", ylab="", xlab="Errores")
```




```
dev.off()

## null device
##           1
```

Durbin-Watson

Se realiza la prueba de Durbin-Watson:

$H_{0:}La\ \ autocorrelacin\ \ es\ \ igual\ \ a\ \ cero.$

$H_A : La\ \ autocorrelacin\ \ es\ \ diferente\ \ de\ \ cero.$

```
dwtest(mla)

##
##  Durbin-Watson test
##
## data:  mla
## DW = 1.9397, p-value = 0.3429
## alternative hypothesis: true autocorrelation is greater than 0
```

A un nivel de significncia $\alpha = 0{,}05$, se concluye que no existe suficiente evidencia para rechazar la hipótesis nula, ya que el valor $P = 0{,}3429$ es mayor que α.

Jarque-Bera

A continuación se realiza la prueba de Jarque-Bera, para probar la normalidad de los errores

$H_{0:}Los\ \ errores\ \ se\ \ distribuyen\ \ de\ \ forma\ \ normal.$

$H_{A:}Los\ \ errores\ \ no\ \ se\ \ distribuyen\ \ de\ \ forma\ \ normal.$

```
jarque.bera.test(erra)

##
##  Jarque Bera Test
##
## data:  erra
## X-squared = 4.5584, df = 2, p-value = 0.1024
```

Dado que el valor de $P = 0{,}1024$ es mayor que el valor de $\frac{\alpha}{2} = 0{,}025$, se concluye que no existe evidencia para rechazar H_0.

Breusch-Pagan

El siguiente paso, es hacer la prueba de homocedasticidad, empelando el estadístico Breusch-Pagan.

$H_{0:}$: *Los errores son homocedsticos.*

$H_{A:}$: *Los errores no son homocedsticos.*

```
bptest(mla)

##
##  studentized Breusch-Pagan test
##
## data:  mla
## BP = 17.45, df = 15, p-value = 0.2927
```

Dado que el valor $P = 0{,}2927$ es mayor que el nivel de $\frac{\alpha}{2} = 0{,}025$, se concluye que los errores son homocedásticos.

Una vez hecha la prueba de los errores, se procede a verificar qué coeficientes son estadísticamente significativos, en otras palabras se realiza la siguiente prueba de hipótesis para cada uno de los coeficientes estimados:

$$H_{0:} : \beta_i = \beta^*$$

$$H_{A:} : \beta_i \neq \beta^*$$

A un nivel de significancia $\alpha = ,05$, se puede apreciar que los coeficientes que son estadísticamente significativos fueron:

Promedio Bachillerato (PROMBACH),

Ingreso (INGRES10000),

Cantidad de días a la semana que ingiere alcohol (DIASALCOHOL),

Si se considera un nivel de significancia $\alpha = 0{,}10$, entonces también entraría:

Número sesiones que falta a la semana (FALTAS).

Lo que significa que en este estudio, el resto de las variables no son estdísticamente significativas, como el genero, la edad, si el padre o la madre tiene carrera universitaria, si tiene o no novio(a), el avance, si cuentan o no con Lap Top, si trabajan o el tiempo que emplean para trasladarse de su casa a la escuela y regresar, así como si realizan alguna actividad adicional como deportes.

```
summary(mla)
## Call:
## lm(formula = PROMACT ~ EDAD + SEXO + PROMBACH + AVANCE + TRABAJA +
##     TIEMPO + TRANSESC + LAPTOP + PADRELIC + MADRELIC + INGRES10000 +
##     FALTAS + NOVIA.O + DIASALCOHOL + DEPORTE)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.69266 -0.17827 -0.01994  0.17736  1.02149
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.872e+00  2.622e+00   1.476  0.15466
## EDAD        -6.431e-02  5.576e-02  -1.153  0.26174
## SEXO         2.089e-01  2.191e-01   0.953  0.35136
## PROMBACH     6.177e-01  2.081e-01   2.969  0.00733 **
## AVANCE      -1.414e-05  1.123e-02  -0.001  0.99901
## TRABAJA      8.020e-01  4.698e-01   1.707  0.10255
## TIEMPO      -6.176e-01  4.497e-01  -1.373  0.18410
## TRANSESC     8.262e-02  9.322e-02   0.886  0.38550
## LAPTOP       2.876e-01  2.243e-01   1.282  0.21368
## PADRELIC    -1.991e-02  3.130e-01  -0.064  0.94987
## MADRELIC    -4.023e-01  3.780e-01  -1.064  0.29923
## INGRES10000  6.930e-01  3.149e-01   2.201  0.03908 *
## FALTAS      -3.743e-01  1.840e-01  -2.034  0.05478 .
## NOVIA.O     -3.001e-01  2.228e-01  -1.347  0.19244
## DIASALCOHOL  1.627e-01  6.820e-02   2.386  0.02650 *
## DEPORTE      3.675e-02  2.130e-01   0.173  0.86468
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4348 on 21 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.7146, Adjusted R-squared:  0.5107
## F-statistic: 3.505 on 15 and 21 DF,  p-value: 0.004379
```

3.5. Ejemplo Integrador Modelo 2

Se correrá un nuevo modelo considerando sólo las variables significativas a un nivel $\alpha = 0{,}05$.

```
mlb<-lm(PROMACT~PROMBACH+INGRES10000+DIASALCOHOL)
```

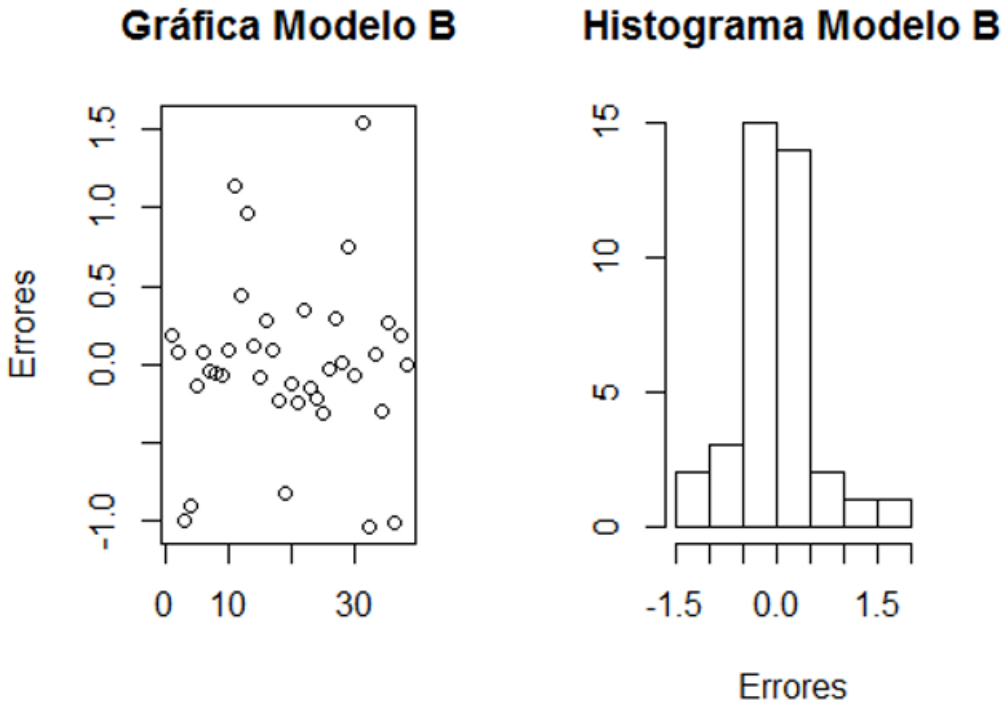
#Los coeficientes asociados a este modelo son

```
mlb
```

```
##
## Call:
## lm(formula = PROMACT ~ PROMBACH + INGRES10000 + DIASALCOHOL)
##
## Coefficients:
## (Intercept)     PROMBACH  INGRES10000  DIASALCOHOL
##      5.6879       0.2966       0.4474       0.1420
```

#A continuación se extraen los errores del modelo ajustado y se grafican nuevamente

```
errb<-residuals(mlb)

par(mfrow=c(1,2))
plot(errb, main="Gráfica Modelo B", ylab="Errores", xlab="")
hist(errb, main="Histograma Modelo B", ylab="", xlab="Errores")
```




```
dev.off()

## null device
##           1
```

Durbin-Watson

Se realiza la prueba de Durbin-Watson:

```
dwtest(mlb)

##
##  Durbin-Watson test
##
## data:  mlb
## DW = 1.999, p-value = 0.491
## alternative hypothesis: true autocorrelation is greater than 0
```

Nuevamente no hay suficiente evidencia para rechazar la hipótesis nula de que no existe auto correlación entre los errores, ya que el valor $P = 0{,}491$ es mayor que el nivel de significancia $\alpha = 0{,}05$

Jarque-Bera

Aplicando la prueba de normalidad de Jarque-Bera:

```
jarque.bera.test(errb)

##
##  Jarque Bera Test
##
## data:  errb
## X-squared = 2.9742, df = 2, p-value = 0.226
```

En este caso el valor de $P = 0{,}226$ es mayor al nivel de significancia $\alpha = 0{,}05$, por lo que no existe suficiente evidencia para rechazar normalidad.

Bresuch-Pagan

Finalmente se hará la prueba de homocedasticidad.

```
bptest(mlb)

##
##  studentized Breusch-Pagan test
##
## data:  mlb
## BP = 3.4939, df = 3, p-value = 0.3216
```

En este caso no existe suficiente evidencia para rechazar la hipótesis de homocedasticidad en los errores, ya que el valor de $P = 0{,}3216$ es mayor que el nivel de significancia propuesto $\alpha = 0{,}05$.

En estas notas no se tratan los temas de qué pasa cuando los errores no son normales, están correlacionados o son heterocedásticos

```
summary(mlb)

##
## Call:
## lm(formula = PROMACT ~ PROMBACH + INGRES10000 + DIASALCOHOL)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.04214 -0.20103 -0.01903  0.18065  1.54638
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.68788    1.41777   4.012 0.000313 ***
## PROMBACH     0.29662    0.15829   1.874 0.069566 .
## INGRES10000  0.44743    0.18938   2.363 0.024012 *
## DIASALCOHOL  0.14200    0.07059   2.012 0.052227 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5666 on 34 degrees of freedom
## Multiple R-squared:  0.2914, Adjusted R-squared:  0.2289
## F-statistic: 4.661 on 3 and 34 DF,  p-value: 0.007817
```

Con el nivel de significancia propuesto sólo el coeficiente de ingreso es estadísticamente significativo, sin embargo al nivel de significancia $\alpha = 0{,}10$, los coeficientes de promedio de bachillerato y días que se consume alcohol son estadísticamente significativos, por lo que se procede al análisis.

$$PROMACT = 5{,}68788 + 0{,}29662 * PROMBACH + 0{,}44743 * INGRESO10000 + 0{,}14200 * DIASALCOHOL.$$

Finalmente se obtiene el valor de los intervalos de confianza para cada uno de los coeficientes estimados

```
confint(mlb, level = 0.90)

##                     5 %      95 %
## (Intercept) 3.29053317 8.0852214
## PROMBACH    0.02895663 0.5642743
## INGRES10000 0.12719686 0.7676644
## DIASALCOHOL 0.02264439 0.2613547
```

La interpretación que se le puede dar a estos intervalos de confianza, por ejemplo, para el promedio en bachillerato, existe un 90 % de probabilidad de que el coeficiente poblacional, real, pero desconocido, se encuentre entre 0.029 y 0.56. Para el caso del ingreso, con la misma probabilidad, el coeficiente se encontrará entre 0.13 y 0.77. Finalmente para los días que se consume alcohol, con la misma probabilidad, los valores están entre 0.02 y 0.26.

Conclusión

Con el modelo propuesto, podemos concluir que el promedio adquirido en el bachillerato es un factor que puede ser empleado para determinar el promedio actual en la carrera ya que al aumentar en una unidad el promedio en el bachillerato, mejoraría el promedio actual en 0.3, otro factor es el ingreso familiar, ya que al tener un ingreso promedio mensual familiar por arriba de 10,000.00 MNX mejora el promedio de los estudiantes en aproximadamente 0.45, la última variable empleada parece no ser consistente, ya que al aumentar en un día el número de días en el que se consume alcohol, mejora también el promedio en 0.14, aunque el coeficiente es el menor de los tres, sin embargo la pregunta que se realizó, no especifica la cantidad de alcohol ingerido, habrá que estimar esta variable en función de la cantidad y entonces suponer que el coeficiente deberá ser negativo.

Capítulo 4

Diseño de Experimentos

4.1. El Experimento y sus Fines (Conceptos Básicos de Experimentos)

En alguna ocasión se planteó el siguiente problema en el curso de estadística Aplicada:

Una empresa que se dedica a la manufactura de tenis tiene que decidir entre cuatro proveedores de material para la suela de sus tenis, llamados proveedor A, B, C y D. Se le encomienda la selección de uno de ellos, con el que realizará un contrato por cinco años para que le entregue el material para la suela de sus tenis, si el costo del material de cada uno de los proveedores fuera el mismo, entonces el criterio de selección estaría basado en el desgaste del material, esto es, qué material se desgasta menos y asegura una mayor vida del producto.

El problema anterior puede ser resuelto con ayuda del diseño de un experimento. En este caso los alumnos propusieron un mecanismo con el cual pudieran probar el desgaste que sufriría cada material, esto es, con ayuda de una máquina que tuviera un disco con una superficie abrasiva, se haría girar a un cierto número de revoluciones constantes y se colocarían muestras del material aplicando una presión constante en un determinado tiempo fijado y se registraría el desgaste sufrido por el material sujeto a la fricción, a través de la medida del espesor del mismo. De manera simple, podemos decir que el criterio para seleccionar al proveedor sería a través de la elección del material que tuviera el menor desgaste, o el que mantuviera un mayor espesor.

4.2. Diseño del Experimento

El ejemplo anterior puede ayudar a ilustrar el concepto de diseño de un experimento. Con ayuda de una prueba de hipótesis se podría establecer lo siguiente:

$$H_0: \quad \mu_A = \mu_B = \mu_C = \mu_D$$

$$H_A: \quad Alguna \quad media \quad es \quad diferente$$

Las hipótesis anteriores pueden ayudar a definir si existe alguna diferencia entre los materiales, y si la hay, entonces es posible encontrar aquel que sea mejor. También se puede encontrar que no haya diferencia entre ellos y por lo tanto se deberá establecer otro criterio de selección.

El diseño de experimentos debe permitir detectar si existen o no diferencias entre los materiales, de tal suerte que se evite cualquier clase de sesgo en el mismo, que pueda favorecer o perjudicar a alguno de los proveedores.

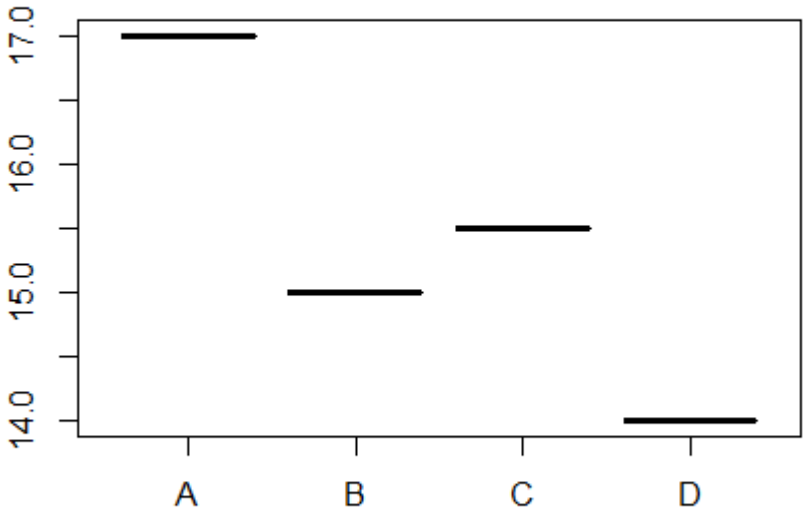
4.3. Número de Ensayos

Si se probara un pedazo de cada material con las mismas dimensiones cada uno, y se midiera su espesor, es probable que se tendría una conclusión errónea al decidir cuál escoger, ya que al tomar una sola muestra de tamaño uno de cada material, esta no sería representativa de la población a la que pertenecen cada material, por lo que se hace necesario que se realicen un número de ensayos lo suficientemente grande, esto es, tomar un tamaño de muestra de cada material lo suficientemente grande como para que la prueba sea significativa. Suponga que las poblaciones de los desgastes de cada uno de los cuatro materiales tienen una distribución normal con las siguientes medias en milímetros y que sus desviaciones, también en milímetros, son iguales.

$$\mu_A = 17, \qquad \mu_B = 15, \qquad \mu_C = 15{,}5, \qquad \mu_D = 14$$
$$\sigma_{A,B,C,D} = 2$$

En este caso es obvio que el material del proveedor A es el que tiene el menor desgaste al poseer el mayor espesor si se pudieran probar todos los materiales, y el peor es D con el menor espesor final, después de la prueba. Gráficamente esto sería:

Figura 4.1: Medias poblacionales de los cuatro materiales.



Como se pude observar, probar todos los materiales no es conveniente, por lo que se sugiere tomar un tamaño de muestra lo suficientemente grande para que la prueba sea válida.

4.4. Análisis de Variaciones

Con todo lo anterior, lo que se espera, es que en el diseño del experimento, las variaciones debidas al materia sean más grandes que las variaciones debidas a los errores, que en este caso son variables aleatorias, para poder distinguir qué material es el mejor.

4.5. Análisis de Resultados

Finalmente, los análisis de los resultados obtenidos con el experimento permitirán concluir si existen diferencias entre los materiales, para posteriormente comenzar la búsqueda de cuál es el mejor.

4.6. Modelos de Análisis de varianza por uno y dos criterios de variación.

En este tipo de modelos, se comparan más de dos poblaciones $K > 2$, de los cuales desconocemos sus medias poblacionales; $\mu_1, ..., \mu_i, ..., \mu_k$.

Lo que persigue es hacer inferencias sobre las medias de las K poblaciones, a través de una serie de muestras de cada una de ellas.

Población 1. $x_{11}, x_{12}, ..., x_{1n_1}$

...

Población i. $x_{i1}, x_{i2}, ..., x_{in_i}$

...

Población k. $x_{k1}, x_{k2}, ..., x_{kn_k}$

Se dice que el experimento está balanceado cuando los tamaños de muestra son iguales $n_1 =, ..., n_i = ...n_k$. Cuando alguno(s) de los tamaños de las muestras difieren de los demás, se dice que el experimento está desbalanceado.

El tamaño de muestra total resulta de:

$$n_T = n_1 + ...n_i + ...n_k$$

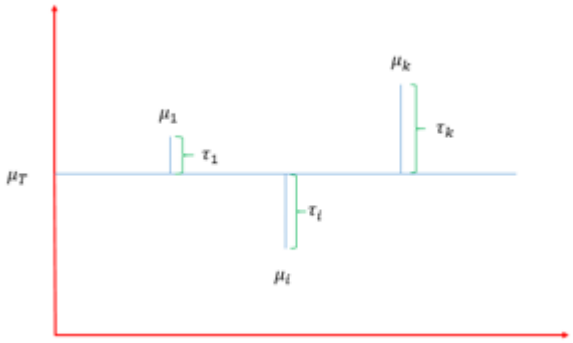
Las hipótesis que se establecen en el diseño de experimento, son las siguientes:

$$H_0 : \mu_1 = ...\mu_i = ...\mu_k$$

$$H_A : \mu_i \neq \mu_j, para \quad alguna \quad i \neq j$$

La siguiente figura ejemplifica lo anterior:

Figura 4.2: Promedio de las medias poblacionales



En la figura anterior se aprecian las variables τ, las cuales son resultado de $\tau_i = \mu_T - \mu_i$, por lo que las hipótesis se pueden replantear de la siguiente manera:

$H_0 : \quad \tau_1 = ...\tau_i = ...\tau_k = 0$

$H_A : \quad \tau_i \neq 0 \quad para \quad alguna \quad i = 1, 2, ..., k$

La representación lineal estadística de las variables x_{ij}, es la siguiente:

$$x_{ij} = \mu_T + \tau_i + \epsilon_{ij}$$

donde ϵ_{ij} es el término de error.

Lo que se espera es que la variabilidad debida al tratamiento sea mayor a la variabiliad debida al error. Lo anterior se puede expresar de la manera siguiente:

$$SC_T = SC_{TRAT} + SC_E$$

donde

SC_T Suma de los Cuadrados Totales

SC_{TRAT} Suma de los Cuadrados del Tratamiento

SC_E Suma de los Cuadrados de los Errores

Por otra parte los cuadrados medios de los tratamientos y de los errores se expresan de la siguiente forma

$$CM_{TRAT} = \frac{SC_{TRAT}}{k-1}$$

donde

CM_{TRAT} es el Cuadrado Medio de los Tratamientos

$$CM_E = \frac{SC_E}{n_\tau - k}$$

donde

CM_E es el Cuadrado Medio de los Errores

Finalmente, el estadísitico de prueba empleado es la F de Fisher, el cual se obtiene a través de:

$$F_0 = \frac{CM_{TRAT}}{CM_E}$$

El supuesto base para poder llevar a cabo las pruebas de hipótesis, es que las varianzas de todas las poblaciones a ser comparadas, son iguales.

Ejemplo **4.1** *Se tienen cuatro poblaciones A, B, C y D, de los cuales se extrajeron muestras de tamaño 10 cada una, el experimento está balanceado. se generarán números aleatorios con diferentes medias y misma varianza, para emplear el análisis de varianza, los cálculos se muestran a continuación:*

```r
#Se ingresa una semilla para generar los mismos valores
set.seed(123)

#Se generan las 10 muestras de las cuatro poblaciones
A<-rnorm(10,40,5) #Media poblacional de 40
B<-rnorm(10,45,5) #Media poblacional de 45
C<-rnorm(10,35,5) #Media poblacional de 35
D<-rnorm(10,50,5) #Media poblacional de 50


#Se muestran los datos en una tabla
datos<-data.frame(A,B,C,D)
datos
```

```
##            A        B        C        D
## 1   37.19762 51.12041 29.66088 52.13232
## 2   38.84911 46.79907 33.91013 48.52464
## 3   47.79354 47.00386 29.86998 54.47563
## 4   40.35254 45.55341 31.35554 54.39067
## 5   40.64644 42.22079 31.87480 54.10791
## 6   48.57532 53.93457 26.56653 53.44320
## 7   42.30458 47.48925 39.18894 52.76959
## 8   33.67469 35.16691 35.76687 49.69044
## 9   36.56574 48.50678 29.30932 48.47019
## 10 37.77169 42.63604 41.26907 48.09764
```

```r
#Se preparan los datos para ser introducidos a la función aov (Analysis
of Variance)
VP<-c(A,B,C,D)
id<-c("A", "B", "C", "D")
idt<-rep(id, each=10)
```

```r
#Se muestra la estructura que deben tener los datos
datos1<-data.frame(idt, VP)
head(datos1)
```

```
##   idt       VP
## 1   A 37.19762
## 2   A 38.84911
## 3   A 47.79354
## 4   A 40.35254
## 5   A 40.64644
## 6   A 48.57532
```

```r
tail(datos1)
```

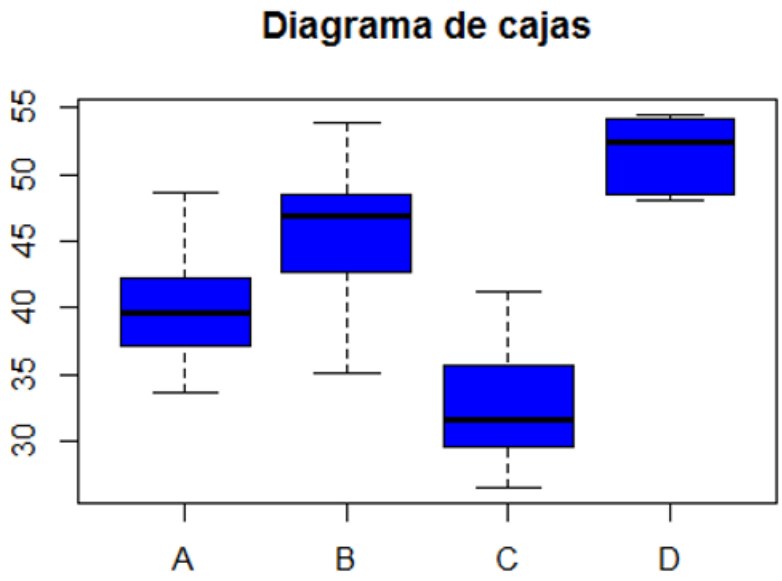
```
##    idt       VP
## 35   D 54.10791
## 36   D 53.44320
## 37   D 52.76959
## 38   D 49.69044
## 39   D 48.47019
## 40   D 48.09764
```

```r
#Se introducen a la función
a<-aov(VP ~ idt, data=datos1)

#Se muestran los resultados
summary(a)
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## idt          3 1924.7   641.6   32.78 2.15e-10 ***
## Residuals   36  704.6    19.6
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Del análisis de varianza, con base a el valor $P = 2{,}15X10^{-10}$, se concluye, que existe suficiente evidencia para rechazar la hipótesis nula, a un nivel de significancia $\alpha = 0{,}05$. Por lo tanto, podemos suponer que al menos una de las medias poblacionales es diferente a las demás. A continuación se muestra el diagrama de caja para cada muestra.

```
boxplot(VP ~ idt, col="Blue", main="Diagrama de cajas")
```



El díagrama de cajas permite visulaizar las diferencias entre las pobalciones. Para probar la eficiencia del modelo, se hará el mismo ejemplo, pero con diferencia de una unidad en los parámetros poblacionales de las medias.

En este caso no habrá suficiente evidencia para rechazar H_0, al nivel de significancia $\alpha = 0{,}05$, sin embargo se podrá rechazar si el nivel de significancia sugerido fuera de $\alpha = 0{,}10$. A continuación se presenta el diagramas de caja para este caso.

```r
#Se ingresa una semilla para generar los mismos valores
set.seed(321)

#Se generan las 10 muestras de las cuatro poblaciones
A1<-rnorm(10,40,5) #Media poblacional de 40
B1<-rnorm(10,41,5) #Media poblacional de 41
C1<-rnorm(10,39,5) #Media poblacional de 39
D1<-rnorm(10,42,5) #Media poblacional de 42


#Se muestran los datos en una tabla
datos<-data.frame(A1,B1,C1,D1)
datos
```

```
##          A1       B1       C1       D1
## 1  48.52452 42.73851 43.58628 34.34794
## 2  36.43981 48.42296 38.46469 44.07857
## 3  38.61008 41.94163 43.94168 45.17099
## 4  39.40175 53.21630 33.63881 48.15424
## 5  39.38020 35.23280 35.20992 41.22718
## 6  41.34092 36.97664 39.47500 42.57270
## 7  43.63421 43.28035 27.34534 30.86868
## 8  41.16568 43.10166 41.08758 50.17895
## 9  41.69557 43.88792 33.39836 41.20340
## 10 37.24043 43.23178 36.62658 42.14134
```

```r
#Se preparan los datos para ser introducidos a la función aov (Analysis
of Variance)
VP1<-c(A1,B1,C1,D1)
id1<-c("A1", "B1", "c1", "D1")
idt1<-rep(id, each=10)
```

```r
#Se muestra la estructura que deben tener los datos
datos2<-data.frame(idt1, VP1)
head(datos2)
```

```
##   idt1      VP1
## 1    A 48.52452
## 2    A 36.43981
## 3    A 38.61008
## 4    A 39.40175
## 5    A 39.38020
## 6    A 41.34092
```

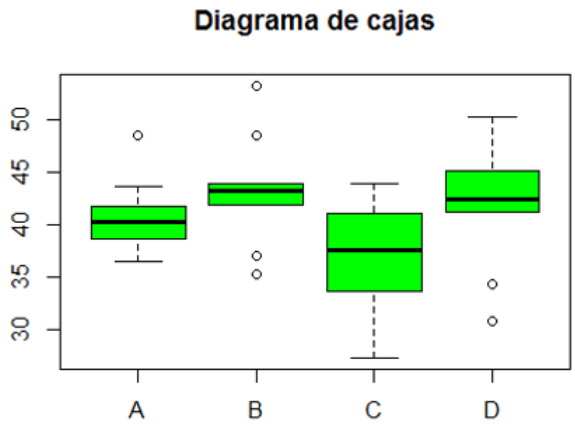
```r
tail(datos2)
```

```
##    idt1      VP1
## 35    D 41.22718
## 36    D 42.57270
## 37    D 30.86868
## 38    D 50.17895
## 39    D 41.20340
## 40    D 42.14134
```

```r
#Se introducen a la función
a1<-aov(VP1 ~ idt1, data=datos2)

#Se muestran los resultados
summary(a1)
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## idt1         3  196.1   65.38   2.671  0.062 .
## Residuals   36  881.2   24.48
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
boxplot(VP1 ~ idt1, col="Green", main="Diagrama de cajas")
```



Los dos ejemplos anteriores nos llevan a la conclusión de que el análisis es preciso, aun cuando la diferencia de las medias no es muy grande y un tamaño de muestra relativamente bajo, por lo que cuando se tenga dudas al respecto, lo mas recomendable es tomar un tamaño de muestra más grande. Lo anterior está en función de la varianza de los datos, ya que el coeficiente de variación se encontraba entre el 12 % y el 15 % para el primer ejemplo y para el segundo se encuentra entre el 11 % y el 13 % aproximadamente.

En el caso de que el experimento no está balanceado.

```r
#Se ingresa una semilla para generar los mismos valores
set.seed(123)

#Se generan las 10 muestras de las cuatro poblaciones
A<-rnorm(10,40,5) #Media poblacional de 40
B<-rnorm(10,45,5) #Media poblacional de 45
C<-rnorm(10,35,5) #Media poblacional de 35
D<-rnorm(10,50,5) #Media poblacional de 50


#Se muestran los datos en una tabla
datos<-data.frame(A,B,C,D)
datos
```

```
##           A        B        C        D
## 1  37.19762 51.12041 29.66088 52.13232
## 2  38.84911 46.79907 33.91013 48.52464
## 3  47.79354 47.00386 29.86998 54.47563
## 4  40.35254 45.55341 31.35554 54.39067
## 5  40.64644 42.22079 31.87480 54.10791
## 6  48.57532 53.93457 26.56653 53.44320
## 7  42.30458 47.48925 39.18894 52.76959
## 8  33.67469 35.16691 35.76687 49.69044
## 9  36.56574 48.50678 29.30932 48.47019
## 10 37.77169 42.63604 41.26907 48.09764
```

```r
#Se preparan los datos para ser introducidos a la función aov (Analysis
of Variance)
VP<-c(A,B,C,D)
id<-c("A", "B", "C", "D")
idt<-rep(id, each=10)
```

Se puede concluir que existe suficiente evidencia para rechazar H_0 ya que $P = 5{,}73x10^{-6}$ es menor al nivel de significancia $\alpha = 0{,}05$.

4.7. Criterios de Comparaciones Múltiples

Entre algunos métodos que podemos mencionar, se encuentran el LSD (por sus siglas en inglés), el método de Tukey y el método de Duncan, por mencionar algunos, todos ellos persiguen establecer cuál de los tratamientos es el que es diferente de los demás. Retomando nuestro ejemplo 4.1 y comenzando con el método LSD.

```
LSD
library("agricolae")

## Warning: package 'agricolae' was built under R version 3.3.2

LSD.test(a,"idt",console=TRUE)

##
## Study: a ~ "idt"
##
## LSD t Test for VP
##
## Mean Square Error:  19.57346
##
## idt,  means and individual ( 95 %) CI
##
##         VP      std  r      LCL      UCL      Min      Max
## A 40.37313 4.768920 10 37.53572 43.21054 33.67469 48.57532
## B 46.04311 5.190367 10 43.20570 48.88052 35.16691 53.93457
## C 32.87721 4.654046 10 30.03980 35.71461 26.56653 41.26907
## D 51.61022 2.636512 10 48.77281 54.44763 48.09764 54.47563
##
## alpha: 0.05 ; Df Error: 36
## Critical Value of t: 2.028094
##
## t-Student: 2.028094
## Alpha    : 0.05
## Least Significant Difference 4.012702
## Means with the same letter are not significantly different
##
##
## Groups, Treatments and means
## a     D   51.61022
## b     B   46.04311
## c     A   40.37313
## d     C   32.87721
```

En este casao, los tratamientos A y D son estadísticamnete diferentes, A con C también lo son.

Tukey

```
TukeyHSD(a)

##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = VP ~ idt, data = datos1)
##
## $idt
##           diff         lwr       upr     p adj
## B-A   5.669982   0.3412749 10.998688 0.0333703
## C-A  -7.495923 -12.8246292 -2.167216 0.0029908
## D-A  11.237095   5.9083879 16.565801 0.0000108
## C-B -13.165904 -18.4946108 -7.837198 0.0000005
## D-B   5.567113   0.2384063 10.895820 0.0377656
## D-C  18.733017  13.4043105 24.061724 0.0000000
```

En este caso se observa que A con D son estadíticamente diferentes, así como B con C y C con D.

Duncan

```
duncan.test(a, "idt",console=TRUE)

##
## Study: a ~ "idt"
##
## Duncan's new multiple range test
## for VP
##
## Mean Square Error:  19.57346
##
## idt,  means
##
##         VP      std  r      Min      Max
## A 40.37313 4.768920 10 33.67469 48.57532
## B 46.04311 5.190367 10 35.16691 53.93457
## C 32.87721 4.654046 10 26.56653 41.26907
## D 51.61022 2.636512 10 48.09764 54.47563
##
## alpha: 0.05 ; Df Error: 36
##
## Critical Range
##        2        3        4
## 4.012702 4.218447 4.352635
##
## Means with the same letter are not significantly different.
##
## Groups, Treatments and means
## a     D   51.61
## b     B   46.04
## c     A   40.37
## d     C   32.88
```

Nuevamente se observa que A con D son estadísticamente diferentes, así como A con B y A con C.

4.8. Modelos de Bloques Completos

En estos modelos se toma en cuenta que existen más de un factor, el cual será denominado como factor de bloque y los factores que se desean probar, serán denominados factores de tratamiento, por lo que la representación del modelo estadístico, quedará como sigue:

$$x_{ij} = \mu_\tau + \tau_i + \gamma_j + \epsilon_{ij}$$

donde:

x_{ij} es el i-ésimo tratamiento del j-ésimo bloque,

μ_τ es la media de las medias poblacionales de los tratamientos ,

τ_i es el efecto debido al tratamiento i,

γ_j es el efecto debido al bloque j,

ϵ_{ij} es término de error que se le atribuye a la variable x_{ij}.

Por lo tanto, se tendrán las sumas de los cuadrados respectivos

$$SC_T = SC_{TRAT} + SC_B + SC_E$$

donde:

SC_B, es la Suma de los Cuadrados del Bloque.

Ejemplo **4.2** *Una compañía desea comprar un nuevo equipo, actualmente cuenta con cuatro equipos de diferentes proveedores, llamados equipos A, B, C y D, que realizan la misma actividad. Se desea saber qué equipo tiene el menor tiempo promedio en realizar dicha actividad, para poder tomar una decisión de compra. Por otro lado, la compañía tiene contratados a cuatro empleados empleado 1, empleado 2, empleado 3 y empleado 4, quienes operan actualmente los equipos. Como lo que se desea es estimar qué equipo es mejor, no se está evaluando las habilidades de los trabajadores. Con el fin de evitar sesgos, se tomaron muestras de los tiempos que tardan en hacer una actividad tomando en cuenta a los diferentes empleados, como se muestra a continuación:*

```
v<-c(7,6,11,9,10,9,17,12,8,10,12,10,9,7,15,8)

dim(v)<-c(4,4)

f<-c("A", "B", "C", "D")
b<-c("Empleado 1", "Empleado 2", "Empleado 3", "Empleado 4")

rownames(v)<-f
colnames(v)<-b

v
```

```
##   Empleado 1 Empleado 2 Empleado 3 Empleado 4
## A          7         10          8          9
## B          6          9         10          7
## C         11         17         12         15
## D          9         12         10          8
```

Se puede apreciar cómo cada empleado operó cada uno de los equipos, con el fin de evitar que la pericia de alguno de ellos sobre los demás creara un sesgo, en caso de que este estuviera operando el equipo más ineficiente de los cuatro e hiciera parecer que ese equipo fuero eficiente. A continuación se realiza la prueba y el análisis del ANOVA.

```
dt4<-c(t(as.matrix(v)))

tm<-rep(f, each=4) #Equipos
bl<-rep(b, times=4) #Empleados

a3<-aov(dt4 ~ tm + bl)

#Se muestran los resultados
summary(a3)

##             Df Sum Sq Mean Sq F value  Pr(>F)
## tm           3   81.5   27.17   13.58 0.00109 **
## bl           3   28.5    9.50    4.75 0.02985 *
## Residuals    9   18.0    2.00
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

En este caso el valor $P = 0{,}00109$, el cual es menor que $\alpha = 0{,}05$, indica que al menos uno de los tratamientos es diferente, mientras que el valor $P = 0{,}02985$ indican que también existen diferencias entre los operadores, al mismo nivel de significanica, pero como se mencionó al inicio del ejemplo, esto no ocasiona un sesgo en el resultado de los tratamientos.

4.9. Modelos de Cuadrados Latinos

En este tipo de diseño se controlan dos bloques y se analiza un tratamiento, la función que se desprende es la siguiente:

$$x_{ijl} = \mu_T + \tau_i + \gamma_j + \delta_l + \epsilon_{ijl}$$

La suma de los cuadrados es la siguiente:

$$SC_T = SC_{TRAT} + SC_{B1} + SC_{B2} + SC_E$$

donde:

SC_{B1} es la Suma de los Cuadrados del Bloque I,

SC_{B2} es la Suma de los Cuadrados del Bloque II.

Los grados de libertad son

$$K^2 = (k-1) + (k-1) + (k-1) + (k-2)(k-1)$$

.

Ejemplo **4.3** *Se desea saber que marca de llanta es mejor para una compañía, la cual está considerando comprar entre cuatro tipos de marcas diferentes, para tal efecto considera que la posición de la llanta, y el tipo de carro son variables que pueden influir, por lo que toma los siguientes datos, en donde el desgaste esta en milésimas de milímetros, después de haber recorrido 30,000 km.*

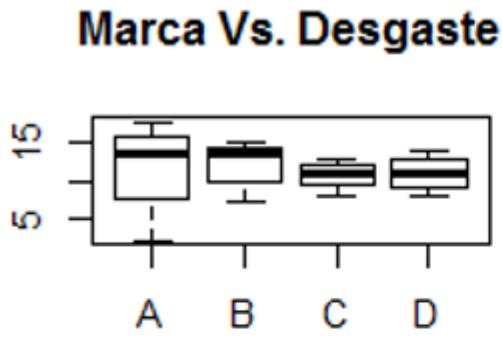
```
carro <- rep(c("c1", "c2", "c3", "c4"), times=4)
pos <- rep(c("p1", "p2", "p3", "p4"), each=4)
marca <- c("C","D","A","B","B","C","D","A","A","B","C","D","D","A","B",
"C")
des <- c(13,10,14,7,15,11,12,2,18,13,11,8,14,13,14,8)

dt4<-data.frame(carro, pos, marca, des)
dt4

##    carro pos marca des
## 1     c1  p1     C  13
## 2     c2  p1     D  10
## 3     c3  p1     A  14
## 4     c4  p1     B   7
## 5     c1  p2     B  15
## 6     c2  p2     C  11
## 7     c3  p2     D  12
## 8     c4  p2     A   2
## 9     c1  p3     A  18
## 10    c2  p3     B  13
## 11    c3  p3     C  11
## 12    c4  p3     D   8
## 13    c1  p4     D  14
## 14    c2  p4     A  13
## 15    c3  p4     B  14
## 16    c4  p4     C   8
```

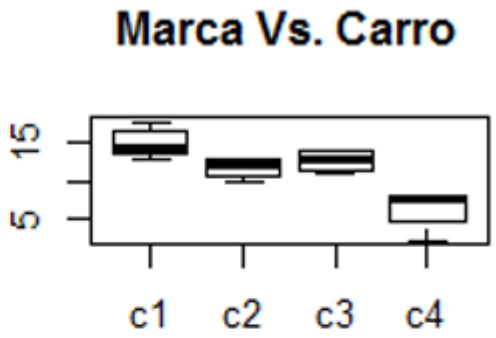
El ANOVA queda como sigue:

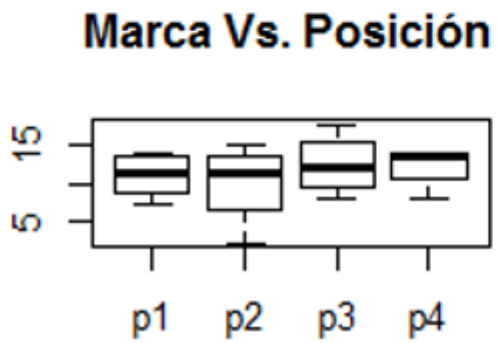
```
a4<-aov(des ~ marca  + pos + carro)
summary(a4)

##             Df Sum Sq Mean Sq F value  Pr(>F)
## marca        3   5.69    1.90   0.374 0.77492
## pos          3  16.19    5.40   1.066 0.43088
## carro        3 165.69   55.23  10.909 0.00765 **
## Residuals    6  30.38    5.06
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Del análisis se concluye a un nivel de significancia del 5 %, que no existen diferencias significativas entre las marcas, sin embargo, si existe diferencia en el tipo de carro y el desgaste ocurrido.

Los siguientes gráficos de caja apoyan la conclusión anterior.







4.10. Modelos de Cuadrados Grecolatinos

En este caso se consideran tres bloques y un tratamiento, por lo que la especificación es la siguiente

$$x_{ijlm} = \mu_T + \tau_i + \gamma_j + \delta_l + \psi_m + \epsilon_{ijlm}.$$

donde

ψ_m: La Suma de los Cuadrados del Bloque III.

La suma de los cuadrados es la siguiente:

$$SC_T = SC_{TRAT} + SC_{B1} + SC_{B2} + SC_{B3} + SC_E$$

Los grados de libertad son:

$$K^2 - 1 = (k-1) + (k-1) + (k-1) + (k-3)(k-1).$$

Ejemplo **4.4** *Del ejemplo de las máquinas y los operadores, imagine que también importa el orden en el que se operan y el lugar donde se hace.*

```
orden<-rep(c("or1","or2","or3", "or4"), times=4)
operador<-rep(c("op1","op2","op3", "op4"), each=4)
met<-c("C","B","A","D","B","C","D","A","D","A","B","C","A","D","C","B")
lugar<-
c("beta","alfa","delta","gama","gama","delta","alfa","beta","delta","gama
","beta","alfa","alfa","beta","gama","delta")
tiempo<-c(11,7,7,10,11,14,15,7,13,6,12,9,8,13,14,7)

dat5<-data.frame(orden,operador,met,lugar,tiempo)
dat5

##    orden operador met lugar tiempo
## 1    or1      op1   C  beta     11
## 2    or2      op1   B  alfa      7
## 3    or3      op1   A delta      7
## 4    or4      op1   D  gama     10
## 5    or1      op2   B  gama     11
## 6    or2      op2   C delta     14
## 7    or3      op2   D  alfa     15
## 8    or4      op2   A  beta      7
## 9    or1      op3   D delta     13
## 10   or2      op3   A  gama      6
## 11   or3      op3   B  beta     12
## 12   or4      op3   C  alfa      9
## 13   or1      op4   A  alfa      8
## 14   or2      op4   D  beta     13
## 15   or3      op4   C  gama     14
## 16   or4      op4   B delta      7
```

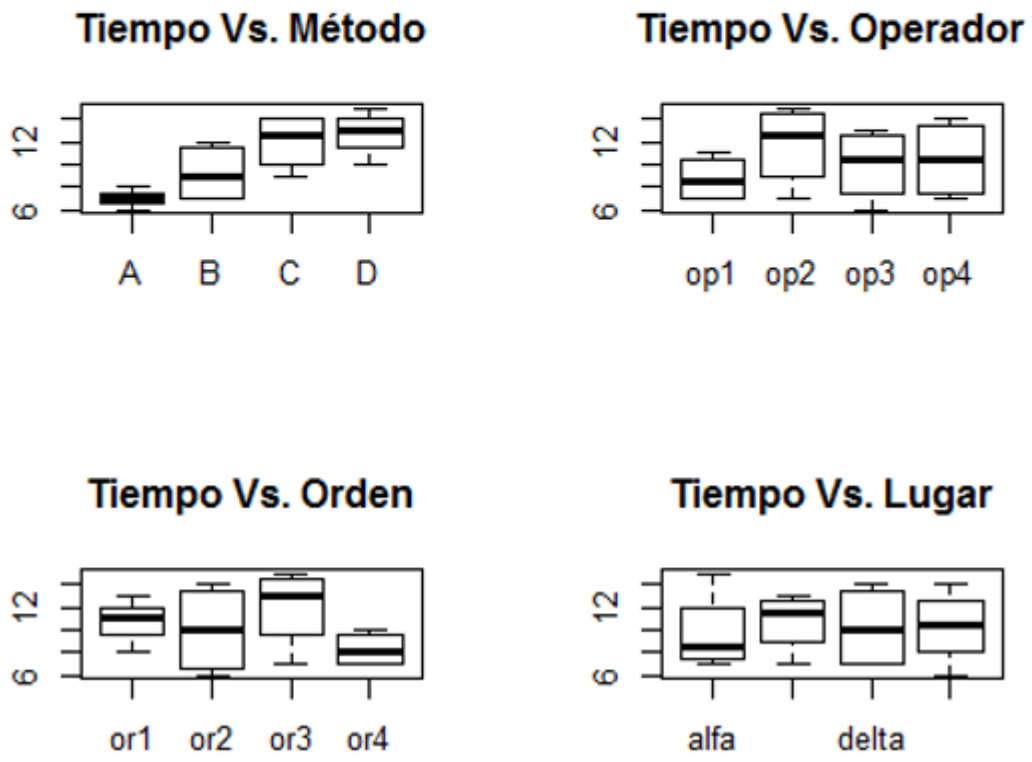
Empleano el ANOVA:

```
a4<-aov(tiempo ~ met + operador + orden  + lugar)
summary(a4)

##             Df Sum Sq Mean Sq F value Pr(>F)
## met          3   83.5  27.833  23.857 0.0135 *
## operador     3   18.5   6.167   5.286 0.1024
## orden        3   29.5   9.833   8.429 0.0567 .
## lugar        3    2.0   0.667   0.571 0.6714
## Residuals    3    3.5   1.167
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

El resultado muestra que al menos un método es diferente a los demás, los otros bloques no son estadísticamente diferentes entre ellos, al nivel de significancia del 5 %, sin embargo, si consideramos el nivel de significancia del 10 %, el orden podría tener diferencias estadísticamente significativas por lo que podría ser importante considerarla.

La siguiente gráfica nos ayuda a confirmar lo anterior:

```
par(mfrow=c(2,2))
{
boxplot(tiempo~met, main="Tiempo Vs. Método")
boxplot(tiempo~operador, main="Tiempo Vs. Operador")
boxplot(tiempo~orden, main="Tiempo Vs. Orden")
boxplot(tiempo~lugar, main="Tiempo Vs. Lugar")
}
```



4.11. Análisis de Covarianza

Retomando el ejemplo 2, corremos dos modelos de regresión, el primero solo con la variable equipo y el segundo con la variable equipo y empleado.

```
f1<-c(7, 10,  8,  9,  6,  9, 10,  7, 11, 17, 12, 15,  9, 12, 10,  8)
ajuste1<-lm(f1 ~ tm) #tm son los equipos
summary(ajuste1)

##
## Call:
## lm(formula = f1 ~ tm)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -2.750 -1.562 -0.125  1.312  3.250
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   8.5000     0.9843   8.636  1.7e-06 ***
## tmB          -0.5000     1.3919  -0.359  0.72568
## tmC           5.2500     1.3919   3.772  0.00266 **
## tmD           1.2500     1.3919   0.898  0.38684
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.969 on 12 degrees of freedom
## Multiple R-squared:  0.6367, Adjusted R-squared:  0.5459
## F-statistic: 7.011 on 3 and 12 DF,  p-value: 0.005591
```

En este caso el coeficiente del equipo C es estadísticamente significativo, con un valor del coeficiente del 5.25.

```
ajuste2<-lm(f1 ~ tm + bl) #bl son los empleados y tm son los equipos
summary(ajuste2)

##
## Call:
## lm(formula = f1 ~ tm + bl)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.7500 -0.8125  0.0000  0.8125  2.0000
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    6.7500     0.9354   7.216    5e-05 ***
## tmB           -0.5000     1.0000  -0.500 0.629071
## tmC            5.2500     1.0000   5.250 0.000528 ***
## tmD            1.2500     1.0000   1.250 0.242824
## blEmpleado 2   3.7500     1.0000   3.750 0.004555 **
## blEmpleado 3   1.7500     1.0000   1.750 0.114044
## blEmpleado 4   1.5000     1.0000   1.500 0.167851
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.414 on 9 degrees of freedom
## Multiple R-squared:  0.8594, Adjusted R-squared:  0.7656
## F-statistic: 9.167 on 6 and 9 DF,  p-value: 0.002054
```

En este caso, además de ser significativo el equipo C, también lo es el empleado 2, con un valor del coeficiente de 3.75.

Lo que significa que tanto el empleado como el equipo afectan al tiempo.

```
drop1(ajuste2, test = "F")

## Single term deletions
##
## Model:
## f1 ~ tm + bl
##        Df Sum of Sq  RSS    AIC F value   Pr(>F)
## <none>              18.0 15.885
## tm      3      81.5 99.5 37.241  13.583 0.001088 **
## bl      3      28.5 46.5 25.070   4.750 0.029846 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

El resultado contrasta la significancia de la variable concomitante (dada la presencia del factor) y la del factor (dada la presencia de la variable concomitante). La diferencia entre máquinas es significativa.

4.12. Diseño Factorial 2K

En esta clase de diseños se consideran k factores con dos niveles cada uno y un total de 2^k tratamientos. Con el diseño completo 2^k, se pueden analizar $2^{(k-1)}$ efectos:

$\binom{k}{1} = k, \quad estos \quad son \quad los \quad efectos \quad parciales,$

$\binom{k}{2} = \frac{k!}{2!(k-2)!}, \quad estas \quad son \quad las \quad interacciones \quad dobles,$

$\binom{k}{a} = \frac{k!}{a!(k-a)!}, \quad esas \quad son \quad las \quad interacciones \quad triples,$

$\binom{k}{k} = 1, \quad finalmente \quad las \quad interacciones \quad de \quad los \quad k \quad factores.$

En este caso y con ayuda de la notación de Yates, se estiman los efectos:

$$Efecto \quad ABC...k = \frac{1}{n2^{k-1}} \quad [Contraste \quad ABC...k]$$

La Sumas de los cuadrados con un grado de libertad:

$$SC_{(ABC...K)} = \frac{1}{n2^k}[ContrasteABC...k]^2$$

La Suma de los Cuadrados Totales es:

$$SC_T = \sum_{i=1}^{n2^k} x_i^2 - \frac{x_i^2}{n2^k}$$

con:

$n2^k$ grados de libertad.

La Suma de los Cuadrados del Error tiene:

$2^k(n-1)$ grados de libertad.

Las hipótesis que se plantean son las siguientes:

H_0 : efectos=0.

H_A : efectos $\neq 0$.

```r
datos <-
matrix(c(560,590,679,640,643,591,652,625,1047,1042,759,858,1085,1053,719,
870),byrow=T,ncol=2)
dim(datos)<-c(8,2)

dimnames(datos)<-
list(c("(1)","a","b","ab","c","ac","bc","abc"),c("Rep1","Rep2"))

A <- rep(c(-1,1),4)
B <- rep(c(-1,-1,1,1),2)
C <- c(rep(-1,4),rep(1,4))

Total <- apply(datos,1,sum)

cbind(A,B,C,datos,Total)
```

```
##      A  B  C Rep1 Rep2 Total
## (1) -1 -1 -1  560  590  1150
## a    1 -1 -1  679  640  1319
## b   -1  1 -1  643  591  1234
## ab   1  1 -1  652  625  1277
## c   -1 -1  1 1047 1042  2089
## ac   1 -1  1  759  858  1617
## bc  -1  1  1 1085 1053  2138
## abc  1  1  1  719  870  1589
```

```r
datosv <- c(t(datos))

Af <- rep(as.factor(A),rep(2,8))
Bf <- rep(as.factor(B),rep(2,8))
Cf <- rep(as.factor(C),rep(2,8))

datosml <- lm(datosv ~ Af*Bf*Cf)
```

```
anova(datosml)

## Analysis of Variance Table
##
## Response: datosv
##           Df Sum Sq Mean Sq  F value    Pr(>F)
## Af         1  40905   40905  16.5670 0.0035821 **
## Bf         1    248     248   0.1005 0.7593822
## Cf         1 376076  376076 152.3151 1.731e-06 ***
## Af:Bf      1   2576    2576   1.0431 0.3369830
## Af:Cf      1  95018   95018  38.4835 0.0002584 ***
## Bf:Cf      1     28      28   0.0112 0.9184567
## Af:Bf:Cf   1    150     150   0.0608 0.8114799
## Residuals  8  19752    2469
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Los efectos Af, Cf y la combinación Af:Cf, son estadísticamente significativas a un nivel de significancia $\alpha = ,05$.

Capítulo 5

Estadística no Paramétrica

La estadáitica no paramétrica basa sus decisiones no en parámetros como su nombre lo indica, sino más bien en la forma en la que se distribuyen los datos, normalmente se emplea cuando no hay suficinete información y no existe certeza de que los datos se distribuyan de forma normal. También es empleda cuando los datos se presentan en una escala nominal.

Entre las pruebas que se emplean podemos mencionar:

Prueba de los Signos.

Prueba de Rangos con Signos de Wilcoxon.

Prueba de Kruskal – Wallis.

Prueba de Rachas.

Prueba de Kolmogorov – Smirnov.

Prueba de Correlacion de Rangos.

5.1. Prueba de los Signos

Ejemplo **5.1** *Los siguientes datos, un total de 11, representan el número de días que un celular opera antes de recibir una recarga.*

```
x<-c(1.5, 2.1, 0.8, 1.2, 1.9, 1.5, 1.7, 1.4, 1.9, 1.1, 1.6)
x

##  [1] 1.5 2.1 0.8 1.2 1.9 1.5 1.7 1.4 1.9 1.1 1.6
```

Se desea probar la hipótesis de que los celulares tardan para ser recargados 1.7 días.

$H_0 : \mu^{\sim} = 1{,}7$

$H_a : \mu^{\sim} \neq 1{,}7$

Con un valor de $\alpha = 0{,}05$.

```
v<-1.7              #Lo que se desea probar.
a<-length(x[x>v])   #Tamaño de valores mayores a 1.7 días.
a

## [1] 3

b<-length(x[x==v])  #Tamaño de valores exactamente igual a 1.7 días.
b

## [1] 1

d<-length(x)-b      #Tamaño de valores que no son iguales a 1.7 días
d

## [1] 10

binom.test(a, d)

##
##  Exact binomial test
##
## data:  a and d
## number of successes = 3, number of trials = 10, p-value = 0.3438
## alternative hypothesis: true probability of success is not equal to 0.
5
## 95 percent confidence interval:
##  0.06673951 0.65245285
## sample estimates:
## probability of success
##                    0.3
```

Como el valor $P = 0{,}3438$ del estadístico es mayor que el nivel de significancia α, no existe suficiente evidencia para recharzar H_0.

Ejemplo **5.2** *Se desea saber sí las llantas "llr"son mejores que las llantas "llc", para lo que se tomó la siguiente muestra en años que tardaron en cambiar las llantas.*

```
##    Auto_N llr llc
## 1       1 4.2 4.1
## 2       2 4.7 4.9
## 3       3 6.6 6.2
## 4       4 7.0 6.9
## 5       5 6.7 6.8
## 6       6 4.5 4.4
## 7       7 5.7 5.7
## 8       8 6.0 5.8
## 9       9 7.4 6.9
## 10     10 4.9 4.9
## 11     11 6.1 6.0
## 12     12 5.2 4.9
## 13     13 5.7 5.3
## 14     14 6.9 6.5
## 15     15 6.8 7.1
## 16     16 4.9 4.8
```

Solución

Se plantean las siguientes hipótesis:

$H_0 : \mu_1^{\sim} - \mu_2^{\sim} = 0.$

$H_a : \mu_1^{\sim} - \mu_2^{\sim} > 0.$
Con un nivel de significancia $\alpha = 0{,}05$.

```
x1<-llr-llc

x1

##  [1]  0.1 -0.2  0.4  0.1 -0.1  0.1  0.0  0.2  0.5  0.0  0.1  0.3  0.4
0.4
## [15] -0.3  0.1

v1<-0
a1<-length(x1[x1>v1])
b1<-length(x1[x1==v1])
d1<-length(x1)-b1

binom.test(a1, d1)

##
##  Exact binomial test
##
## data:  a1 and d1
## number of successes = 11, number of trials = 14, p-value = 0.05737
## alternative hypothesis: true probability of success is not equal to 0.
5
## 95 percent confidence interval:
##  0.4920243 0.9534207
## sample estimates:
## probability of success
##              0.7857143
```

Como el valor $P = 0{,}05737$ de la prueba resulto ligeramente mayor que el nivel de significancia α, no existe suficiente evidencia para rechazar H_0.

Otra forma de abordar estos problemas es a través de la probabilidad de que ocurra un evento.

Ejemplo **5.3** *Se le preguntan a algunas parejas de recién casados, por separado, cuántos hijos desean tener, se muestran los resultados:*

```
##    Pareja H M
## 1       1 1 2
## 2       2 2 1
## 3       3 1 0
## 4       4 1 1
## 5       5 0 0
## 6       6 1 0
## 7       7 0 1
## 8       8 2 1
## 9       9 0 1
## 10     10 1 1
```

Se desea establecer que:

$H_0 : p = 0{,}5$ Esto es, que coinciden en el número de hijos.

$H_1 : P < 0{,}5$ Esto sería que no coinciden.

Resultado:

```
x2<-M-H
x2

##  [1]  1 -1 -1  0  0 -1  1 -1  1  0

v2<-0

a2<-length(x2[x2>v2])
a2

## [1] 3

b2<-length(x2[x2==v2])
b2

## [1] 3

d2<-length(x2)-b2
d2

## [1] 7

binom.test(a2, d2)$p.value/2

## [1] 0.5
```

Por lo tanto no existe suficiente evidencia para suponer que no desean la misma cantidad de hijos.

5.2. Prueba de Rangos con Signos de Wilcoxon.

En esta prueba a diferencia de la de signos, si se consideran las magnitudes de las diferencias, cuando los valores se encuentran en escala ordinal.

Ejemplo **5.4** *Se presentan las calificaciones de 14 alumnos, tomados al azar, que obtiuvieron en su primer parcial "pp" y su segundo parcial "sp", determine si existen diferencias entre los parciales.*

```
##    Alumno_N pp sp
## 1         1 70 95
## 2         2 70 85
## 3         3 70 90
## 4         4 80 97
## 5         5 70 85
## 6         6 72 90
## 7         7 70 85
## 8         8 80 95
## 9         9 75 85
## 10       10 90 90
## 11       11 90 90
## 12       12 70 85
## 13       13 70 95
## 14       14 70 85
```

Se empleará la prueba no paramétrica de rangos con signos de Wilcoxon.

```
wilcox.test(tabla1$pp, tabla1$sp, paired=TRUE)

## Warning in wilcox.test.default(tabla1$pp, tabla1$sp, paired = TRUE): c
annot
## compute exact p-value with ties

## Warning in wilcox.test.default(tabla1$pp, tabla1$sp, paired = TRUE): c
annot
## compute exact p-value with zeroes

##
##  Wilcoxon signed rank test with continuity correction
##
## data:  tabla1$pp and tabla1$sp
## V = 0, p-value = 0.002192
## alternative hypothesis: true location shift is not equal to 0
```

A un nivel de significancia del 5 % se puede concluir que ambas pruebas no pertenencen a la misma población, dado que su valor $P = 0{,}002192$.

5.3. Prueba de Kruskal-Wallis

Ejemplo **5.5** *Se desea determinar si hay diferencias de cantidad de SO_2 en el centro de la CDMX a través de los 365 días de los 12 meses del año 2016, para lo que se consultó la siguiente página: http://www.aire.cdmx.gob.mx/default.php?opc = %27aqBjnmU = %27.*

```
contaminante<-read.csv(file.choose(), header=TRUE) #Se extrajeron los dat
os y se convirtieron a uan archivo de excel con extensión csv.
head(contaminante)

##   Mes SO2
## 1   1  20
## 2   1  19
## 3   1  16
## 4   1  16
## 5   1  12
## 6   1   4

tail(contaminante)

##     Mes SO2
## 361  12   4
## 362  12   7
## 363  12   8
## 364  12  12
## 365  12   8
## 366  12   9

kruskal.test(SO2 ~ Mes, data = contaminante)

##
##  Kruskal-Wallis rank sum test
##
## data:  SO2 by Mes
## Kruskal-Wallis chi-squared = 77.238, df = 11, p-value = 5.034e-12
```

Se puede concluir a un 5 % de nivel de significancia que el ozono no es idéntico entre los meses, dado que el valor $P = 5{,}034X10^{-12}$.

5.4. Prueba de Rachas

Esta prueba demuestra que una secuencia de datos que se toman de una muestra en forma consecutiva no son aleatorios, por ejemplo tomemos la siguiente serie de datos.

```
x1<-c(1,2,3,4,5,6,7,8,9,10)
x1

##  [1]  1  2  3  4  5  6  7  8  9 10
```

A todas luces la secuencia de datos no son aleatorios, al aplicar la prueba tenemos:

```
library(tseries)

## Warning: package 'tseries' was built under R version 3.3.2

runs.test(as.factor(x1>median(x1)))

##
##  Runs Test
##
## data:  as.factor(x1 > median(x1))
## Standard Normal = -2.6833, p-value = 0.00729
## alternative hypothesis: two.sided
```

En este caso como el valor $P = 0{,}00729$ es menor a un valor predeterminado de α, por ejemplo 0.05, existe evidencia suficiente para decir que la secuencia de datos no es aleatoria. Ahora considere la siguiente secuencia de datos:

```
set.seed(123)
x2<-runif(10)
x2

##  [1] 0.2875775 0.7883051 0.4089769 0.8830174 0.9404673 0.0455565 0.528
1055
##  [8] 0.8924190 0.5514350 0.4566147
```

Los datos que se encuentran en x2 son aleatorios, en realidad son seudo aleatorios, pero servirán para hacer la prueba.

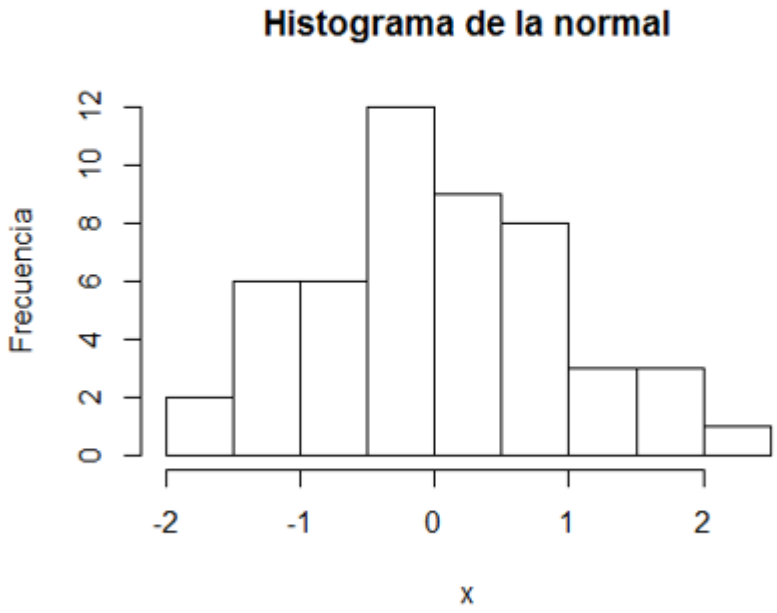
```
runs.test(as.factor(x2>median(x2)))

##
##  Runs Test
##
## data:  as.factor(x2 > median(x2))
## Standard Normal = 0.67082, p-value = 0.5023
## alternative hypothesis: two.sided
```

Como el valor de $P = 0{,}5023$ que es mayor a un nivel de significancia dado, por ejemplo 0.05, no existe suficiente evidencia para decir que la secuencia en x2 no es aleatoria.

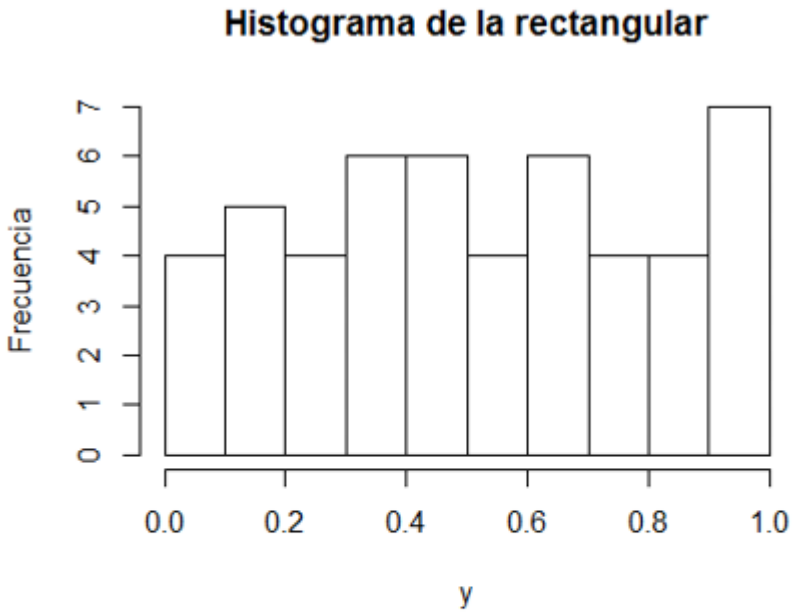
5.5. Prueba de Kolmogorov-Smirnov

Esta prueba, a pesar de que emplea un tamaño de muestra lo suficientemente grande, lo que establece es si una distribución de probabilidad, de una población de la que solo se tiene una muestra, obedece al comportamiento de alguna distribución conocida. Generalmente se emplea cuando se sabe que los datos pertenecen a una función de distribución continua.

```
set.seed(123)
x <- rnorm(50)
y <- runif(50)

hist(x, main="Histograma de la normal", ylab="Frecuencia")
```




```
hist(y, main="Histograma de la rectangular", ylab="Frecuencia")
```

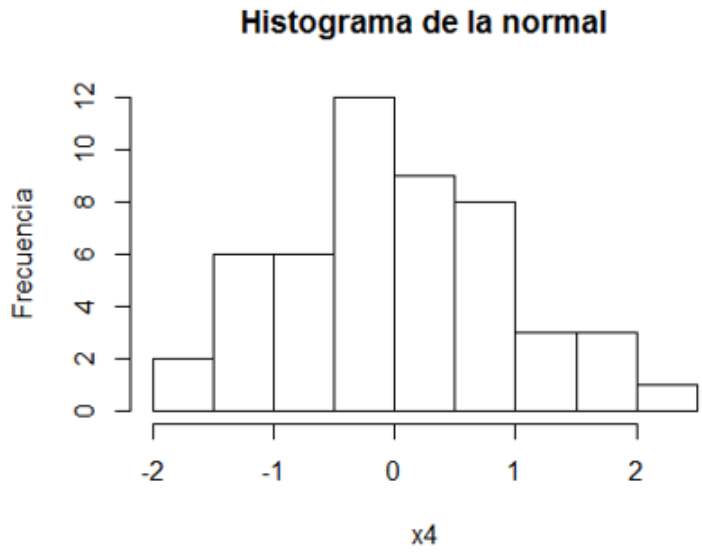



```
ks.test(x, y)

##
##  Two-sample Kolmogorov-Smirnov test
##
## data:  x and y
## D = 0.52, p-value = 1.581e-06
## alternative hypothesis: two-sided
```

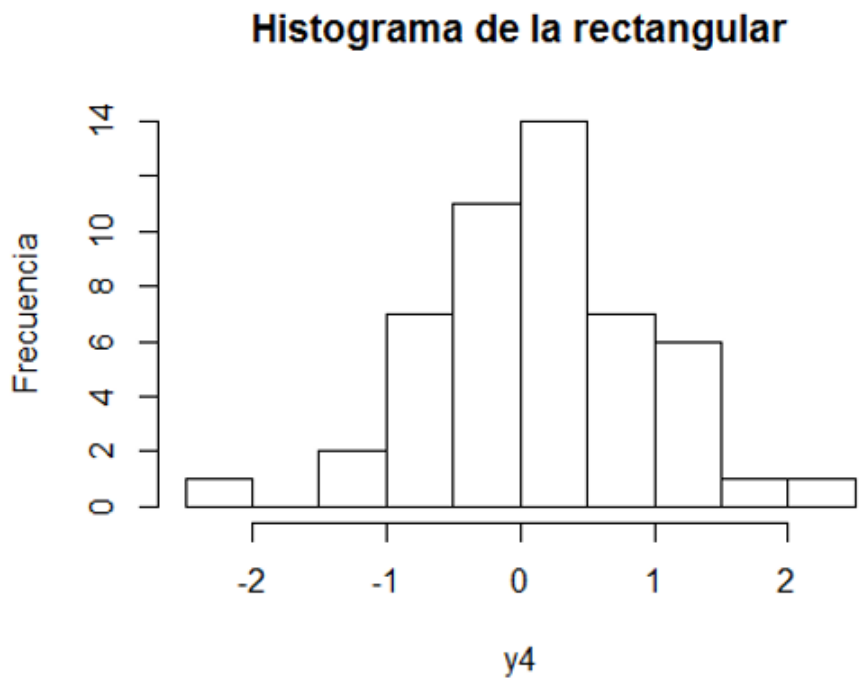
Es claro que ambas gráficas no son iguales, la que corresponde a x, pertenece a una distribución normal, mientras que la que pertenece a y es una distribución rectangular. El valor $P = 1{,}581X10^{-6}$ es menor a un nivel de significancia dado, por ejemplo 0.05, los datos no pertenecen a la misma distribución.

Para demostrar el caso en el que la hipótesis de que una muestra pertenece a una determinada población, tomemos el siguiente caso:

```
set.seed(123)
x4 <- rnorm(50)
x4m<-mean(x4)
x4m

## [1] 0.03440355

x4sd<-sd(x4)
x4sd

## [1] 0.92587

y4<-rnorm(50,x4m,x4sd)

hist(x4, main="Histograma de la normal", ylab="Frecuencia")
```




```
hist(y4, main="Histograma de la rectangular", ylab="Frecuencia")
```




```
ks.test(x4, y4)

##
##  Two-sample Kolmogorov-Smirnov test
##
## data:  x4 and y4
## D = 0.16, p-value = 0.5487
## alternative hypothesis: two-sided
```

En este caso x4 pertenece a una distribución normal con media cero y desviación estándar de uno, mientras que y4 tiene media de 0,03440355 y desviación estándar de 0,92587, así mismo las gráficas de x4 y de y4 varían en su forma, sin embargo ambas fueron generadas de una distribución normal, por lo que se espera que la prueba establezca que no existen diferencias significativas entra ambas distribuciones, lo que se comprueba al tener que el valor $P = 0{,}5487$, que es mayor a un nivel de significacnia dado, por ejemplo a un 0.05, por lo que no hay evidencia para decir que no se distribuya de forma normal.

5.6. Prueba de Correlación de Rangos

Esta prueba establece que:

$H_0 : \rho = 0.$

$H_A : \rho \neq 0.$

Por ejemplo, se puede pensar en la eficiencia de un equipo y en la satisfacción de los clientes valorado de cero a diez.

```
set.seed(123)
N<-c(1:10)
Efi<-c(7,8,10,6,5,7,8,7,7,6) #Eficiencia
Satis<-sample(c(1:10),10) #Satisfcción
tabla3<-data.frame(N, Efi, Satis)
tabla3
```

```
##     N Efi Satis
## 1   1   7     3
## 2   2   8     8
## 3   3  10     4
## 4   4   6     7
## 5   5   5     6
## 6   6   7     1
## 7   7   8    10
## 8   8   7     9
## 9   9   7     2
## 10 10   6     5
```

```
cor.test(Efi, Satis, method="spearman")
```

```
## Warning in cor.test.default(Efi, Satis, method = "spearman"): Cannot
## compute exact p-value with ties
```

```
##
##  Spearman's rank correlation rho
##
## data:  Efi and Satis
## S = 146.31, p-value = 0.7553
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.1132882
```

Es claro que los valores de satisfacción son generados aleatoriamente, por lo que se espera que no haya suficiente evidencia para rechazar la hipótesis nula, en el que el coeficiente de correlación a nivel poblacional sea de cero entre Eficiencia y Satisfacción, lo cual se comprueba con el valor $P = 0{,}7553$ del estadístico de prueba que es mayor a un nivel de significancia dado, por ejemplo de 0.05.

Capítulo 6

Confiabilidad

La confiabilidad es un área de la estadística, que nos permite identificar la probabilidad de que un sistema no falle, ya que dicho sistema se encuentra interrelacionado, dependiendo de la secuencia que siga, ya sea en paralelos, en serie o una combinación de ambas, la confiabilidad del mismo, dependiendo de cada caso, tendrá cierta probabilidad de no fallar.

En este caso a la confiabilidad la denotaremos con la letra r".

6.1. Confiabilidad, Usos y Aplicaciones

La confiabilidad es la probabilidad de que un sistema no falle, como anteriormente se mencionó, sus usos y aplicaciones son variadas, en el caso de la Ingeniería Industrial, en la cual una de sus principales áreas de aplicación son los procesos, podemos apreciar que la confiabilidad es de gran importancia, ya que si por alguna razón dicho proceso falla en alguna de sus partes, dependiendo de la composición del mismo, en serie o paralelo, podrá afectar la producción que se planea llevar a cabo, trayendo consigo pérdidas monetarias a las empresas.

Una de sus aplicaciones la podemos encontrar en los trenes que se forman para ensamblar vehículos, en los cuales, por cuestiones de calidad cada una de sus partes son importantes, si por alguna razón llegarán a fallar, tendrían serias repercusiones en la cantidad de vehículos a producir. Otro ejemplo de procesos, son la manufactura de bienes o la prestación de servicios, los cuales se caracterizan por tener una secuencia lógica para poder llegar a un fin.

6.2. Distribuciones del Tiempo de Falla

Entre las principales distribuciones empleadas, por sus propiedades, podemos mencionar a la distribución Exponencial, la distribución Gama y la distribución Weibull.

6.3. Distribución Exponencial

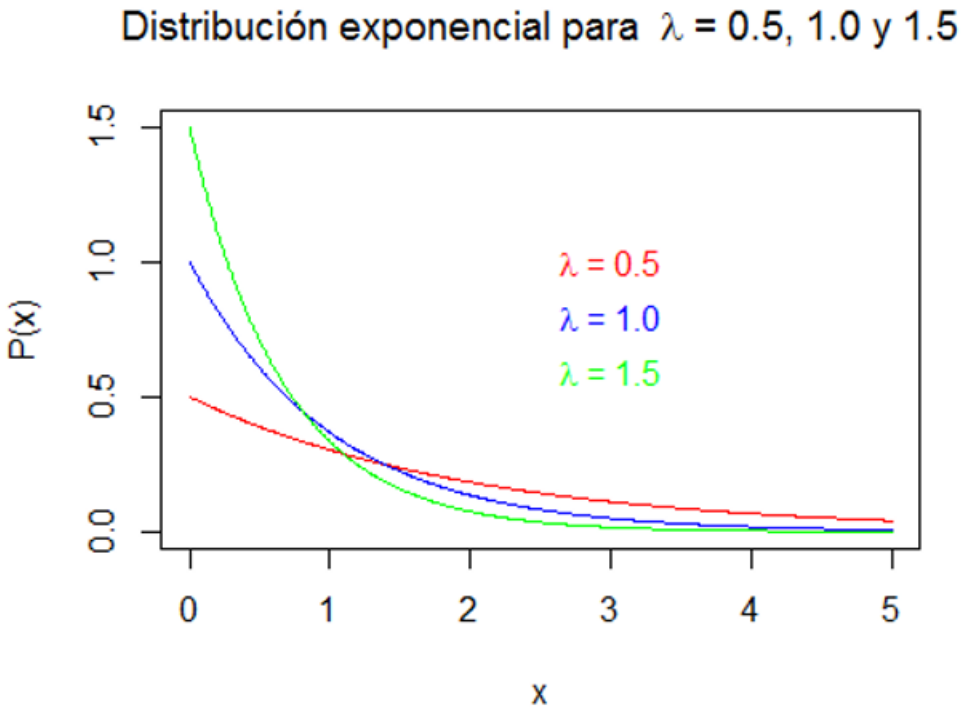
Esta distribución se emplea para determinar la probabilidad de falla a través del tiempo y se puede caracterizar básicamente por su parámetro λ. A continuación, se muestra la función que permite calcular la probabilidad de falla en el tiempo t.

$$F(t) = 1 - e^{-\lambda t}$$

El tiempo promedio de falla es $1/\lambda$.

$F(t) = 1 - e^{-\lambda t}$

El tiempo promedio de falla es $1/\lambda$.



6.4. Distribución Gama

La distribución Gama tiene parámetros k y λ, el tiempo promedio de falla se obtiene como k/λ.

6.5. Distribución Weibull

Tiene como parámetros α y λ, la probabilidad de falla en el tiempo t, puede obtenerse como:

$$F(t) = 1 - e^{\lambda t^a}$$

El tiempo promedio de falla se puede calcular a través de:

$$\frac{1}{\lambda}\Gamma(1 + \frac{1}{a})$$

6.6. Sistemas en Serie y en Paralelo

Para calcular la confiabilidad de un sistema en serie, se emplea la siguiente expresión:

$$r = (r_1)(r_2)...(r_n)$$

Para calcular la confiabilidad de un sistema en paralelo, se emplea la siguiente expresión:

$$r = 1 - (1 - r_1)(1 - r_2)...(1 - r_n)$$

En el caso de un sistema combinado se resuelve primero el sistema en paralelo y posteriormente se calcula como si fuera un sistema en serie.

6.7. Modelo Exponencial en Confiabilidad

En este caso la confiabilidad se puede obtener a través de la distribución exponencial, con ayuda de la siguiente expresión:

$$r(t) = e^{-\lambda t}$$

En el caso de una actividad específica del proceso:

$$r_i(t) = e^{-\lambda t}$$

Empleando la expresión para el cálculo de la confiabilidad en un sistema en serie:

$$r(t) = (e^{(} - \lambda_1 t))(e^{-\lambda_2 t})...(e^{-\lambda_n t}) = e^{-\lambda t}$$

donde:

$$\lambda = \lambda_1 + \lambda_2 + ... + \lambda_n$$

En el caso de que el sistema se encuentre en paralelo, tenemos:

$$r(t) = 1 - (1 - e^{-\lambda_1 t})(1 - e^{-\lambda_2 t})...(1 - e^{-\lambda_n t})$$

Bibliografías

- ✓ Canavos, George C. Probabilidad y Estadística. Aplicaciones y Métodos. Trad. Edmundo Gerardo Urbina Medal. 5ta ed. Vol. 2500. México: McGraw-Hill, 2003. Impreso. págs. 572-592.

- ✓ Gujarati, Damodar N. y Dawn C. Porter. Econometría. 5ta ed. México: McGraw-Hill, 2010. Impreso. págs.13-466 .

- ✓ Gutiérrez Pulido, Humberto y Román de la Vara S. Análisis y diseño de experimentos. 2da ed. México: McGraw-Hill, 2008. Impreso. págs. 60-235.

- ✓ Jeffrey M. Wooldridge. Introducción a la econometría. Un enfoque moderno. 4ta ed. México: McGraw-Hill, 2008. Impreso. págs. 60-235.

- ✓ Lind, Douglas A., William G. Marchal, y Samuel A. Wathen. Estadística Aplicada a Los Negocios y La Economía. 13ava ed. México: Cengage Learning, 2011. Impreso. págs. 260-367.

- ✓ Scheaffer, R., Mendenhall W., y Ott, Lyman. Elementos de Muestreo. Trad. Gilberto Rendón Sánchez, Dr. José Roberto Gómez Aguilar. 3ra ed. México: Grupo Editorial Iberoamérica, 1987. Impreso. págs. 39-122.

- ✓ Walpole, Ronald E., y Raymond H. Myers. Probabilidad Y Estadística. Trad. Gerardo Maldonado Vázquez. 4ta ed. México: McGraw-Hill, 1993. Impreso. págs. 643-675

Referencias electrónicas

- ✓ Fitting Linear Models. (n.d.). Consultado Abril 23, 2017, desde https://stat.ethz.ch/R-manual/R-devel/library/stats/html/lm.html
  Referencias:
  - Chambers, J. M. (1992) Linear models. Chapter 4 of Statistical Models in S eds J. M. Chambers and T. J. Hastie, Wadsworth & Brooks/Cole.
  - Wilkinson, G. N. and Rogers, C. E. (1973) Symbolic descriptions of factorial models for analysis of variance. Applied Statistics, 22, 392–9.

- ✓ Fit an Analysis of Variance Model. (n.d.). Consultado Abril 23, 2017, desde https://stat.ethz.ch/R-manual/R-devel/library/stats/html/aov.html
  Referencias:
  - Chambers, J. M., Freeny, A and Heiberger, R. M. (1992) *Analysis of variance; designed experiments.* Chapter 5 of *Statistical Models in S* eds J. M. Chambers and T. J. Hastie, Wadsworth & Brooks/Cole.

- ✓ Kruskal-Wallis Rank Sum Test. (n.d.). Consultado Abril 23, 2017, desde https://stat.ethz.ch/R-manual/R-devel/library/stats/html/kruskal.test.html
  Referencias:
  - Myles Hollander and Douglas A. Wolfe (1973), *Nonparametric Statistical Methods.* New York: John Wiley & Sons. Pages 115–120.

- ✓ Wilcoxon Rank Sum and Signed Rank Tests. (n.d.). Consultado Abril 23, 2017, desde https://stat.ethz.ch/R-manual/R-devel/library/stats/html/wilcox.test.html
  Referencias:
  - David F. Bauer (1972), Constructing confidence sets using rank statistics. *Journal of the American Statistical Association* **67**, 687–690.
  - Myles Hollander and Douglas A. Wolfe (1973), *Nonparametric Statistical Methods.* New York: John Wiley & Sons. Pages 27–33 (one-sample), 68–75 (two-sample). Or second edition (1999).

- ✓ Exact Binomial Test. (n.d.). Consultado Abril 23, 2017, desde https://stat.ethz.ch/R-manual/R-devel/library/stats/html/binom.test.html
  Referencias:
  - Clopper, C. J. & Pearson, E. S. (1934). The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, **26**, 404–413.
  - William J. Conover (1971), *Practical nonparametric statistics*. New York: John Wiley & Sons. Pages 97–104.
  - Myles Hollander & Douglas A. Wolfe (1973), *Nonparametric Statistical Methods.* New York: John Wiley & Sons. Pages 15–22.

- ✓ Kolmogorov-Smirnov Tests. (n.d.). Consultado Abril 23, 2017, desde https://stat.ethz.ch/R-manual/R-devel/library/stats/html/ks.test.html
  Referencias:
  - Z. W. Birnbaum and Fred H. Tingey (1951), One-sided confidence contours for probability distribution functions. *The Annals of Mathematical Statistics*, **22**/4, 592–596.
  - William J. Conover (1971), *Practical Nonparametric Statistics*. New York: John Wiley & Sons. Pages 295–301 (one-sample Kolmogorov test), 309–314 (two-sample Smirnov test).
  - Durbin, J. (1973), *Distribution theory for tests based on the sample distribution function*. SIAM.
  - George Marsaglia, Wai Wan Tsang and Jingbo Wang (2003), Evaluating Kolmogorov's distribution. *Journal of Statistical Software*, **8**/18. http://www.jstatsoft.org/v08/i18/.

- ✓ Test for Association/Correlation Between Paired Samples. (n.d.). Consultado Abril 23, 2017, desde http://astrostatistics.psu.edu/su07/R/html/stats/html/cor.test.html
  Referencias:
  - D. J. Best & D. E. Roberts (1975), Algorithm AS 89: The Upper Tail Probabilities of Spearman's *rho*. *Applied Statistics*, **24**, 377–379.
  - Myles Hollander & Douglas A. Wolfe (1973), *Nonparametric statistical inference*. New York: John Wiley & Sons. Pages 185–194 (Kendall and Spearman tests).

R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.