

Modelos no paramétricos

Sofía Villers Gómez Dulce María Reyes Varela
Carlos Fernando Vásquez Guerra

Índice general

I	Introducción	8
1.	Escalas de Medición	10
1.1.	Variables categóricas	10
1.1.1.	Escala nominal:	10
1.1.2.	Escala ordinal:	10
1.2.	Variables cuantitativas	11
1.2.1.	Escala de intervalo:	11
1.2.2.	Escala de razón:	11
II	Pruebas Binomiales	12
2.	Prueba de Proporciones	14
2.1.	Datos	14
2.2.	Supuestos	14
2.3.	Estadístico de Prueba	14
2.4.	Hipótesis	15
2.5.	Intervalos de Confianza	17
2.6.	Ejemplo	18
2.7.	Ejemplo en R-Studio	19
3.	Prueba de Cuantiles	21
3.1.	Datos	21
3.2.	Supuestos	21
3.3.	Estadístico de Prueba	21
3.4.	Hipótesis	22
3.5.	Ejemplo	24
3.6.	Ejemplo en R-Studio	25

4. Prueba de Signos	27
4.1. Datos	27
4.2. Supuestos	27
4.3. Estadístico de Prueba	27
4.4. Hipótesis	28
4.5. Ejemplo	30
4.6. Ejemplo en R-Studio	31
5. Prueba Mc Nemar	33
5.1. Datos	33
5.2. Supuestos	33
5.3. Estadístico de Prueba	34
5.4. Hipótesis	34
5.5. Ejemplo	35
5.6. Ejemplo en R-Studio	36
6. Prueba de Cox Stuart	39
6.1. Datos	39
6.2. Supuestos	39
6.3. Estadístico de Prueba	39
6.4. Hipótesis	40
6.5. Ejemplo	42
6.6. Ejemplo en R-Studio	43
7. Ejercicios	46
 III Prueba de Rango	 49
8. Prueba U-Mann y Witney	51
8.1. Datos	51
8.2. Supuestos	51
8.3. Estadístico de Prueba	52
8.4. Hipótesis	52
8.5. Ejemplo	54
8.6. Ejemplo en R-Studio	56
9. Intervalo de confianza para la diferencia entre dos medias	58
9.1. Datos	58
9.2. Supuestos	58
9.3. Método	58
9.4. Ejemplo	59

10.Prueba de Kruskal-Wallis	60
10.1. Datos	60
10.2. Supuestos	60
10.3. Hipótesis	61
10.4. Estadístico de prueba	61
10.5. Regla de decisión	62
10.6. Ejemplo	62
10.7. Ejemplo en R-Studio	65
11.Prueba de Igualdad de Varianzas	67
11.1. Datos	67
11.2. Supuestos	67
11.3. Estadístico de Prueba	68
11.4. Hipótesis	69
11.5. Ejemplo	70
11.6. Ejemplo en R-Studio	72
12.Prueba para más de dos Muestras	76
12.1. Datos	76
12.2. Hipótesis	76
12.3. Estadístico de Prueba	77
12.4. Regla de decisión	77
12.5. Comparación múltiple	77
12.6. Ejemplo	78
12.7. Ejemplo en R-Studio	79
13.Ejercicios	81
 IV Tablas de Contingencia	 84
14.Tablas de Contingencia de 2x2	86
14.1. Datos	86
14.2. Supuestos	86
14.3. Estadístico de Prueba	87
14.4. Hipótesis	87
14.5. Ejemplo	89
14.6. Ejemplo en R-Studio	90

15.Prueba de Independencia	92
15.1. Datos	92
15.2. Supuestos	92
15.3. Estadístico de Prueba	92
15.4. Hipótesis	93
15.5. Ejercicio	93
15.6. Ejemplo en R-Studio	95
16.Tablas de Contingencia de $r \times c$	97
16.1. Datos	97
16.2. Supuestos	98
16.3. Estadístico de Prueba	98
16.4. Hipótesis	98
16.5. Ejercicio	99
16.6. Ejemplo en R-Studio	100
17.Prueba de la Mediana	102
17.1. Datos	102
17.2. Supuestos	102
17.3. Estadístico de Prueba	103
17.4. Hipótesis	103
17.5. Comparación Múltiple	103
17.5.1. Ejemplo	104
17.6. Ejemplo en R-Studio	105
18.Ejercicios	108
V Bondad de Ajuste	110
19.Prueba de la Ji-cuadrada	112
19.1. Datos	112
19.2. Hipótesis	112
19.3. Estadístico de Prueba	113
19.4. Ejemplo	113
19.5. Ejemplo en R-Studio	114
20.Prueba Kolmogorov	117
20.1. Datos	118
20.2. Supuestos	118
20.3. Estadístico de Prueba	119
20.4. Hipótesis	119

20.5. Ejemplo	120
20.6. Ejemplo en R-Studio	123
20.7. Otro ejemplo en R	124
21.Prueba Kolmogorov-Smirnov	125
21.1. Hipótesis	125
21.1.1. Caso A (Prueba de 2 colas)	125
21.2. Ejemplo	126
22.Prueba Lilliefors para Normalidad	133
22.1. Datos	133
22.2. Supuestos	134
22.3. Hipótesis	134
22.4. Estadístico de Prueba	134
22.5. Ejemplo	134
22.6. Ejemplo en R-Studio	140
23.Pueba de Lilliefors Exponencial	143
23.1. Datos	143
23.2. Supuestos	143
23.3. Hipótesis	143
23.4. Estadístico de Prueba	144
23.5. Ejemplo	144
23.6. Ejemplo en R-Studio	149
24.Prueba Anderson-Darling	151
24.1. Datos	151
24.2. Supuestos	151
24.3. Hipótesis	151
24.4. Estadístico de Prueba	152
24.5. Regla de Decisión	152
24.5.1. Ejemplo	152
24.6. Ejemplo en R-Studio	156
25.Otras estadísticas	158
26.Ejercicios	160

Prefacio

El curso de Modelos No Paramétricos y de Regresión es el segundo curso de estadística en el mapa curricular de la carrera de actuaría en la facultad, por lo que al llegar a este curso es muy probable que ya hayan cursado inferencia estadística en donde se revisan temas de estimación de parámetros (estimación puntual, estimación por intervalos, pruebas de hipótesis sobre las estimaciones de dichos parámetros). Se puede notar que las metodologías estudiadas en el curso de inferencia se basan en el supuesto de que conocemos la distribución que sigue nuestra variable de interés y por lo tanto las observaciones provienen de una cierta familia paramétrica de distribuciones y entonces con las observaciones (muestra) podemos hacer inferencia estadística acerca de los valores de los parámetros de dicha distribución.

Sin embargo, hay casos en donde desconocemos la distribución paramétrica de la que provienen nuestros datos y aún así quisiéramos poder realizar algún tipo de inferencia sobre la variable de interés o de la población. ¿Se puede?

La respuesta es sí, pero utilizando procedimientos distintos o realizando algunas variaciones a las metodologías vistas en el curso de inferencia estadística. Esos son los temas que se revisan en la primer parte de este curso denominado modelos no paramétricos.

Objetivos

- Proporcionar a los alumnos el material para cubrir los temas del curso de modelos no paramétricos y de regresión.
- Reforzar las bases teóricas con contenido electrónico completado con herramientas de R-Studio.

Este libro fue escrito con bookdown usando RStudio.

Esta versión fue escrita con:

Licencia

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

This is a human-readable summary of (and not a substitute for) the license. Please see <https://creativecommons.org/licenses/by-sa/4.0/legalcode> for the full legal text.

You are free to:

- **Share**—copy and redistribute the material in any medium or format
- **Remix**—remix, transform, and build upon the material for any purpose, even commercially.

The licensor cannot revoke these freedoms as long as you follow the license terms.

Under the following terms:

- **Attribution**—You must give appropriate credit, provide a link to the license, and indicate if changes were made. You may do so in any reasonable manner, but not in any way that suggests the licensor endorses you or your use.

- **ShareAlike**—If you remix, transform, or build upon the material, you must distribute your contributions under the same license as the original.
- **No additional restrictions**—You may not apply legal terms or technological measures that legally restrict others from doing anything the license permits.

Notices:

You do not have to comply with the license for elements of the material in the public domain or where your use is permitted by an applicable exception or limitation.

No warranties are given. The license may not give you all of the permissions necessary for your intended use. For example, other rights such as publicity, privacy, or moral rights may limit how you use the material.

Parte I

Introducción

Introducción

Los métodos paramétricos tienen la característica de que siempre se trabaja con muestras aleatorias provenientes de cierta distribución conocida. Por otro lado, los métodos no paramétricos suelen requerir suposiciones menos restrictivas acerca del nivel de medición de los datos y sobre la forma de las distribuciones de probabilidad generadas por los datos muestrales, por lo que se pueden usar en situaciones en donde los métodos paramétricos no son aplicables.

Una de las consideraciones para determinar si lo apropiado es un método paramétrico o un método no paramétrico es la escala de medición empleada para generar los datos. Todos los datos son generados por una de las cuatro escalas de medición: nominal, ordinal, de intervalo o de razón.

Capítulo 1

Escalas de Medición

Esta sección es una traducción de Conover (1998).

A continuación se definirán los tipos de escalas que se le asigna a una variable. Hay cuatro escalas básicas las cuales son: **nominal**, **ordinal**, **de intervalo** y **de razón**. A las variables que se consideran con escala ordinal y nominal también se les conoce como **variables categóricas o cualitativas**, mientras que a las variables que se consideran con escala de intervalo y razón se conocen como **variables cuantitativas**.

1.1. Variables categóricas

1.1.1. Escala nominal:

La escala nominal utiliza números simplemente como un medio para separar las propiedades o elementos en diferentes clases o categorías. El número asignado a la observación sirve sólo como un “**nombre**” para la categoría a la que pertenece la observación, de ahí el título “**nominal**”. Por ejemplo, utilizamos la escala nominal de medición cuando definimos una variable aleatoria que equivalía a 1 si una moneda caía “**sol**”, y 0 si la moneda caía “**águila**”. Podríamos, igual bien, haber usado los números **7.3** y **3.9** para representar el sol y el águila, respectivamente. Nuestra elección de 1 y 0 fue principalmente por conveniencia. Otro ejemplo, cuando 12 sujetos se numeran arbitrariamente del 1 al 12, se utiliza una escala de medición nominal y la asignación de los números es una forma de variable aleatoria. Al clasificar objetos según el color, las categorías pueden etiquetarse como 1, 2, 3 o azul, amarillo, rojo o A, B, C. Los números son simplemente nombres de categoría. Los números pueden ser reemplazados por otros números no utilizados, siempre que las categorías permanezcan intactas.

1.1.2. Escala ordinal:

La escala ordinal se refiere a mediciones en las que sólo son relevantes las comparaciones “**mayor**”, “**menor**” o “**igual**” entre ellas. El valor numérico se usa sólo como un medio para organizar los elementos que se miden en orden, de menor a mayor. Ésta capacidad de ordenar los elementos, en función del tamaño relativo de sus medidas, le da el nombre de Si algunos de los elementos son iguales entre sí, decimos que existen vínculos. Cuando se le pide a una persona que asigne el número 1 a la más preferida de las tres marcas, el número 3 a la menos preferida y el número 2 a la marca restante, está usando una escala ordinal y está usando los números 1, 2, 3; ésta podría haber usado tres números, digamos 16, 20, 75, siempre y cuando los números se asignen a las marcas de tal manera que el orden relativo del número represente la preferencia relativa de la marca.

1.2. Variables cuantitativas

1.2.1. Escala de intervalo:

La escala de intervalo, considera como información pertinente **no sólo** el orden relativo de las mediciones como en la escala ordinal sino también el **tamaño del intervalo** entre mediciones, es decir, el tamaño de la diferencia (en un sentido de resta) entre dos mediciones. La escala de intervalo implica el concepto de una unidad de distancia, y la distancia entre dos mediciones cualquiera puede expresarse como un cierto número de unidades. Un buen ejemplo es la escala por la cual generalmente representamos la temperatura. El aumento de una unidad (grado) de temperatura se define por un cambio particular en el volumen de mercurio en un termómetro; en consecuencia, la diferencia entre dos temperaturas cualquiera puede medirse en **unidades o grados**. El valor numérico real de la temperatura es simplemente una comparación con un punto arbitrario llamado “cero grados”. La escala de intervalo requiere un punto cero y una unidad de distancia (no es posible tener este último sin el primero), pero no es importante cómo se definan los “ceros” y la unidad de distancia.

La temperatura ha sido medida de manera bastante adecuada durante algún tiempo por las escalas Fahrenheit y Celsius, que tienen diferentes temperaturas cero y diferentes definiciones de 1 grado o unidad. El principio de las mediciones de intervalo no se viola por un cambio en la escala o la ubicación o ambos.

1.2.2. Escala de razón:

La escala de razón se usa cuando **no sólo el orden y el tamaño del intervalo son importantes**, sino que también la **razón entre dos medidas** es significativa. Si es razonable hablar de que una cantidad es “dos veces” otra cantidad, la escala de proporción es apropiada para la medición, como cuando se miden los rendimientos de cultivos, distancias, pesos, alturas, ingresos, etc. En realidad, la única distinción entre la escala de razón y la escala de intervalo es que la escala de razón tiene una **medida natural** “cero”, mientras que la medida de cero se define **arbitrariamente** en la escala de intervalo. Como en la escala de intervalo, la unidad de distancia de la escala de razón se define arbitrariamente.

No existe un acuerdo universal entre los científicos que prefieren usar escalas adicionales, y algunas mediciones no se incluyen claramente en una de las cuatro escalas recién definidas. Por lo tanto, ésta clasificación de escalas de medida puede ser demasiado simplista, pero es suficiente para los propósitos de este curso.

La mayoría de los métodos estadísticos paramétricos habituales requieren una escala de medición de intervalo (o más fuerte). La mayoría de los métodos no paramétricos suponen que la escala nominal o la escala ordinal son apropiadas.

Parte II

Pruebas Binomiales

Introducción

La distribución de probabilidad binomial se introdujo para describir las probabilidades asociadas con el número de caras cuando se lanza una moneda n veces. En su forma más general, cada uno de n ensayos independientes da como resultado "*éxito*", con probabilidad p , o "*fracaso*", con probabilidad $q = 1 - p$. La distribución binomial describe la probabilidad de obtener exactamente " k " éxitos.

Capítulo 2

Prueba de Proporciones

Supongamos que un científico desea saber si la tasa de mortalidad va en aumento, si la tasa de pobreza está cambiando o si la tasa de algún grupo cívico está a favor de una política en particular. En muchos casos existe una proporción hipotética p de una población en estudio y una proporción específica p^* y se pretende llevar a cabo una comparación para saber si la proporción hipotética es igual, menor o mayor que la proporción específica p^* . Una prueba de proporciones puede ser útil para ayudar a responder este tipo de preguntas.

2.1. Datos

Los datos consisten en una muestra

$X_1, X_2, X_3, \dots, X_n$ una muestra aleatoria de tamaño n de la población.

cuyo resultado se puede clasificar en dos y sólo dos categorías “categoría 1” o “categoría 2”. El número de observaciones en la categoría 1 es O_1 y el número de observaciones en la categoría 2 es $O_2 = n - O_1$

2.2. Supuestos

- 1) Los n ensayos son mutuamente independientes.
- 2) La probabilidad p de que el resultado de cada ensayo caiga en la categoría 1 es la misma en cada uno de los ensayos.

2.3. Estadístico de Prueba

Como nos preocupa la probabilidad del resultado “clase 1”, dejaremos que el estadístico de prueba T sea el número de veces que el resultado es “**clase 1**”. Es decir, si $n < 20$ utilizar el estadístico

$$T = O_1$$
$$T \sim \text{Bin}(n, p^*)$$

Donde p^* es la probabilidad especificada en la hipótesis nula de nuestra prueba a realizar y n es el tamaño de la muestra.

Por otro lado, si $n \geq 20$ puede resultar más sencillo utilizar una aproximación normal para realizar la prueba, en dicho caso se puede utilizar el siguiente cuantil:

$$t = np + Z_q \sqrt{np(1-p)}$$

Donde Z_q es el cuantil de una distribución normal estándar que se puede obtener en la tabla correspondiente¹.

Dependiendo del planteamiento de nuestro problema a resolver se formulan las hipótesis:

2.4. Hipótesis

Caso A (Prueba de dos colas)

$$\mathbf{H}_0 : p = p^*$$

vs

$$\mathbf{H}_a : p \neq p^*$$

Regla de decisión

Rechazamos H_0 a un nivel de significancia α si

$$T \leq t_1 \quad \text{o} \quad T > t_2$$

Elegimos α_1 y α_2 , los tamaños de la cola inferior y superior, respectivamente. El tamaño de la prueba es $\alpha = \alpha_1 + \alpha_2$.

Debemos encontrar t_1 y t_2 tales que:

$$\mathbf{P}[Y \leq t_1] = \alpha_1 \quad \mathbf{P}[Y > t_2] = \alpha_2$$

Donde $Y \sim \text{Bin}(n, p^*)$.

y calculamos el p -value de la siguiente manera:

$$p\text{-value} = 2 * \min\{\mathbf{P}[Y \leq T], \mathbf{P}[Y \geq T]\}$$

Sugerimos que si $n > 20$, el p -value puede obtenerse usando:

$$\mathbf{P}[Y \leq t_{obs}] \approx \mathcal{N}\left(\frac{t_{obs} - np^* + 0.5}{\sqrt{np^*(1-p^*)}}\right)$$

y

$$\mathbf{P}[Y \geq t_{obs}] \approx 1 - \mathcal{N}\left(\frac{t_{obs} - np^* - 0.5}{\sqrt{np^*(1-p^*)}}\right)$$

Con $\mathcal{N}(\cdot)$ la función de distribución de la normal estándar.

¹Véase que de acuerdo al tipo de cuantil (de cola inferior o superior), el signo del cuantil Z_q cambiará.

Caso B (Prueba de cola inferior)

$$\mathbf{H}_0 : p \geq p^*$$

vs

$$\mathbf{H}_a : p < p^*$$

Regla de decisión

Rechazamos H_0 a un nivel de significancia α si

$$T \leq t$$

Elegimos α , el tamaño de la prueba y debemos encontrar t tal que

$$\mathbf{P}[Y \leq t] = \alpha$$

Donde $Y \sim \text{Bin}(n, p^*)$.

y calculamos el $p - \text{value}$ de la siguiente manera:

$$p - \text{value} = \mathbf{P}[Y \leq T]$$

Sugerimos que si $n > 20$, el $p - \text{value}$ puede obtenerse usando:

$$\mathbf{P}[Y \leq t_{obs}] \approx \mathcal{N}\left(\frac{t_{obs} - np^* + 0.5}{\sqrt{np^*(1-p^*)}}\right)$$

Caso C (Prueba de cola superior)

$$\mathbf{H}_0 : p \leq p^*$$

vs

$$\mathbf{H}_a : p > p^*$$

Regla de decisión

Rechazamos H_0 a un nivel de significancia α si

$$T > t$$

Elegimos α , el tamaño de la prueba y debemos encontrar t tal que

$$\mathbf{P}[Y \leq t] = 1 - \alpha$$

Donde $Y \sim \text{Bin}(n, p^*)$.

y calculamos el $p - value$ de la siguiente manera:

$$p - value = \mathbf{P}[Y \geq T]$$

Sugerimos que si $n > 20$, el $p - value$ puede obtenerse usando:

$$\mathbf{P}[Y \geq t_{obs}] \approx \mathcal{N}\left(\frac{t_{obs} - np^* - 0.5}{\sqrt{np^*(1-p^*)}}\right)$$

2.5. Intervalos de Confianza

En este enfoque lo que se busca es estimar cotas inferiores y superiores (L, U) tales que el intervalo formado por las mismas contenga al parámetro de interés (en este caso una proporción) con una confianza $1 - \alpha$ especificada por el usuario.

Datos

Los datos a los que se les puede aplicar éste método son del mismo que en las pruebas de hipótesis, es decir, una colección $\{X_i\}_{i=1}^n$ de ensayos Bernoulli independientes en dónde se debe poder suponer que la probabilidad de éxito p se mantiene constante en todos los ensayos.

Supuestos

- 1) Los n ensayos son independientes.
- 2) La probabilidad de éxito permanece constante durante todos los ensayos.

La idea detrás del método es encontrar todos los valores de p^* tales que rechazamos la hipótesis nula.

Hipótesis

$$\mathbf{H}_0 : p = p^*$$

vs

$$\mathbf{H}_a : p \neq p^*$$

Procedimiento

- 1) Fijar el nivel de confianza $1 - \alpha$
- 2) Hallar p_1 y p_2 tales que:

$$\mathbf{P}[Y \leq O_1 | p = p_1] = \alpha_1 \quad y \quad \mathbf{P}[Y \geq O_1 | p = p_2] = \alpha_2$$

donde $\alpha = \alpha_1 + \alpha_2$ y $Y \sim \text{Bin}(n, p)$.

- 3) Hacer $L = p_1$ y $U = p_2$

Recordemos que mucha de la teoría abarcada en este material esta basado en el libro Conover (1998) y en dicha referencia se pueden encontrar tablas con los valores de los intervalos de confianza para muestras menores a 30, cuando se tienen más grandes, es decir si $n > 30$, se pueden utilizar las siguientes expresiones para el cálculo de los intervalos.

$$L = \frac{O_1}{n} - Z_{1-\alpha/2} \sqrt{\frac{O_1(n-O_1)}{n^3}} \quad U = \frac{O_1}{n} + Z_{1-\alpha/2} \sqrt{\frac{O_1(n-O_1)}{n^3}}$$

Ahora aplicaremos lo anterior en un ejemplo ilustrativo:

2.6. Ejemplo

Se tienen 20 graduados del Tecnológico de Texas que presentaron el examen general de leyes y 18 de ellos lo pasaron. Si esta muestra es aleatoria y representativa de todos los estudiantes graduados del Tecnológico de Texas, ¿esto prueba que la probabilidad de que un graduado de esa escuela pase el examen general de leyes es más alto que el promedio del estado, que es del 70 %?

Paso 1 Escribimos la prueba a utilizar

La prueba a utilizar **Prueba de proporciones caso C cola superior**

Paso 2 Formulamos nuestras hipótesis en contexto al problema planteado

H_0 : La probabilidad de que un graduado pase el examen es menor/igual al 70 %.

vs

H_a : La probabilidad de que un graduado pase el examen es mayor al 70 %.

De manera alternativa:

$H_0 : p \leq p^*$ es decir, $p \leq .70$ *vs* $H_a : p > p^*$ es decir, $p > .70$

Paso 3 Estadístico de prueba

Utilizaremos el estadístico

$T = O_1$ número de observaciones de la clase 1.

$T = 18$ número de graduados que pasaron el examen

$T \sim \text{Bin}(20, 0.70)$

Paso 4 Procedimiento completo

Supuestos:

1. Muestra aleatoria de tamaño 20.
2. Tomaremos como “éxito” al acreditar el examen.
 - $T = O_1 = 18$ número de éxitos.
3. Tomaremos $\alpha = 5\% = 0.05$ el nivel de significancia.
 - $n = 20$ tamaño de la muestra
 - $p^* = 70\% = 0.70$

Paso 5 Regla de decisión

Rechazo H_0 si $T > t_2$ y Rechazo H_0 si $p\text{-value} < \alpha$

ya que $n \geq 20$ puede resultar más sencillo utilizar una aproximación normal para realizar la prueba, en dicho caso se podemos utilizar:

$$t_2 = np^* + Z_q \sqrt{np^*(1-p^*)}$$

$$t_2 = 20 * (0.7) + 1.65 * \sqrt{20 * (0.7) * (0.3)} = 17.38$$

\therefore como $T = 18 > 17 = t_2$ entonces rechazo H_0 .

y por otro lado calculamos el p – *value* de la siguiente manera:

$$p - value = \mathbf{P}[Y \geq T] = \mathbf{P}[Y \geq 18] = 1 - \mathbf{P}[Y < 18] = 0.035$$

\therefore como $p - value = 0.035 < \alpha = 0.05$ entonces rechazo H_0 .

Paso 6 Conclusión

Existe información suficiente para decir que los alumnos que acreditaron el examen están por arriba del promedio del Estado que es del 70 %, en otras palabras la probabilidad de que un graduado del Tecnológico de Texas pase el examen general de leyes es mayor al 70 %.

2.7. Ejemplo en R-Studio

Ahora haremos la réplica en R.

La estadística de prueba será $T = O_1$

```
# Datos

T=18                                #Número de éxitos
alpha=0.05                          #Nivel de significancia
n=20                                #Tamaño de la muestra
p=0.70                              #Proporción

t=qbinom(.95,n,p)                   #Valor crítico
t

[1] 17

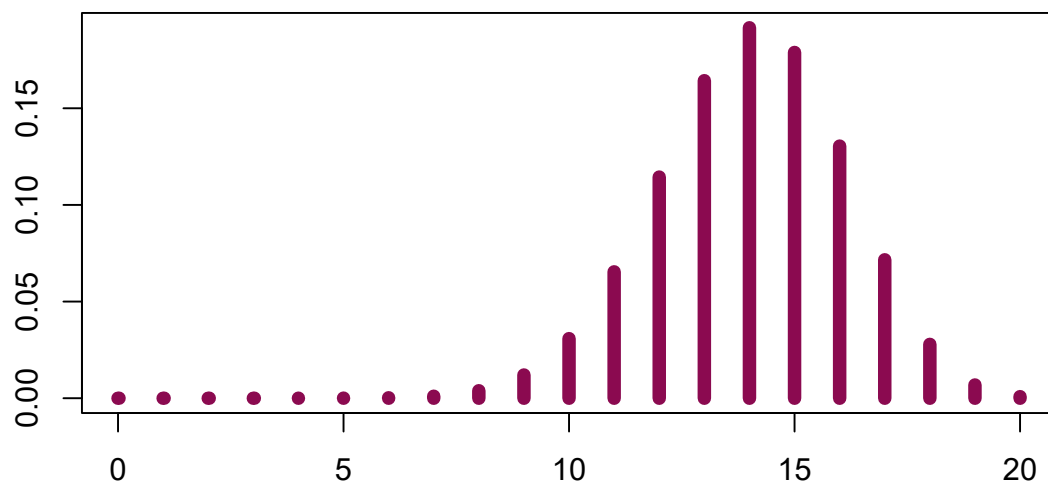
pvalue=1-pbinom(17,20,.7)           #P-value
#Otra opción sería: 1-pbinom(17,20,.7, lower.tail = F)
pvalue
```

```
[1] 0.03548313
```

Según el planteamiento de las hipótesis, este es un Caso C (de cola superior), por lo que siguiendo la regla de decisión tenemos que como $T = 18 > 17 = t$, entonces se rechaza H_0 y por lo tanto se concluye que hay información suficiente para decir que la probabilidad de que un graduado del tecnológico de Texas pase el examen es mayor al 70 %.

Graficamos la función binomial

Función de distribución de una variable Binomial(20,0.7)



Utilizando la función de R, debemos tener cuidado ya que nuestra muestra es pequeña:

```
binom.test(T,n, p = 0.7, alternative = c("greater"))
```

Exact binomial test

data: T and n

number of successes = 18, number of trials = 20, p-value = 0.03548

alternative hypothesis: true probability of success is greater than 0.7

95 percent confidence interval:

0.7173815 1.0000000

sample estimates:

probability of success

0.9

Capítulo 3

Prueba de Cuantiles

Lo que nos interesa en ésta prueba binomial es hacer inferencia sobre los cuantiles de una variable aleatoria. Por ejemplo, muchos examinamos una muestra aleatoria de valores de alguna variable x para ver si la mediana de x es igual a 17 (por ejemplo).

La escala de medición suele ser al menos ordinal para ésta prueba de hipótesis, aunque la prueba binomial solo requiere la escala nominal que es más débil para su medición. Esto se debe a que los cuantiles tienen poco significado con las mediciones de escala nominal.

Recordatorio

El cuantil de valor p de una variable aleatoria X es un número x tal que:

$$\mathbf{P}[X < x] \leq p$$

- En el caso de una variable aleatoria continua se da la igualdad.

3.1. Datos

Los datos consisten en una muestra

$$X_1, X_2, X_3, \dots, X_n \quad \text{una muestra aleatoria de tamaño } n \text{ de la población}$$

3.2. Supuestos

- 1) Los datos a los que se les aplica esta prueba son una muestra aleatoria independiente e idénticamente distribuida.
- 2) La escala de medida de las X_i s es al menos ordinal.

3.3. Estadístico de Prueba

$$T_1$$

El número de observaciones en la muestra que son **menores o iguales** al valor x^* sobre el que se va a hacer la hipótesis.

$$T_2$$

El número de observaciones en la muestra que son **estrictamente menores** a x^* .

Cuando $n \leq 20$ la distribución nula de éstos estadísticos es $Bin(n, p = p^*)$ con n el tamaño de la muestra y p^* dada en la hipótesis nula (recuerde que queremos hacer pruebas relacionadas con el cuantil x_p de la *v.a.* en cuestión).

Por otro lado, si $n > 20$ puede resultar más sencillo utilizar una aproximación normal para realizar la prueba, en dicho caso se puede utilizar el cuantil:

$$t = np + Z_q \sqrt{np(1-p)}$$

Donde Z_q es el cuantil de una distribución normal estándar que se puede obtener en la tabla correspondiente.

Dependiendo del planteamiento de nuestro problema a resolver se formulan las hipótesis:

3.4. Hipótesis

Caso A (Prueba de dos colas)

$$\mathbf{H}_o : x_p = x^* \text{ es equivalente decir : } \mathbf{P}[X \leq x^*] = p^*$$

vs

$$\mathbf{H}_a : x_p \neq x^* \text{ es equivalente decir : } \mathbf{P}[X \leq x^*] \neq p^*$$

Regla de decisión

Rechazamos H_0 a un nivel de significancia α si

$$T_1 \leq t_1 \quad o \quad T_2 > t_2$$

Elegimos $\alpha_1, \alpha_2 \geq 0$, tales que $\alpha_1 + \alpha_2 = \alpha$ el tamaño de la prueba y debemos encontrar t_1 y t_2 tales que:

$$\mathbf{P}[Y \leq t_1] = \alpha_1 \quad y \quad \mathbf{P}[Y \leq t_2] = 1 - \alpha_2$$

Donde $Y \sim Bin(n, p^*)$.

y calculamos el *p-value* de la siguiente manera:

$$p - value = 2 * \min\{\mathbf{P}[Y \leq T_1], \mathbf{P}[Y \geq T_2]\}$$

Sugerimos que si $n > 20$, el *p-value* puede obtenerse usando:

$$\mathbf{P}[Y \leq T_1] \approx \mathcal{N}\left(\frac{T_1 - np^* + 0.5}{\sqrt{np^*(1-p^*)}}\right)$$

y

$$\mathbf{P}[Y \geq T_2] \approx 1 - \mathcal{N}\left(\frac{T_2 - np^* - 0.5}{\sqrt{np^*(1-p^*)}}\right)$$

Con $\mathcal{N}(\cdot)$ la función de distribución de la normal estándar.

Caso B (Prueba de cola inferior)

$$\mathbf{H}_0 : x_p \leq x^* \text{ es equivalente a decir: } \mathbf{P}[X < x^*] \geq p^*$$

vs

$$\mathbf{H}_a : x_p > x^* \text{ es equivalente a decir : } \mathbf{P}[X < x^*] < p^*$$

Regla de decisión

Rechazamos H_0 a un nivel de significancia α si

$$T_1 \leq t_1$$

Elegimos α , el tamaño de la prueba y debemos encontrar t_1 tal que

$$\mathbf{P}[Y \leq t_1] = \alpha$$

Donde $Y \sim \text{Bin}(n, p^*)$.

y calculamos el $p - \text{value}$ de la siguiente manera:

$$p - \text{value} = \mathbf{P}[Y \leq T_1]$$

Sugerimos que si $n > 20$, el $p - \text{value}$ puede obtenerse más sencillo usando la aproximación normal:

$$\mathbf{P}[Y \leq T_1] \approx \mathcal{N}\left(\frac{T_1 - np^* + 0.5}{\sqrt{np^*(1-p^*)}}\right)$$

Caso C (Prueba de cola superior)

$$\mathbf{H}_0 : x_p \geq x^* \text{ es equivalente a decir : } \mathbf{P}[X < x^*] \leq p^*$$

vs

$$\mathbf{H}_a : x_p < x^* \text{ es equivalente a decir : } \mathbf{P}[X < x^*] > p^*$$

Regla de decisión

Rechazamos H_0 a un nivel de significancia α si

$$T_2 > t_2$$

Elegimos α , el tamaño de la prueba y debemos encontrar t_2 tal que

$$\mathbf{P}[Y \leq t_2] = 1 - \alpha$$

Donde $Y \sim \text{Bin}(n, p^*)$.

y calculamos el p -value de la siguiente manera:

$$p\text{-value} = \mathbf{P}[Y \geq T_2]$$

Sugerimos que si $n > 20$, el p -value puede obtenerse más sencillo usando la aproximación normal:

$$\mathbf{P}[Y \geq T_2] \approx 1 - \mathcal{N}\left(\frac{T_2 - np^* - 0.5}{\sqrt{np^*(1-p^*)}}\right)$$

Ahora aplicaremos lo anterior en un ejemplo ilustrativo:

3.5. Ejemplo

El intervalo de tiempo entre las erupciones del geiser Old Faithful se registra 112 veces, de las cuales 8 son menores a 60 minutos y una es exactamente 60 minutos. Se quiere desea verificar que la mediana del intervalo es mayor a 60 minutos.

Paso 1 Escribimos la prueba a utilizar

La prueba a utilizar **Prueba de cuantiles caso B cola inferior**

Paso 2 Formulamos nuestras hipótesis en contexto al problema planteado

$$\mathbf{H}_0 : \mathbf{P}[X \leq 60] \geq 0.50$$

vs

$$\mathbf{H}_a : \mathbf{P}[X \leq 60] < 0.50$$

Donde X es el intervalo de tiempo entre las erupciones, asumiendo que ambos intervalos son independientes e idénticamente distribuidas.

Paso 3 Estadístico de prueba

La estadística de prueba será $T_1 = 9$ y $T_2 = 8$. Tomaremos $\alpha = .1$

$T_1 = 9$ número de intervalos que son menores o iguales a 60 minutos.

$T_2 = 8$ número de intervalos que son estrictamente menores a 60 minutos.

$$T_1 \sim \text{Bin}(112, 0.50)$$

Paso 4 Procedimiento completo

Supuestos:

1. Muestra aleatoria de tamaño 112.
 - Tomaremos $\alpha = 5$ el nivel de significancia.
 - $n = 112$ tamaño de la muestra
 - $p^* = 0.50$

Paso 5 Regla de decisión

Rechazamos H_0 a un nivel de significancia α si $T_1 \leq t_1$ y Rechazo H_0 si $p - value < \alpha$

$$T_1 = 9 \leq 47.3 = t_1$$

como $n > 20$ puede resultar más sencillo utilizar una aproximación normal para realizar la prueba, en dicho caso se puede utilizar el estadístico:

$$t_1 = np + Z_q \sqrt{np(1-p)} = (112)(0.5) + (-1.645) \sqrt{112(0.5)(1-0.5)}$$

$$t_1 = 47.3$$

\therefore Rechazamos H_0

y calculamos el $p - value$ de la siguiente manera:

$$p - value = \mathbf{P}[Y \leq T_1] = \mathbf{P}[Y \leq 9]$$

Como $n > 20$, el $p - value$ puede obtenerse más sencillo usando la aproximación normal:

$$\begin{aligned} \mathbf{P}[Y \leq 9] &\approx \mathcal{P} \left[Z \leq \frac{T_1 - np^* + 0.5}{\sqrt{np^*(1-p^*)}} \right] = \\ &\left(Z \leq \frac{9 - (112)(0.5) + 0.5}{\sqrt{(112)(0.5)(1-0.5)}} \right) = \mathbf{P}[Z \leq -8.7876] \approx 0.0001 \end{aligned}$$

\therefore Rechazo H_0 ya que $p - value = 0.0001 < 0.05 = \alpha$. es decir, el $p - value$ es “muy pequeño”.

Paso 6 Conclusión

Existe información suficiente para decir que la mediana del intervalo de erupciones es mayor a 60 minutos.

3.6. Ejemplo en R-Studio

Ahora haremos la réplica en R.

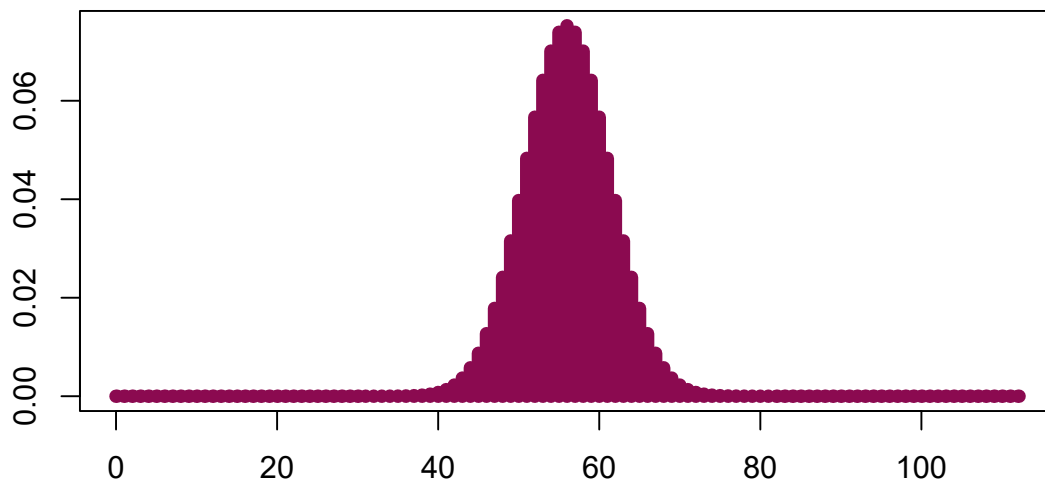
La estadística de prueba será $T_1 = 9$ y $T_2 = 8$. Tomaremos $\alpha = .1$

```
#Datos
T_1=9          #Observaciones menores/iguales a 60 minutos
T_2=8          #Observaciones que son menores estrictamente a 60 minutos
alpha=0.05     #Nivel de significancia
n=112          #Tamaño de la muestra
p=0.5          #Cuantil en este caso la mediana
```

Según el planteamiento de las hipótesis, este es un Caso B (de cola inferior), por lo que siguiendo la regla de decisión, se rechaza H_0 si $T_1 \leq t_1$ donde t_1 será el cuantil que acumule 10 % en la distribución binomial

Podemos graficar la función de distribución:

Función de distribución de una variable Binomial(112,0.5)



A continuación calculamos t_1 y el p -value:

```
t_1=qbinom(.10,n,p)      #Cuantil a comparar con el estadístico de prueba
t_1
```

```
[1] 49
```

```
pvalue=pbinom(T_1,n,p)  #P-value correspondiente
pvalue
```

```
[1] 1.157256e-21
```

Tenemos que como $T_1 = 9 \leq 49 = t_1$, entonces se rechaza H_0 y por lo tanto se concluye que hay información suficiente para decir que la mediana de los intervalos de tiempo entre erupciones es mayor a 60 minutos.

Al mismo resultado llegamos si en lugar de buscar t_1 calculamos el p -value de la estadística T_1 la cual da un valor muy cercano a cero por lo tanto cae en la región de rechazo y se concluye que la mediana de los intervalos de tiempo entre erupciones es mayor a 60 minutos.

Finalmente podemos utilizar la función en R.

```
binom.test(T_1,n,p=0.5,alternative = "less")
```

```
Exact binomial test
```

```
data: T_1 and n
number of successes = 9, number of trials = 112, p-value < 2.2e-16
alternative hypothesis: true probability of success is less than 0.5
95 percent confidence interval:
 0.000000 0.136034
sample estimates:
probability of success
 0.08035714
```

Capítulo 4

Prueba de Signos

La prueba de signos merece una consideración especial debido a su versatilidad, su gran utilidad y simplicidad. Ésta es una prueba de proporciones cuando el valor específico $p^* = 1/2$ y para los casos de dos colas, cola inferior y cola superior la máxima probabilidad para rechazar la hipótesis nula H_0 se da cuando $p = 1/2$. Frecuentemente la prueba de signos también es apropiada para analizar datos de un vector aleatorio (X, Y) y ver si alguna de sus entradas tiene valores más grandes que la otra. De esta manera, si una variable tiende a tener valores mayores que la otra, se puede utilizar la prueba de signos para determinar si las medias de éstas variables son diferentes.

4.1. Datos

Los datos consisten en observaciones bivariadas aleatorias:

$$(X_1, Y_1), (X_2, Y_2), (X_3, Y_3), \dots, (X_{n'}, Y_{n'})$$

Dentro de cada par de datos en la muestra se clasificará de la siguiente manera:

- Por un signo “+” cuando $X_i < Y_i$
- Por un signo “-” cuando $X_i > Y_i$
- Se omitirán las parejas cuando $X_i = Y_i$.

El tamaño de la muestra después de quitar los empates será n .

4.2. Supuestos

- 1) Las variables aleatorias bivariadas (X_i, Y_i) son mutuamente independientes.
- 2) La escala de medida es al menos ordinal dentro de cada par.

4.3. Estadístico de Prueba

$$T = \text{Total de signos " + "}$$

La distribución nula de T es una distribución binomial con n el número de parejas de la muestra sin empates y $p = 1/2$.

$$T \sim \text{Bin}(n, 1/2)$$

Dependiendo del planteamiento de nuestro problema a resolver se formulan las hipótesis:

4.4. Hipótesis

Caso A (Prueba de dos colas)

$$H_0 : P[\text{obtener } +] = P[\text{obtener } -]$$

vs

$$H_a : P[\text{obtener } +] \neq P[\text{obtener } -]$$

Regla de decisión

Rechazamos H_0 a un nivel de significancia α si

$$T \leq t \quad \text{o} \quad T > n - t$$

Elegimos α , el tamaño de la prueba y debemos encontrar t tal que

$$P[Y \leq t] = \alpha/2$$

Donde $Y \sim \text{Bin}(n, 1/2)$.

Por otro lado, si $n > 20$ puede resultar más sencillo utilizar una aproximación normal para realizar la prueba, en dicho caso se puede utilizar el siguiente cuantil:

$$t = \frac{1}{2} (n + z_{\alpha/2} \sqrt{n})$$

Donde $z_{\alpha/2}$ es el cuantil de una distribución normal estándar que se puede obtener en la tabla correspondiente.

y calculamos el p -value de la siguiente manera:

$$p\text{-value} = 2 * \min\{P[Y \leq T], P[Y \geq T]\}$$

Sugerimos que si $n > 20$, el p -value puede obtenerse más sencillo usando la aproximación normal:

$$P[Y \leq T] \approx \mathcal{N}\left(\frac{2T - n + 1}{\sqrt{n}}\right)$$

$$P[Y \geq T] \approx 1 - \mathcal{N}\left(\frac{2T - n - 1}{\sqrt{n}}\right)$$

Caso B (Prueba de cola inferior)

$$H_0 : P[\text{obtener } +] \geq P[\text{obtener } -]$$

vs

$$H_a : P[\text{obtener } +] < P[\text{obtener } -]$$

Regla de decisión

Rechazamos H_0 a un nivel de significancia α si

$$T \leq t$$

Elegimos α , el tamaño de la prueba y debemos encontrar t tal que:

$$\mathbf{P}[Y \leq t] = \alpha$$

Donde $Y \sim \text{Bin}(n, 1/2)$.

y calculamos el $p - \text{value}$ de la siguiente manera:

$$p - \text{value} = \mathbf{P}[Y \leq T]$$

Sugerimos que si $n > 20$, el $p - \text{value}$ puede obtenerse más sencillo usando la aproximación normal:

$$\mathbf{P}[Y \leq T] \approx \mathcal{N}\left(\frac{2T - n + 1}{\sqrt{n}}\right)$$

Caso C (Prueba de cola superior)

$$\mathbf{H}_0 : \mathbf{P}[\text{obtener } +] \leq \mathbf{P}[\text{obtener } -]$$

vs

$$\mathbf{H}_a : \mathbf{P}[\text{obtener } +] > \mathbf{P}[\text{obtener } -]$$

Regla de decisión

Rechazamos H_0 a un nivel de significancia α si

$$T > n - t$$

Elegimos α , el tamaño de la prueba y debemos encontrar t tal que

$$\mathbf{P}[Y \leq t] = \alpha$$

Donde $Y \sim \text{Bin}(n, 1/2)$.

y calculamos el $p - \text{value}$ de la siguiente manera:

$$p - \text{value} = \mathbf{P}[Y \geq T]$$

Sugerimos que si $n > 20$, el $p - \text{value}$ puede obtenerse más sencillo usando la aproximación normal:

$$\mathbf{P}[Y \geq T] \approx 1 - \mathcal{N}\left(\frac{2T - n - 1}{\sqrt{n}}\right)$$

Ahora aplicaremos lo anterior en un ejemplo ilustrativo:

4.5. Ejemplo

Un grupo de 6 amigos se puso a dieta, en un intento para perder peso obtuvieron los siguientes resultados:

Nombre	Peso Antes	Peso Después
<i>Ed</i>	174	165
<i>Jim</i>	191	186
<i>Max</i>	188	183
<i>Ray</i>	182	178
<i>Abdul</i>	201	203
<i>Phil</i>	188	181

El grupo de amigos desea ver si la dieta que están realizando es efectiva.

Recordando que cada par de datos en la muestra se clasificará de la siguiente manera:

- " + " cuando $X_i < Y_i$
- " - " cuando $X_i > Y_i$

y se omitirán las parejas cuando $X_i = Y_i$. Y el tamaño de la muestra después de quitar los empates será n , haremos lo siguiente:

Nombre	Peso Antes X_i	Peso Después Y_i	Signo
<i>Ed</i>	174	165	"
<i>Jim</i>	191	186	"
<i>Max</i>	188	183	"
<i>Ray</i>	182	178	"
<i>Abdul</i>	201	203	-
<i>Phil</i>	188	181	"

Paso 1 Escribimos la prueba a utilizar

La prueba a utilizar **Prueba de signos Caso B cola inferior**

Paso 2 Formulamos nuestras hipótesis en contexto al problema planteado

$$H_0 : P[\text{obtener } +] \geq P[\text{obtener } -]$$

vs

$$H_a : P[\text{obtener } +] < P[\text{obtener } -]$$

es decir,

H_0 : En promedio los pesos antes de la dieta son mayores a los pesos después de la dieta.

vs

H_a : En promedio los pesos después de la dieta son mayores a los pesos antes de la dieta.

Paso 3 Estadístico de prueba

Utilizaremos el estadístico

$$T = 1 \text{ número de signos " + "}$$

$$T \sim \text{Bin}(6, 1/2)$$

Paso 4 Procedimiento completo

Supuestos:

1. Muestra aleatoria de tamaño 6.
 - Tomaremos como “éxito” los que si perdieron peso, en este caso los signos “+”
 - Tomaremos $\alpha = 5\% = 0.05$ el nivel de significancia.
 - $n = 6$ tamaño de la muestra
 - $p = 1/2$

Paso 5 Regla de decisión

Rechazamos H_0 a un nivel de significancia α si $T \leq t$

$$T = 1 \leq 1 = t$$

\therefore Rechazamos H_0

y calculamos el $p - value$ de la siguiente manera:

$$p - value = \mathbf{P}[Y \leq T] = \mathbf{P}[Y \leq 1] = 0.1093$$

$\therefore p - value = 0.1093 > 0.05 = \alpha$ entonces no rechazamos H_0 .

Nos preguntaremos, ¿Cómo es que pasa esto?, bueno esto se debe a que nuestra muestra es muy pequeña (son 6 datos), claro no siempre vamos a tener muestras así. Pero, debemos tener en cuenta que al tener muestras reducidas le estamos reduciendo potencia a la prueba de hipótesis utilizada, ya que la capacidad de detectar diferencias significativas entre los datos disminuye.

Paso 6 Conclusión

Existe evidencia suficiente para decir que los pesos después de la dieta son menores a los pesos antes de la dieta.

4.6. Ejemplo en R-Studio

Ahora haremos la réplica en R.

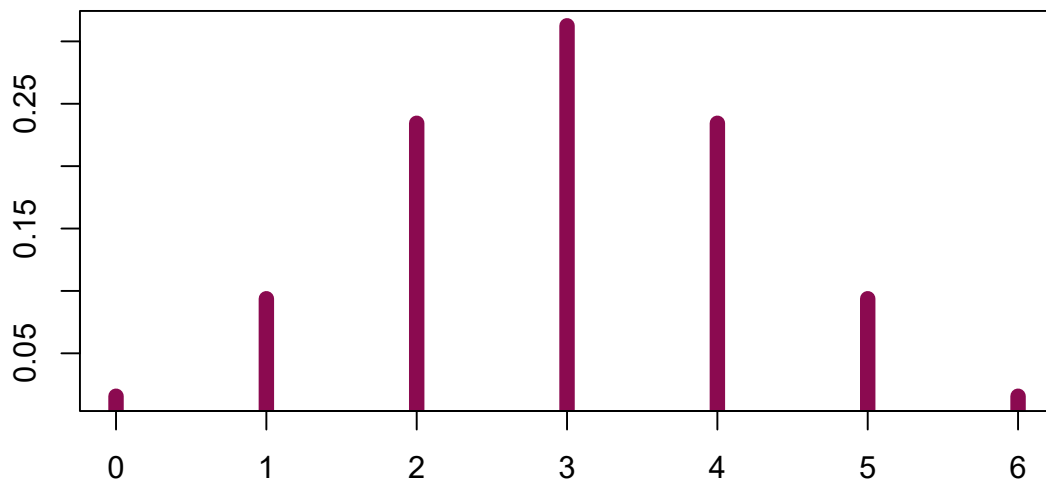
La estadística de prueba será $T = 1$. Tomaremos $\alpha = 5\%$

```
#Datos
T=1           #Número de éxitos (+)
alpha=0.05    #Nivel de
n=6           #Tamaño de la muestra
p=0.5         #Para la prueba de signos "p" siempre será 1/2
```

Según el planteamiento de las hipótesis, este es un Caso B (de cola inferior), por lo que siguiendo la regla de decisión, se rechazará H_0 si $T \leq t$ donde t será en cuantil que acumule 5% en la distribución binomial

Podemos graficar la función de distribución:

Función de distribución de una variable Binomial(6,0.5)



Calculamos t y el p -value:

```
t=qbinom(.05,n,p)      #Valor a comparar con nuestro estadístico
t
```

```
[1] 1
```

```
pvalue=pbinom(T,n,p)   #P-value
pvalue
```

```
[1] 0.109375
```

Tenemos que como $T = 1 \leq 1 = t$, entonces se rechaza H_0 y por lo tanto se concluye que hay información suficiente para decir que en promedio los pesos después de la dieta son menores a los pesos antes de la dieta. Por lo tanto la dieta parece funcionar.

Finalmente utilizaremos la función de R:

```
binom.test(T,n,p=0.5,alternative = "less")
```

Exact binomial test

```
data: T and n
number of successes = 1, number of trials = 6, p-value = 0.1094
alternative hypothesis: true probability of success is less than 0.5
95 percent confidence interval:
 0.0000000 0.5818034
sample estimates:
probability of success
 0.1666667
```

Capítulo 5

Prueba Mc Nemar

El objetivo de esta prueba es ver el efecto que tuvo cierto “tratamiento” sobre un sujeto cuya “condición” se observa antes y después del mismo. Por ejemplo se puede usar para analizar el efecto que un debate tiene en la decisión de una asamblea. Para esto se puede hacer una encuesta para registrar la opinión de los miembros de la misma en categorías **a favor** o **en contra** de una propuesta y después de realizado el debate se les vuelve a preguntar su opinión sobre el tema. Nos interesa estudiar a los individuos que cambiaron de opinión, es decir, los que antes estaban **a favor** y ahora están **en contra** y los que estaban **en contra** y ahora están **a favor**.

La manera de estudiar esto es identificando a los individuos encuestados con parejas ordenadas de la forma $(0,0), (0,1), (1,0)$ o $(1,1)$ donde la primera entrada representa la **postura** del individuo antes del debate y la segunda representa su postura después. (“0”, por ejemplo, puede representar **en contra** y “1” **a favor**). Luego se ingresan estos datos en una tabla de contingencia de la siguiente forma:

	ahora en contra	ahora a favor
antes en contra	Número de $(0, 0)$	Número de $(0, 1)$
antes a favor	Número de $(1, 0)$	Número de $(1, 1)$

5.1. Datos

Los datos consisten en observaciones bivariadas aleatorias:

$$(X_1, Y_1), (X_2, Y_2), (X_3, Y_3), \dots, (X_{n'}, Y_{n'})$$

Dentro de cada par de datos en la muestra se tiene X e Y que sólo tomarán los valores 0 y 1.

En general los datos se resumen en una tabla de contingencia de la forma:

	después 0	después 1
antes 0	a = Número de $(0, 0)$	b = Número de $(0, 1)$
antes 1	c = Número de $(1, 0)$	d = Número de $(1, 1)$

5.2. Supuestos

- 1) Cada pareja (X_i, Y_i) son mutuamente independientes.
- 2) La escala de medida es nominal con 2 categorías para toda X_i y Y_i .
- 3) La probabilidad de que se observe $(0,1)$ es \geq que la de $(1,0)$ o la de $(1,0)$ es $\geq (0,1)$ para todos los elementos de la muestra.

5.3. Estadístico de Prueba

$$T_1 = \frac{(b-c)^2}{(b+c)} \approx \chi_{(1)}^2$$

Es decir, la distribución de T_1 es aproximadamente $\chi_{(1)}^2$.

Por otro lado, si $b+c \leq 20$, es preferible utilizar el siguiente estadístico:

$$T_2 = b \sim \text{Bin}(b+c, p)$$

Es decir, la distribución de T_2 es exactamente $\text{Bin}(b+c, p)$

Dependiendo del planteamiento de nuestro problema a resolver se formulan las hipótesis:

5.4. Hipótesis

Caso A (Prueba de dos colas)

$$\mathbf{H}_0 : \mathbf{P}[X_i = 0, Y_i = 1] = \mathbf{P}[X_i = 1, Y_i = 0]$$

vs

$$\mathbf{H}_a : \mathbf{P}[X_i = 0, Y_i = 1] \neq \mathbf{P}[X_i = 1, Y_i = 0]$$

Es decir, se quiere ver si la probabilidad de que el individuo pase de 0 a 1 es la misma que la probabilidad de que pase de 1 a 0. En otras palabras, se quiere comprobar si el efecto del tratamiento fue neutro o si hay alguna tendencia tras el mismo.

Regla de decisión

Rechazamos H_0 a un nivel de significancia α si

$$T_2 \leq t \quad \text{o} \quad T_2 > n - t$$

Elegimos α , el tamaño de la prueba y debemos encontrar t tal que

$$\mathbf{P}[Y \leq t] = \alpha/2$$

Donde $Y \sim \text{Bin}(b+c, 1/2)$.

Por otro lado, si $b+c > 20$

Rechazamos H_0 a un nivel de significancia α si

$$T_1 > t_{1-\alpha}$$

Elegimos α , el tamaño de la prueba y debemos encontrar $t_{1-\alpha}$ tal que

Donde $t_{1-\alpha}$ se encuentra en la tabla correspondiente ya que $T_1 \sim \chi_{(1)}^2$.

y calculamos el p -value de la siguiente manera:

$$p\text{-value} = 2 * \min\{\mathbf{P}[Y \leq T_2], \mathbf{P}[Y \geq T_2]\}$$

Ahora aplicaremos lo anterior en un ejemplo ilustrativo:

5.5. Ejemplo

May y Johnson (1997) publicaron el resultado de un estudio en el cual los investigadores querían determinar el efecto de la hipnosis en reducir el dolor asociado con la venopunción en pacientes juveniles de cáncer. Los datos se observan en la siguiente tabla:

<i>Antes de la hipnosis</i>	<i>Después de la hipnosis</i>		<i>Total</i>
	Sin Dolor	Con Dolor	
Sin Dolor	a = 18	b = 4	22
Con Dolor	c = 12	d = 5	17
<i>Total</i>	30	9	39

Paso 1 Escribimos la prueba a utilizar

La prueba a utilizar **Prueba de Mc Nemar Caso A de dos colas**

Paso 2 Formulamos nuestras hipótesis en contexto al problema planteado

$$\mathbf{H}_0 : \mathbf{P}[X_i = 0, Y_i = 1] = \mathbf{P}[X_i = 1, Y_i = 0]$$

vs

$$\mathbf{H}_a : \mathbf{P}[X_i = 0, Y_i = 1] \neq \mathbf{P}[X_i = 1, Y_i = 0]$$

es decir,

$$\mathbf{H}_0 : \text{Se mantiene el sentir dolor después de la hipnosis.}$$

vs

$$\mathbf{H}_a : \text{Hay un cambio considerable al no sentir dolor después de la hipnosis.}$$

Paso 3 Estadístico de prueba

Recordando que $b + c = 4 + 12$ entonces $b + c = 16 \leq 20$ entonces utilizaremos el estadístico:

$$T_2 = b = 4$$

$$T_2 \sim \text{Bin}(16, 1/2)$$

Paso 4 Procedimiento completo

Supuestos

1. Muestra aleatoria de tamaño $n=16$.
 - Tomaremos $\alpha = 5\% = 0.05$ el nivel de significancia.
 - $n = 16$ tamaño de la muestra
 - $p=1/2$

Paso 5 Regla de decisión

Rechazamos H_0 a un nivel de significancia α si $T_2 \leq t$ o $T_2 > n - t$

Debemos encontrar t tal que

$$\mathbf{P}[Y \leq t] = \alpha/2 = 0.025$$

$$T_2 = 4 \leq 5 = t \quad o \quad 5 > 16 - 4$$

$$T_2 = 4 \not\leq 5 = t \quad y \quad 5 \not> 12$$

\therefore Rechazamos H_0

Paso 6 Conclusión

Existe evidencia suficiente para decir que hay un cambio considerable a no sentir dolor después de recibir hipnosis.

5.6. Ejemplo en R-Studio

Ahora haremos la réplica en R.

Veamos que $b + c = 4 + 12 = 16 \leq 20$ por lo tanto el estadístico de prueba será

$$T_2 = b = 4 \quad T_2 \sim \text{Bin}(16, 1/2)$$

Tomaremos $\alpha = 5\%$

	Despues hipnosis	
Antes hipnosis	Sin dolor	dolor
Sin dolor	18	4
dolor	12	5

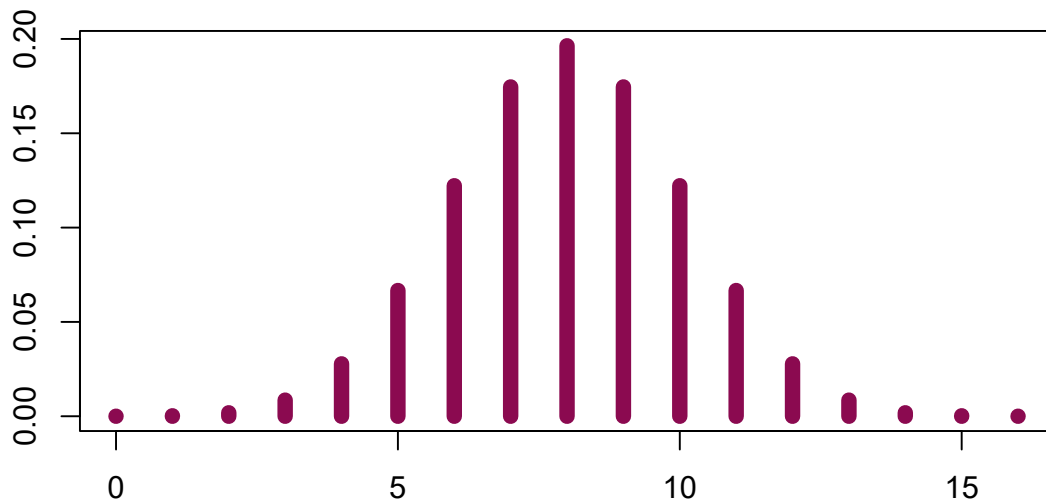
```
#Datos

T_2=4           #Es la posición b
n=16            #Tamaño de la muestra
alpha=0.05      #Nivel de significancia
p=0.5           #Para la prueba Mc nemar p es 1/2
```

Según el planteamiento de las hipótesis, este es un Caso A (dos colas), por lo que siguiendo la regla de decisión tenemos que rechazaremos H_0 si $T_2 \leq t$ o $T_2 > n - t$ donde t será en cuantil que acumule 5% en la distribución binomial

Podemos graficar la función de distribución:

Función de distribución de una variable Binomial(16,0.5)



Calculamos t y el p -value:

```
t=qbinom(.05,n,p)      #Valor con el que vamos a comparar el estadístico
t
```

```
[1] 5
```

```
pvalue=2*min(c(pbinom(T_2,n,p), pbinom(T_2,n,p,lower.tail = F))) #P-value
pvalue
```

```
[1] 0.07681274
```

Tenemos que como $T_2 = 4 \leq 5 = t$, entonces se rechaza H_0 como ya rechazamos en la cola inferior no es necesario probar la cola superior, pero si NO hubieramos rechazado con la primer ecuación debemos probar si $T_2 > n - t$.

Por lo tanto se concluye que hay información suficiente para decir que hay un cambio considerable al no sentir dolor después de la hipnosis.

Finalmente utilizaremos la función en R:

```
binom.test(4,n,p,"t")
```

Exact binomial test

data: 4 and n

number of successes = 4, number of trials = 16, p-value = 0.07681

alternative hypothesis: true probability of success is not equal to 0.5

95 percent confidence interval:

0.07266204 0.52377082

sample estimates:

probability of success

0.25

```
#En el caso de tener una n mayor a 20 utilizamos la siguiente  
#prueba
```

Con corrección por continuidad:

```
mcnemar.test(data, correct = TRUE) #CON CORRECCION POR CONTINUIDAD
```

McNemar's Chi-squared test with continuity correction

data: data

McNemar's chi-squared = 3.0625, df = 1, p-value = 0.08012

Sin corrección por continuidad:

```
mcnemar.test(data, correct = FALSE) #SIN CORRECCION POR CONTINUIDAD
```

McNemar's Chi-squared test

data: data

McNemar's chi-squared = 4, df = 1, p-value = 0.0455

Capítulo 6

Prueba de Cox Stuart

La siguiente prueba tiene como objetivo identificar tendencias en series de observaciones. Si $\{X_i\}_{i=1}^n$ es una serie de observaciones consecutivas, una manera de intentar descubrir si hay o no una tendencia es fijarse en las diferencias del tipo $X_{i+c} - X_i$ con $c = \frac{n}{2}$ si n es par y $c = \frac{(n+1)}{2}$ si es impar. Si $\{X_i\}_{i=1}^n$ tuviera una tendencia creciente, esperaríamos que las diferencias $X_{i+c} - X_i$ fuesen, en promedio, positivas, en cambio, si la tendencia fuese decreciente, se esperaría que las $X_{i+c} - X_i$ fueran negativas, en promedio. Note que la elección de c produce las mayores distancias entre índices temporales para diferencias homogéneas de la serie de tiempo.

6.1. Datos

Los datos son observaciones consecutivas de una serie de tiempo $\{X_i\}_{i=1}^n$.

En esta prueba se agrupan en parejas de la forma (X_i, X_{i+c}) a las que se les asigna el signo " + " si $X_{i+c} > X_i$, el signo " - " si $X_{i+c} < X_i$ o "0" si hay empates.

6.2. Supuestos

- 1) Las variables aleatorias X_1, X_2, \dots, X_n son mutuamente independientes.
- 2) La escala de medida de las X_i s es al menos ordinal.
- 3) Las X_i s son todas idénticamente distribuidas o existe una tendencia creciente o decreciente.

6.3. Estadístico de Prueba

$$T = \text{Total de signos " + "}$$

La distribución nula de T es una distribución binomial con n el número de parejas de la muestra sin empates y $p = 1/2$.

$$T \sim \text{Bin}(n, 1/2)$$

Dependiendo del planteamiento de nuestro problema a resolver se formulan las hipótesis:

6.4. Hipótesis

Caso A (Prueba de dos colas)

$$H_0 : P[\text{obtener } +] = P[\text{obtener } -]$$

vs

$$H_a : P[\text{obtener } +] \neq P[\text{obtener } -]$$

Regla de decisión

Rechazamos H_0 a un nivel de significancia α si

$$T \leq t \quad \text{o} \quad T > n - t$$

Elegimos α , el tamaño de la prueba y debemos encontrar t tal que

$$P[Y \leq t] = \alpha/2$$

Donde $Y \sim \text{Bin}(n, 1/2)$.

Por otro lado, si $n > 20$ puede resultar más sencillo utilizar una aproximación normal para realizar la prueba, en dicho caso se puede utilizar el estadístico:

$$t = \frac{1}{2} (n + z_{\alpha/2} \sqrt{n})$$

Donde $z_{\alpha/2}$ es el cuantil de una distribución normal estándar que se puede obtener en la tabla correspondiente.

y calculamos el p - value de la siguiente manera:

$$p - \text{value} = 2 * \min\{P[Y \leq T], P[Y \geq T]\}$$

Sugerimos que si $n > 20$, el p - value puede obtenerse más sencillo usando la aproximación normal:

$$P[Y \leq T] \approx \mathcal{N}\left(\frac{2T - n + 1}{\sqrt{n}}\right)$$

$$P[Y \geq T] \approx 1 - \mathcal{N}\left(\frac{2T - n - 1}{\sqrt{n}}\right)$$

Caso B (Prueba de cola inferior)

$$H_0 : P[\text{obtener } +] \geq P[\text{obtener } -]$$

vs

$$H_a : P[\text{obtener } +] < P[\text{obtener } -]$$

Regla de decisión

Rechazamos H_0 a un nivel de significancia α si

$$T \leq t$$

Elegimos α , el tamaño de la prueba y debemos encontrar t tal que

$$\mathbf{P}[Y \leq t] = \alpha$$

Donde $Y \sim \text{Bin}(n, 1/2)$.

y calculamos el $p - \text{value}$ de la siguiente manera:

$$p - \text{value} = \mathbf{P}[Y \leq T]$$

Sugerimos que si $n > 20$, el $p - \text{value}$ puede obtenerse más sencillo usando la aproximación normal:

$$\mathbf{P}[Y \leq T] \approx \mathcal{N}\left(\frac{2T - n + 1}{\sqrt{n}}\right)$$

Caso C (Prueba de cola superior)

$$\mathbf{H}_0 : \mathbf{P}[\text{obtener } +] \leq \mathbf{P}[\text{obtener } -]$$

vs

$$\mathbf{H}_a : \mathbf{P}[\text{obtener } +] > \mathbf{P}[\text{obtener } -]$$

Regla de decisión

Rechazamos H_0 a un nivel de significancia α si

$$T > n - t$$

Elegimos α , el tamaño de la prueba y debemos encontrar t tal que

$$\mathbf{P}[Y \leq t] = \alpha$$

Donde $Y \sim \text{Bin}(n, 1/2)$.

y calculamos el $p - \text{value}$ de la siguiente manera:

$$p - \text{value} = \mathbf{P}[Y \geq T]$$

Sugerimos que si $n > 20$, el $p - \text{value}$ puede obtenerse más sencillo usando la aproximación normal:

$$\mathbf{P}[Y \geq T] \approx 1 - \mathcal{N}\left(\frac{2T - n - 1}{\sqrt{n}}\right)$$

Ahora aplicaremos lo anterior en un ejemplo ilustrativo:

6.5. Ejemplo

Una hidroeléctrica esta muy interesada en seguir las tasas promedio de descarga de agua de las corrientes que lo alimentan. Se tienen los registros mensuales de estas tasas (en pies cúbicos por segundo) durante un período de 24 meses. La empresa sospecha que la tasa esta disminuyendo. ¿Podemos corroborar la sospecha con un nivel de significancia del 5 %?

Los datos están en la siguiente tabla, los cuales estan emparejados por mes ya que estas tasa de descarga siguen un ciclo anual, Se sabe que la tasa de descarga sigue un ciclo anual, por lo que no se puede emparejar las descargas de corriente durante dos meses diferentes; sin embargo, al emparejar los mismos meses consecutivos, se puede investigar la existencia de una tendencia. Los datos son los siguientes::

Mes	Primer Año	Segundo Año	Mes	Primer Año	Segundo Año
<i>Enero</i>	14.6	14.2	<i>Julio</i>	92.8	88.1
<i>Febrero</i>	12.2	10.5	<i>Agosto</i>	74.4	80
<i>Marzo</i>	104	123	<i>Septiembre</i>	75.4	75.6
<i>Abril</i>	220	190	<i>Octubre</i>	51.7	48.8
<i>Mayo</i>	110	138	<i>Noviembre</i>	29.3	27.1
<i>Junio</i>	86	98.1	<i>Diciembre</i>	16	15.7

Los datos ya estan emparejados ahora sólo debemos asignar los signos considerando X_i el primer año y Y_i el segundo año.

Recordando que cada par de datos en la muestra se clasificará por un signo "+" cuando $X_i < Y_i$, por un signo "-" cuando $X_i > Y_i$ y se omitirán las parejas cuando $X_i = Y_i$. Y el tamaño de la muestra después de quitar los empates será n , haremos lo siguiente:

Mes	Primer Año	Segundo Año	Signo
<i>Enero</i>	14.6	14.2	—
<i>Febrero</i>	12.2	10.5	—
<i>Marzo</i>	104	123	+
<i>Abril</i>	220	190	—
<i>Mayo</i>	110	138	+
<i>Junio</i>	86	98.1	+
<i>Julio</i>	92.8	88.1	—
<i>Agosto</i>	74.4	80	+
<i>Septiembre</i>	75.4	75.6	+
<i>Octubre</i>	51.7	48.8	—
<i>Noviembre</i>	29.3	27.1	—
<i>Diciembre</i>	16	15.7	—

Paso 1 Escribimos la prueba a utilizar

La prueba a utilizar **Prueba Cox-Stuart Caso B cola inferior**

Paso 2 Formulamos nuestras hipótesis en contexto al problema planteado

H_0 : La tasa promedio de descarga de agua no está disminuyendo.

vs

H_a : La tasa promedio de descarga de agua está disminuyendo.

Paso 3 Estadístico de prueba

Utilizaremos el estadístico

$$T = 5 \text{ número de signos " + "}$$

$$T \sim \text{Bin}(12, 1/2)$$

Paso 4 Procedimiento completo**Supuestos:**

1. Muestra aleatoria de tamaño 12.
 - Tomaremos como “éxito” los que si están disminuyendo, en este caso los signos “+”.
 - Tomaremos $\alpha = 5\% = 0.05$ el nivel de significancia.
 - $n = 12$ tamaño de la muestra
 - $p = 1/2$

Paso 5 Regla de decisión

Rechazamos H_0 a un nivel de significancia α si $T \leq t$

$$T = 5 \not\leq 3 = t$$

\therefore No rechazamos H_0

y calculamos el p – value de la siguiente manera:

$$p - value = \mathbf{P}[Y \leq T] = \mathbf{P}[Y \leq 5] = 0.3872$$

$\therefore p - value = 0.3872 > 0.05 = \alpha$ entonces no rechazamos H_0 .

Paso 6 Conclusión

Existe evidencia suficiente para decir que la tasa promedio de descarga de agua no está disminuyendo.

6.6. Ejemplo en R-Studio

Ahora haremos la réplica en R.

La estadística de prueba será $T = 5$. Tomaremos $\alpha = 5\%$

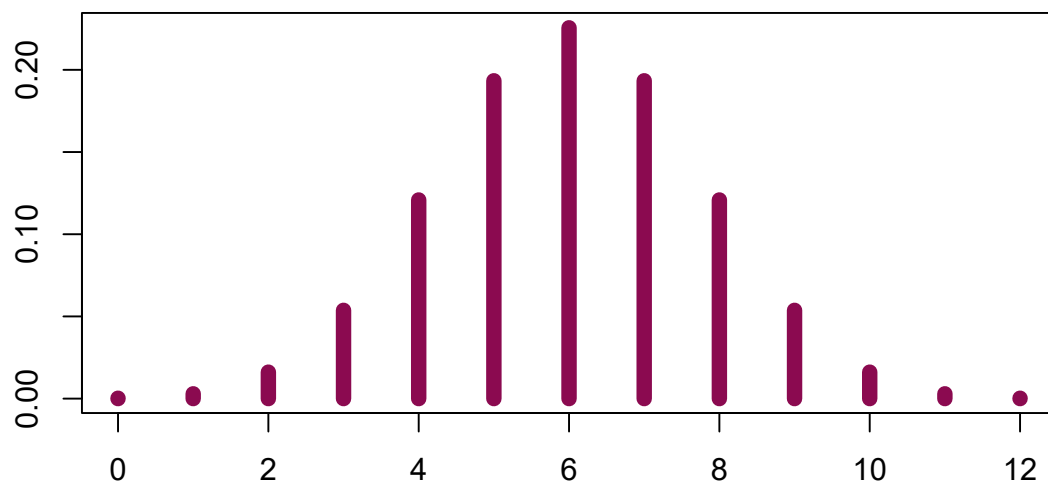
```
#Datos

T_1=5           #Número de éxitos (+)
alpha=0.05      #Nivel de significancia
n=12            #Tamaño de la muestra
p=0.5           #Para la prueba de cox "p" siempre será 1/2
```

Según el planteamiento de las hipótesis, este es un Caso B (de cola inferior), por lo que siguiendo la regla de decisión se rechaza H_0 si $T \leq t$ donde t será en cuantil que acumule 5 % en la distribución binomial

Podemos graficar la función de distribución:

Función de distribución de una variable Binomial(12,0.5)



Calculamos t y el p -value:

```
t=qbinom(.05,n,p)      #Cuantil a comparar con el estadístico
t
```

```
[1] 3
```

```
pvalue=pbinom(T_1,n,p) #p-value
pvalue
```

```
[1] 0.387207
```

Tenemos que como $T = 5 \not\leq 3 = t$, entonces NO se rechaza H_0 y por lo tanto se concluye que NO hay información suficiente para decir que la tasa promedio de descarga mensual este disminuyendo.

Finalmente utilizaremos la función en R:

```
binom.test(5,n,p=0.5,alternative = "less")
```

Exact binomial test

```
data: 5 and n
number of successes = 5, number of trials = 12, p-value = 0.3872
alternative hypothesis: true probability of success is less than 0.5
95 percent confidence interval:
 0.0000000 0.6847622
sample estimates:
probability of success
 0.4166667
```

```
x=c(14.6,14.2,12.2,10.5,104,123,220,190,110,138,86,98.1,92.8,88.1,74.4,80,75.4,75.6,51.7,48.8,29.3,2  
  
library(randtests)  
cox.stuart.test(x)
```

Cox Stuart test

```
data: x  
statistic = 4, n = 12, p-value = 0.3877  
alternative hypothesis: non randomness
```

Capítulo 7

Ejercicios

1. Un fabricante de teléfonos móviles afirma que sólo el 5 % de todas las unidades que vende sufre una falla durante el primer mes de operación normal. Una organización de consumidores ha pedido a 45 consumidores que han adquirido estos teléfonos móviles, que reporten cualquier mal funcionamiento durante el primer mes. Al final de éste sólo siete consumidores reportaron mal funcionamiento. Si la organización de consumidores cree que la proporción de teléfonos que sufrirán alguna falla es mayor al valor afirmado por el fabricante. ¿Con una α del 10 % podría la organización sustentar su creencia?
2. Un candidato a través de encuestas propias afirma que el 65 % o más de los votantes están a su favor. Sin embargo, a través de una encuesta a 20 personas, 10 están a su favor. Probar que a un nivel de significancia del 5 %, la hipótesis de que el candidato a sobrestimado los votos a su favor.
3. En una muestra de 150 partidos de básquetbol universitario, el equipo de casa ganó 98 partidos. Realice una prueba para determinar si los datos sustentan la hipótesis de que en el básquetbol universitario el equipo de casa tiene ventaja. ¿A qué conclusión llega con $\alpha = 0.05$?
4. Una muestra aleatoria de niños de tercer año de secundaria mostró las siguientes observaciones de peso (kg)

64	60	44	54	61
46	66	56	42	62
39	54	68	64	65
75	38	53	58	46

Probar:

- a) La mediana de los pesos es 46kg.
 - b) El cuartíl superior es al menos 60kg.
 - c) El tercer decil no es mayor a 45kg.
5. La siguiente es una muestra de 15 departamentos nuevos de 2 recamaras con estacionamiento en la colonias Roma, Condesa y Escandón. Los datos están en millones de pesos.

6.4	5	4.2
4.6	4.4	5.6
3.5	3.8	4.5
7.5	5.6	4.2
8.1	5.8	6.3

Probar:

- a) Cuando menos 50 % de las observaciones están por debajo de los 4.3 millones.

b) No mas del 20 % de las observaciones tienen un costo mayor a 7 millones.

6. En una encuesta anual para determinar si los salarios en el sector federal son proporcionales con los pagos en el sector privado, los trabajadores publicos y privados fueron emparejados tan cerca como fue posible (con respecto al tipo de trabajo, formación académica, años de experiencia, etc.) los salarios se ordenaron en parejas.

Pareja i	Privado	Gobierno
1	12,500	11,750
2	22,300	20,900
3	14,500	14,800
4	32,300	29,900
5	20,800	20,500
6	19,200	18,400
7	15,800	14,500
8	17,500	17,900
9	23,300	21,400
10	42,100	43,200
11	16,800	15,200
12	14,500	14,200

Probar la hipótesis nula de que los salarios son iguales contra la hipótesis alternativa de que el salario de los trabajadores federales es generalmente menor a la contraparte en el sector privado. Utiliza $\alpha = 0.1$.

7. Una oficina tiene dos computadoras : A y B. En un estudio del uso del ordenador, la compañía ha recabado registro de las tasas de uso por 5 semanas. La meta es decidir cual computadora se pone bajo un contrato de servicio porque tiene una tasa alta de uso. Con los datos de las tasas de uso de la siguiente tabla se puede hacer una recomendación preliminar respecto a que computadora contratar?

Semanas	Computadora A	Computadora B
1	15.7	32.4
2	10.8	41.2
3	45	35.1
4	12.3	25
5	8.2	8.2

8. Se tienen dos trituradores de alimentos y se tiene la sospecha de que el aparato B es mas eficiente que el aparato A. Para probar dicha sospecha, se probaron en ambos trituradores diferentes alimentos y se registró el tiempo en minutos que le tomaba a cada aparato convertir el alimento en puré. Los resultados fueron los siguientes.

Alimento	A	B
1	0.5	0.6
2	1	0.9
3	1.2	1.2
4	0.8	0.9
5	0.4	0.5
6	1.5	1.8
7	0.3	0.4
8	1.2	1.4
9	0.7	0.9

Con los datos observados se puede corroborar la sospecha en cuanto a la eficiencia de los trituradores? Use $\alpha = 5\%$

9. Se toma una muestra aleatoria de 135 ciudadanos de E.U. y se les preguntó su opinión con respecto a cierta política. El estudio registró a 43 ciudadanos que estaban en contra de esa política. Después de varias semanas, durante las cuales los ciudadanos recibieron cartas informativas, se les volvió a preguntar su opinión; 37 estuvieron en contra, y 30 de las 37 originalmente no estaban en contra de la política. ¿Es significativo el cambio en el número de personas en contra de la política?

10. Un investigador intenta determinar si un fármaco tiene un efecto sobre una enfermedad particular. Se cuenta con la información de los individuos en el estudio con el diagnóstico (enfermedad: presente o ausente) antes del tratamiento, y el diagnóstico después del tratamiento

Antes de tratamiento	Después de tratamiento		Total
	Presente	Ausente	
Presente	101	121	222
Ausente	59	33	92
Total	160	154	314

11. Un programa ecológico sobre la contaminación de un río tomó 5 muestras de agua de diferentes lugares de un río antes y después de dos años, obteniéndose los siguientes resultados. Los números representan la media de la contaminación, donde medidas grandes indican alta contaminación.

Numero de Muestras	Medidas iniciales	Medidas después de 2 años
1	88.4	87.1
2	81.3	79.4
3	68.4	69.1
4	100.5	91.1
5	93.2	95.3

Se está interesado en saber si el programa de rehabilitación ecológica ha tenido efecto en la reducción de la contaminación. Use $\alpha = 1\%$.

12. Se presentan a continuación los tipos de cambio MXN-USD de los últimos 30 días.

24.48 , 24.07 , 23.66 , 23.70 , 23.93 , 24.34 , 24.22 , 23.87 , 23.96 , 24.05 , 23.78 , 23.74 , 23.21 , 22.88 , 22.74 , 22.73 , 22.58 , 22.21 , 22.33 , 22.21 , 22.17 , 22.21 , 22.04 , 21.79 , 21.76 , 21.91 , 21.58 , 21.60 , 21.50 , 21.91

Se desea corroborar la aseveración del presidente respecto a que el peso mexicano esta recuperando fuerza. Use $\alpha = 10\%$.

Parte III

Prueba de Rango

Introducción

Ya nos empapamos de la pruebas binomiales ahora en esta seccion se presentan tres de las pruebas de rangos más utilizadas en modelos no paramétricos, la prueba de Mann y Witney, la prueba de Kruskal-Wallis y la prueba de Igualdad de Varianzas.

Estas pruebas son otros métodos no paramétrico que se usan para determinar si hay diferencia entre poblaciones. El procedimiento se basa en ordenar las observaciones pertenecientes a muestras aleatorias de cada población, asignarles rangos según sus valores y posteriormente construir la prueba sobre dichos rangos.

Capítulo 8

Prueba U-Mann y Witney

Esta prueba está diseñada para determinar si dos muestras han sido extraídas de la misma población. La prueba también es conocida como Mann-Whitney-Wilcoxon o suma de rangos de Wilcoxon.

A diferencia de las pruebas binomiales bivariadas vistas en la sección anterior, esta no se basa en una muestra por pares. En lugar usa dos muestras independientes, una de cada población a probar.

8.1. Datos

Se considera que los datos provienen de dos muestras aleatorias:

$x_1, x_2, x_3, \dots, x_n$ una muestra aleatoria de tamaño n de la población.

$y_1, y_2, y_3, \dots, y_m$ una muestra aleatoria de tamaño m de la población.

De forma conveniente, tomamos $N = n + m$.

Sea:

$F(t)$ la función de distribución de probabilidad de X .

$G(t)$ la función de distribución de probabilidad de Y .

8.2. Supuestos

- 1) Independencia dentro de cada muestra.
- 2) Independencia entre muestras.
- 3) La escala de medida es al menos ordinal.

Pasos a seguir

1. Formar una nueva muestra combinada de los $n + m$ datos
2. Ordenar de menor a mayor
3. Asignamos el rango correspondiente
4. Si hay empates entonces asignamos el promedio de los rangos

8.3. Estadístico de Prueba

- Si no hay empates

$$T = \sum_{i=1}^n R(X_i)$$

- Si hay empates

$$T_1 = \frac{T - n \frac{N+1}{2}}{\sqrt{\frac{nm}{N(N-1)} \sum_{i=1}^N R_i^2 - \frac{nm(N+1)^2}{4(N-1)}}$$

$$T_1 \sim N(0, 1)$$

Dependiendo del planteamiento de nuestro problema a resolver se formulan las hipótesis:

8.4. Hipótesis

Caso A (Prueba de dos colas)

$$\mathbf{H}_0 : F(x) = G(x) \quad \forall x, \quad \text{alternativo} \quad \mathbf{H}_0 : \mathbf{E}[X] = \mathbf{E}[Y]$$

vs

$$\mathbf{H}_a : F(x) \neq G(x) \quad \text{para alguna } x, \quad \text{alternativo} \quad \mathbf{H}_a : \mathbf{E}[X] \neq \mathbf{E}[Y]$$

Regla de decisión

Rechazamos H_0 a un nivel de significancia α si

$$T < t_{\frac{\alpha}{2}} \quad \text{o} \quad T > t_{1-\frac{\alpha}{2}}$$

Donde $t_{\frac{\alpha}{2}}$ el cuantil inferior se busca en las tablas correspondientes a nuestra prueba.

Y $t_{1-\frac{\alpha}{2}}$ el cuantil superior se calcula de la siguiente manera:

$$\omega_p = n(n+m+1) - \omega_{1-p}$$

donde ω_{1-p} es el cuantil inferior obtenido de la tabla correspondiente a nuestra prueba.

Los cuantiles aproximados en el caso de no tener empates, y n o m mayores a 20, se encuentra la aproximación normal.

$$\omega_p = \frac{n(N+1)}{2} + z_p \sqrt{\frac{nm(N+1)}{12}}$$

donde el cuantil z_p se obtiene de la tabla de la distribución normal.

y calculamos el p - *valuede* la siguiente manera:

$$p - value = 2 * \mathbf{P} \left(Z \leq \frac{T + \frac{1}{2} - n \frac{N+1}{2}}{\sqrt{\frac{nm(N+1)}{12}}} \right)$$

Donde Z es la variable aleatoria normal estándar.

- Si ocupamos T_1 , rechazamos H_0 si:

$$T_1 < z_{\frac{\alpha}{2}} \quad o \quad T_1 > z_{1-\frac{\alpha}{2}}$$

Donde $z_{\frac{\alpha}{2}}$ y $z_{1-\frac{\alpha}{2}}$ son cuantiles que se buscan en la tabla de una normal estándar.

y calculamos el $p - value$ de la siguiente manera:

$$p - value = 2 * \min\{\mathbf{P}[Z \leq T_1], \mathbf{P}[Z \geq T_1]\}$$

Donde Z es la variable aleatoria normal estándar.

Caso B (Prueba de cola inferior)

$$\mathbf{H}_0 : F(x) = G(x) \quad vs \quad \mathbf{H}_a : F(x) > G(x), \quad \text{alternativo} \quad \mathbf{H}_a : \mathbf{E}[X] < \mathbf{E}[Y]$$

Regla de decisión

Rechazamos H_0 a un nivel de significancia α si

$$T < t_\alpha$$

Donde t_α el cuantil inferior se busca en las tablas correspondientes a nuestra prueba.

Los cuantiles aproximados en el caso de no tener empates, y n o m mayores a 20, se encuentra la aproximacion normal.

$$\omega_p = \frac{n(N+1)}{2} + z_p \sqrt{\frac{nm(N+1)}{12}}$$

donde el cuantil z_p se obtiene de la tabla de la distribución normal.

y calculamos el $p - value$ de la siguiente manera:

$$p - value = \mathbf{P} \left(Z \leq \frac{T + \frac{1}{2} - n \frac{N+1}{2}}{\sqrt{\frac{nm(N+1)}{12}}} \right)$$

* Si ocupamos T_1 , rechazamos H_0 si:

$$T_1 < z_\alpha$$

Donde z_α es el cuantil que se buscan en la tabla de una normal estándar.

y calculamos el $p - value$ de la siguiente manera:

$$p - value = \mathbf{P}[Z \leq T_1]$$

Donde Z es la variable aleatoria normal estándar.

Caso C (Prueba de cola superior)

$$\mathbf{H}_0 : F(x) = G(x) \text{ vs } \mathbf{H}_a : F(x) < G(x), \text{ alternativo } \mathbf{H}_a : \mathbf{E}[X] > \mathbf{E}[Y]$$

Regla de decisión

Rechazamos H_0 a un nivel de significancia α si

$$T > t_{1-\alpha}$$

Y $t_{1-\alpha}$ el cuantil superior se calcula de la siguiente manera:

$$\omega_p = n(n+m+1) - \omega_{1-p}$$

donde ω_{1-p} es el cuantil inferior obtenido de la tabla correspondiente a nuestra prueba.

Los cuantiles aproximados en el caso de no tener empates, y n o m mayores a 20, se encuentra la aproximación normal.

$$\omega_p = \frac{n(N+1)}{2} + z_p \sqrt{\frac{nm(N+1)}{12}}$$

donde el cuantil z_p se obtiene de la tabla de la distribución normal.

y calculamos el p -value de la siguiente manera:

$$p\text{-value} = \mathbf{P} \left(Z \geq \frac{T + \frac{1}{2} - n \frac{N+1}{2}}{\sqrt{\frac{nm(N+1)}{12}}} \right)$$

- Si ocupamos T_1 , rechazamos H_0 si:

$$T_1 > z_{1-\alpha}$$

Donde $z_{1-\alpha}$ es el cuantil que se buscan en la tabla de una normal estándar.

y calculamos el p -value de la siguiente manera:

$$p\text{-value} = \mathbf{P}[Z \geq T_1] \text{ o } 1 - \mathbf{P}[Z \leq T_1]$$

Donde Z es la variable aleatoria normal estándar.

Tal vez es demasiada información que procesar así que veremos un ejemplo paso por paso:

8.5. Ejemplo

Una agencia publicitaria está investigando a qué tipo de avisos les prestan más atención los adolescentes. Se observan a 11 adolescentes, a 6 de ellos se les muestran anuncios de comida y a los 5 restantes se les muestran anuncios de bebidas; todos los anuncios tienen una duración similar y a continuación se muestra el registro del tiempo de atención (en segundos) de los 11 adolescentes. Utilizaremos $\alpha = 5\%$.

Comida	25	41	42	45	47	50
Bebidas	23	28	30	35	38	

Podemos ver que $m=6$ (en este caso comidas) y $n=5$ (en este caso bebidas)

Paso 1 Escribimos la prueba a utilizar

La prueba a utilizar **Prueba U-Mann-Whitney - Caso A - dos colas**

Paso 2 Formulamos nuestras hipótesis en contexto al problema planteado

H_0 : La distribución del tiempo de atención que prestan los adolescentes a los anuncios sobre comida es igual a la distribución de los anuncios de bebidas.

vs

H_a : La distribución del tiempo de atención que prestan los adolescentes a los anuncios sobre comida es distinta a la distribución de los anuncios de bebidas.

Ordenamos de menor a mayor

Tipo de aviso	B	C	B	B	B	B	C	C	C	C	C
Dato ordenado	23	25	28	30	35	38	41	42	45	47	50
Rango	1	2	3	4	5	6	7	8	9	10	11

Paso 3 Estadístico de prueba

Como no tenemos empates utilizaremos el estadístico T:

$$T = \sum_{i=1}^n R(X_i)$$

Paso 4 Procedimiento completo

$$T = \sum_{i=1}^5 R(X_i) = 1 + 3 + 4 + 5 + 6 = 19$$

Paso 5 Regla de decisión

Rechazamos H_0 a un nivel de significancia $\alpha=.05$ si

$$T < t_{\frac{\alpha}{2}} \quad o \quad T > t_{1-\frac{\alpha}{2}}$$

El valor del cuantil inferior localizado en las tablas es 19, y por otro lado vamos a calcular el cuantil superior:

$$\begin{aligned}\omega_p &= n(n+m+1) - \omega_{1-p} \\ \omega_p &= 5(5+6+1) - 19 \\ \omega_p &= 41\end{aligned}$$

$$19 \not< 19 \quad o \quad 19 \not> 41$$

\therefore No rechazo H_0

Calculamos p-value

$$p - value = 2 * \mathbf{P} \left(Z \leq \frac{T + \frac{1}{2} - n \frac{N+1}{2}}{\sqrt{\frac{nm(N+1)}{12}}} \right)$$

$$p - value = 2 * \mathbf{P} \left(Z \leq \frac{19 + \frac{1}{2} - 5 \frac{12}{2}}{\sqrt{\frac{5*6(12)}{12}}} \right) = 0.028$$

$$p - value = 2 * 0.028$$

$$p - value = 0.056$$

Ya que $p - value > \alpha$

\therefore No rechazo H_0

Paso 6 Conclusión

\therefore No hay diferencia significativa en el tiempo de atención de los adolescentes en los anuncios de comida y bebidas.

8.6. Ejemplo en R-Studio

Ahora haremos la réplica en R.

El primer paso es ordenar los rangos:

```
R_i=c(1,2,3,4,5,6,7,8.5,8.5,10,11)
```

El segundo paso es calcular T que corresponde a la suma de los rangos asignados a los anuncios de “Bebida”.

```
T=1+3+4+5+6 #rangos asignados a los anuncios de bebida
T
```

```
[1] 19
```

```
n=5
m=6
N=n+m
```

Según el planteamiento de las hipótesis, este es un Caso A (de 2 colas), por lo que siguiendo la regla de decisión, se rechazará H_0 si $T \leq t_1$ o $T > t_2$, donde t_1 es el cuantil que acumula 2.5% de probabilidad en una distribución normal estándar, y t_2 es el cuantil que acumula 97.5% en la misma distribución.

```
t1=19
t1
```

```
[1] 19
```

```
t2=41
t2
```

```
[1] 41
```

```
pvalue=2*0.028
pvalue
```

```
[1] 0.056
```

Observamos que $T \leq t_1$ por lo tanto rechazaremos H_0 y concluimos que la distribución del tiempo de atención que prestan los adolescentes a los anuncios sobre comida es distinta a la distribución del tiempo de atención prestada a los anuncios de bebidas.

En R la prueba “**wilcox.test**” también realiza esta prueba con un cálculo ligeramente distinto de la estadística de prueba. En el caso de la prueba de Wilcoxon lo que contaremos es el número de ocasiones que la $x_i > y_j$ para todas las combinaciones de $i = 1, \dots, n$ y $j = 1, \dots, m$.

Si las distribuciones son iguales entonces se esperaría que el número de veces que $x_i > y_j$ sea cercano a la mitad de las combinaciones.

```
B=c(23, 28, 30, 35, 38)      #Datos de bebidas
C=c(25, 41, 42, 45, 47, 50) #Datos de comidas
m1<-wilcox.test(B,C, na.rm=TRUE, paired=FALSE, exact=FALSE)
print(m1)
```

```
Wilcoxon rank sum test with continuity correction
```

```
data: B and C
```

```
W = 4, p-value = 0.05523
```

```
alternative hypothesis: true location shift is not equal to 0
```

Con esta prueba se rechaza la hipótesis nula de que ambas distribuciones son iguales. Y se concluye de la misma forma que la prueba de rangos anterior. La distribución del tiempo de atención que prestan los adolescentes a los anuncios sobre comida es distinta a la distribución del tiempo de atención prestada a los anuncios de bebidas.

Capítulo 9

Intervalo de confianza para la diferencia entre dos medias

9.1. Datos

Los datos consisten en dos muestras aleatorias X_1, \dots, X_n y Y_1, \dots, Y_m de tamaño n y m , respectivamente. Sea X y Y las variables aleatorias con la misma distribución que las X_i y las Y_j , respectivamente.

9.2. Supuestos

- 1) Ambas muestras son muestras aleatorias de sus respectivas poblaciones.
- 2) Además de la independencia dentro de cada muestra, existe mutua independencia entre las dos muestras.
- 3) Las dos funciones de distribución de la población son idénticas, excepto por una posible diferencia en los parámetros de ubicación. Es decir, hay una constante d (por ejemplo) tal que X tiene la misma función de distribución que $Y + d$.

9.3. Método

Determine el cuantil $\alpha/2$ ($\omega_{\alpha/2}$) para n y m de las tablas correspondientes a **Prueba Mann-Whitney**, o si n y m son mayores a 20 se ocupa la aproximación del cuantil, donde $1 - \alpha$ es el coeficiente de confianza. Note que esto se puede utilizar incluso si hay muchos empates.

Luego calcule k , dado por:

$$k = \omega_{\alpha/2} - n(n+1)/2$$

Para todos los pares posibles (X_i, Y_j) , encuentre las k diferencias más grandes $X_i - Y_j$ y encontrar las k diferencias más pequeñas.

Para encontrar las diferencias más grandes y más pequeñas, es conveniente ordenar primero cada muestra, de menor a mayor, y luego formar una matriz de diferencias $X_i - Y_j$ usando las X_s como filas y las Y_s como columnas.

La k -ésima diferencia más grande es el límite superior U y la k -ésima diferencia más pequeña es el límite inferior L .

Entonces el intervalo de confianza es obtenido por:

$$\mathbf{P}[L \leq \mathbf{E}(X) - \mathbf{E}(Y) \leq U] \geq 1 - \alpha$$

Ahora haremos un ejemplo para ilustrar la teoría:

9.4. Ejemplo

Se desea mezclar una masa de pastel hasta que se alcance una consistencia específica. Se mezclan cinco lotes de la mezcla usando la batidora A, y los otros cinco lotes se mezclan usando la batidora B. Los tiempos requeridos para mezclar se dan de la siguiente manera (en minutos):

Batidora A	Batidora A
7.3	7.4
6.9	6.8
7.2	6.9
7.8	6.7
7.2	7.1

Se busca un intervalo de confianza del 95 % para la diferencia de medias en los tiempos de mezcla, más específicamente para $\mathbf{E}(X) - \mathbf{E}(Y)$, donde X se refiere a la *batidora A* y Y se refiere a la *batidora B*.

Paso 1

Encontrar cuantil $\alpha/2$ ($\omega_{\alpha/2}$), para nuestro ejemplo $n=5$, $m=5$, $\alpha = 0.05$, y buscando en la **Tabla Mann-Whitney** tenemos $\omega_{\alpha/2} = \omega_{0.025} = 18$

Paso 2

Calculamos k , dado por:

$$k = \omega_{\alpha/2} - n(n+1)/2$$

$$k = 18 - (5)(6)/2 = 3$$

Paso 3

Ordenamos las muestras de menor a mayor, las X_s las usaremos como filas y las Y_s las usaremos como columnas para formar la matriz de diferencias $X_i - Y_j$:

$X_i \ Y_j$	6.7	6.8	6.9	7.1	7.4
6.9	0.2	0.1	0.0	-0.2	-0.5
7.2	0.5	0.4	0.3	0.1	-0.2
7.2	0.5	0.4	0.3	0.1	-0.2
7.3	0.6	0.5	0.4	0.2	-0.1
7.8	1.1	1.0	0.9	0.7	0.4

Paso 4

Entonces las más grandes y las más pequeñas diferencias son encontradas:

Diferencias Pequeñas	Diferencias Grandes
$6.9 - 7.4 = -0.5$	$7.8 - 6.7 = 1.1$
$6.9 - 7.1 = -0.2$	$7.8 - 6.8 = 1.0$
$7.2 - 7.4 = -0.2 = \mathbf{L}$	$7.8 - 6.9 = 0.9 = \mathbf{U}$

Paso 5

El intervalo de confianza del 95 % (L, U) para la diferencia de medias, es decir, $\mathbf{E}(X) - \mathbf{E}(Y)$ es $(-0.2, 0.9)$.

Capítulo 10

Prueba de Kruskal-Wallis

Se obtienen k *m.a.i.* de k distintas poblaciones ($k \geq 2$) y queremos probar la hipótesis nula de que todas las poblaciones tienen la misma distribución contra la alternativa de que algunas de las poblaciones tienden a tener distribución distinta. Un caso particular, si $k = 2$ se tiene la prueba de **Mann-Whitney**.

10.1. Datos

Se tienen k *m.a.* que pueden tener incluso distintos tamaños, la i -ésima *m.a.* de tamaño n_i es:

$$X_{i1}, X_{i2}, \dots, X_{in_i} \quad (\text{m.a. de la } i\text{-ésima población.})$$

Así:

<i>Muestra 1</i>	<i>Muestra 2</i>	<i>Muestra 3</i>	...	<i>Muestra k</i>
x_{11}	x_{21}	x_{31}	...	x_{k1}
x_{12}	x_{22}	x_{32}	...	x_{k2}
...	\ddots	...
x_{1n_1}	x_{2n_2}	x_{3n_3}	...	x_{kn_k}

sea N el número total de observaciones:

$$N = \sum_{i=1}^k n_i$$

10.2. Supuestos

- 1) Todas las muestras son muestras aleatorias de sus respectivas poblaciones.
- 2) Además de la independencia dentro de las muestras, suponemos que hay independencia entre las muestras (es decir, las poblaciones son independientes).
- 3) Las k poblaciones son idénticas o algunas de las poblaciones tienden a tener valores más grandes que las otras poblaciones.
- 4) La escala de medida es al menos ordinal.

10.3. Hipótesis

H_0 : Las funciones de distribución de las k poblaciones son idénticas.

vs

H_a : Al menos una de las poblaciones tiende a tener valores mayores que

al menos una de las poblaciones.

o equivalentemente :

H_a : Las k poblaciones no tienen medias idénticas.

Asignación de Rangos

- Asignamos el rango 1 a la observación más pequeña de las N observaciones, el rango 2 a la 2da más pequeña, y así sucesivamente hasta llegar a la observación mayor que recibirá el rango N .
- Sea $R(X_{ij})$ el rango asignado a la observación X_{ij}
- Sea R_i la suma de los rangos asignados a la i - ésima muestra (la i - ésima columna):

$$R_i = \sum_{j=1}^{n_i} R(X_{ij}), i = 1, 2, \dots, k$$

NOTA

Se calcula R_i para cada muestra, si hay observaciones repetidas, se le asigna el promedio de los rangos de las observaciones repetidas.

<i>Rangos</i> <i>Muestra 1</i>	<i>Rangos</i> <i>Muestra 2</i>	<i>Rangos</i> <i>Muestra 3</i>	\dots	<i>Rangos</i> <i>Muestra k</i>
$R(X_{11})$	$R(X_{21})$	$R(X_{31})$	\dots	$R(X_{k1})$
$R(X_{12})$	$R(X_{22})$	$R(X_{32})$	\dots	$R(X_{k2})$
\vdots	\vdots	\vdots	\dots	\vdots
$R(X_{1n_1})$	$R(X_{2n_2})$	$R(X_{3n_3})$	\dots	$R(X_{kn_k})$
R_1	R_2	R_3	\dots	R_k

10.4. Estadístico de prueba

$$T = \frac{1}{S^2} \left(\sum_{i=1}^k \frac{R_i^2}{n_i} - \frac{N(N+1)^2}{4} \right)$$

donde N y R_i son como los definimos anteriormente y donde:

$$S^2 = \frac{1}{N-1} \left(\sum_{\text{todos los rangos}} R(X_{ij})^2 - N \frac{(N+1)^2}{4} \right)$$

* Si no hay empates S^2 se simplifica a $N(N+1)/12$, y la estadística de prueba se reduce a:

$$T = \frac{12}{N(N+1)} \sum_{i=1}^k \frac{R_i^2}{n_i} - 3(N+1)$$

10.5. Regla de decisión

Rechazamos H_0 al nivel de significancia α si $T > \omega_{1-\alpha}$, donde $\omega_{1-\alpha}$ se obtiene de la tabla de cuantiles de la distribución de T (tablas Kruskal-Wallis). En caso de no encontrarse en la tabla, los tamaños de muestra correspondientes se utiliza la aproximación Ji-cuadrada con $k-1$ grados de libertad para el estadístico de prueba.

Es decir, rechazamos H_0 a un nivel α si $T > \chi_{(k-1)}^2(1-\alpha)$

y calculamos el p -value de la siguiente manera el p -value es aproximadamente la probabilidad de una variable aleatoria chi-cuadrado con $k-1$ grados de libertad que excede el valor observado de T .

Múltiples comparaciones

Si y sólo si la hipótesis nula es rechazada, podríamos utilizar el siguiente procedimiento para determinar cuales pares de poblaciones tienden a ser diferentes.

Podemos decir que las poblaciones i y j parecen ser diferentes si se satisface la siguiente desigualdad:

$$\left| \frac{R_i}{n_i} - \frac{R_j}{n_j} \right| > t_{1-\alpha/2} \left(S^2 \frac{N-1-T}{N-k} \right)^{1/2} \left(\frac{1}{n_i} + \frac{1}{n_j} \right)^{1/2}$$

Donde R_i y R_j son las sumas de los rangos de las dos muestras, $t_{1-(\frac{\alpha}{2})}$ es el cuantil $1-\frac{\alpha}{2}$ de la distribución t obtenida de las tablas de la distribución con $N-k$ grados de libertad. Este procedimiento es repetido para todas las parejas de poblaciones.

Ahora veamos cómo se hace un ejemplo:

10.6. Ejemplo

Se tienen 4 métodos de cultivo de maíz, para ver si hay diferencia entre uno y otro se utilizan terrenos similares y se aplican los distintos métodos y se ve la cantidad de cosecha por m^2 de tierra de cada terreno obtenido.

Paso 1 Prueba a utilizar **Prueba Kruskal-Wallis**

Método 1	Método 2	Método 3	Método 4
83	91	101	78
91	90	100	82
94	81	91	81
89	83	93	77
89	84	96	79
96	83	95	81
91	88	94	80
92	91		81
90	89		
	84		

Paso 2 Planteamiento de hipótesis

Queremos probar:

Hipótesis:

\mathbf{H}_0 : Las 4 Métodos son idénticos.

VS

\mathbf{H}_a : Al menos alguno de los métodos tiende a producir cosechas diferentes
(mayores o menores) que los otros métodos.

Paso 3 Estadístico de Prueba:

- N = Número total de observaciones

$$T = \frac{1}{S^2} * \left[\sum_{i=1}^k \frac{R_i^2}{n_i} - \frac{N * (N + 1)^2}{4} \right]$$

- Donde

$$S^2 = \frac{1}{N - 1} * \left[\sum_{\text{Todos los rangos}} R(X_{i,j})^2 - N * \frac{(N + 1)^2}{4} \right]$$

- Si NO hay empates:

$$S^2 = \frac{N * (N + 1)}{12}$$

y T se reduce a

$$T = \frac{12}{N * (N + 1)} * \sum_{i=1}^k \frac{R_i^2}{n_i} - 3 * (N + 1)$$

ahora regresando al problema:

Paso 4 Planteamiento completo

Método 1	Rango	Método 2	Rango	Método 3	Rango	Método 4	Rango
83	11	91	23	101	34	78	2
91	23	90	19.5	100	33	82	9
94	28.5	81	6.5	91	23	81	6.5
89	17	83	11	93	27	77	1
89	17	84	13.5	96	31.5	79	3
96	31.5	83	11	95	30	81	6.5
91	23	88	15	94	28.5	80	4
92	26	91	23			81	6.5
90	19.5	89	17				
		84	13.5				
$R_1 = 196.5$		$R_2 = 153$		$R_3 = 207$		$R_4 = 38.5$	

Supuestos:

- Muestra aleatoria de tamaño 34
- $k = 4$ Número de Métodos
- $\alpha = 0.05$ Nivel de significancia
- $N = 34$ Tamaño de la muestra
- $n_1 = 9, n_2 = 10, n_3 = 7, n_4 = 8$, (número de registros por categoría)

Como hubo empates utilizaremos el estadístico:

$$T = \frac{1}{S^2} \left(\sum_{i=1}^k \frac{R_i^2}{n_i} - \frac{N(N+1)^2}{4} \right)$$

donde N y R_i son como los definimos anteriormente y donde:

$$S^2 = \frac{1}{N-1} \left(\sum_{\text{todos los rangos}} R(X_{ij})^2 - N \frac{(N+1)^2}{4} \right)$$

Regresando a nuestro ejemplo:

$$T = \frac{1}{S^2} \left(\sum_{i=1}^k \frac{R_i^2}{n_i} - \frac{N(N+1)^2}{4} \right)$$

- donde N y R_i son como los definimos anteriormente y donde:

$$S^2 = \frac{1}{34-1} \left([(11)^2 + (23)^2 + (28.5)^2 + \dots + (4)^2 + (6.5)^2 +] - 34 \frac{(34+1)^2}{4} \right)$$

$$S^2 = \frac{1}{33} \left[13621.75 - 34 \frac{(35)^2}{4} \right]$$

$$S^2 = 97.25$$

Y nuestro estadístico es:

$$T = \frac{1}{97.25} \left(\left[\frac{196.5^2}{9} + \frac{153^2}{10} + \frac{207}{7} + \frac{38.5^2}{8} \right] - \frac{34(34+1)^2}{4} \right)$$

$$T = \frac{1}{97.25} (12937.71 - 10412.5)$$

$$T = \frac{1}{97.25} (2525.216964)$$

$$T = 25.46$$

Paso 5 Regla de decisión

- Si $k = 3$ y $n_i \leq 5 \dots \forall i$ y NO HAY EMPATES, usar tablas Kruskal-Wallis y rechazo H_0 a un nivel de significancia α si $T > \omega^{1-\alpha}$ donde el cuantil se busca en tablas Kruskal-Wallis, en caso de no encontrar el valor se ocupa:
- En general, rechazamos H_0 a un nivel de significancia α si $T > \chi_{k-1}^2(1-\alpha)$

Por otro lado calculamos:

$$\chi^2_{(3)}(1 - 0.05) = 7.815$$

Por lo que $T = 25.46 > 7.815 = \chi^2_{(3)}(0.95)$, entonces rechazo H_0 .

Ahora calculamos el p -value:

$$P[X > T] = 1 - P[X \leq 25.46] = 0.0000123 < \alpha$$

Por lo tanto rechazamos H_0 a un nivel de significancia de $\alpha = 0.05$,

Paso 6 Conclusión

Entonces existe evidencia suficiente para decir que al menos alguno de los métodos tiende a producir cosechas diferentes (mayores o menores) que los otros métodos.

10.7. Ejemplo en R-Studio

Calculamos el estadístico de prueba T

```
##Prueba Kruskal-Wallis

Metod1=c(83,91,94,89,89,96,91,92,90)
Metod2=c(91,90,81,83,84,83,88,91,89,84)
Metod3=c(101,100,91,93,96,95,94)
Metod4=c(78,82,81,77,79,81,80,81)

calif=list(g1=Metod1,g2=Metod2,g3=Metod3,g4=Metod4)

n1=length(Metod1)
n2=length(Metod2)
n3=length(Metod3)
n4=length(Metod4)
N=n1+n2+n3+n4
rangos=rank(c(Metod1,Metod2,Metod3,Metod4))
rangos2=rangos^2
R1=sum(rangos[1:9])
R2=sum(rangos[10:19])
R3=sum(rangos[20:26])
R4=sum(rangos[27:34])

S2=((N*(N+1))/12)
T=(1/S2)*((R1^2/n1)+(R2^2/n2)+(R3^2/n3)+(R4^2/n4))-3*(N+1)
T
```

```
[1] 25.46437
```

Comparamos con el cuantil

```
qchisq(0.950,3)
```

```
[1] 7.814728
```

Utilizando la función de R

```
kruskal.test(calif)
```

```
Kruskal-Wallis rank sum test
```

```
data: calif
```

```
Kruskal-Wallis chi-squared = 25.629, df = 3, p-value = 1.141e-05
```

El valor de la estadística de prueba es 25.62 y su correspondiente p -value es mucho menor a 0.01 ($\alpha = 1\%$) por lo tanto se rechaza la hipótesis nula y se concluye que al menos uno de los 4 métodos tiene distribución diferente.

Para determinar que método es el distinto se tendrán que hacer las comparaciones dos a dos.

```
Rendimiento=c(83,91,94,89,89,96,91,92,90,91,90,81,83,84,83,88,91,89,84,101,100,91,
               93,96,95,94,78,82,81,77,79,81,80,81)
Metodo=c(1,1,1,1,1,1,1,1,1,2,2,2,2,2,2,2,2,2,3,3,3,3,3,3,3,4,4,4,4,4,4,4)
pairwise.wilcox.test(Rendimiento,Metodo)
```

```
Pairwise comparisons using Wilcoxon rank sum test with continuity correction
```

```
data: Rendimiento and Metodo
```

	1	2	3
2	0.0395	-	-
3	0.0385	0.0047	-
4	0.0036	0.0047	0.0047

```
P value adjustment method: holm
```

En este caso observamos que todos los p -values de las comparaciones dos a dos son menores a 0.05 por lo tanto con un α del 5% podemos concluir que todos los métodos tienen distribuciones de los rendimientos distintas.

Capítulo 11

Prueba de Igualdad de Varianzas

Usualmente para comparar varias poblaciones nos basamos en las medias u otras medidas de ubicación de las poblaciones, en algunas situaciones las varianzas podrían ser nuestro campo de interés. Por ejemplo, se ha afirmado que el efecto de sembrar nubes con yoduro de plata podría aumentar la varianza de la lluvia resultante.

Prueba de Igualdad de Varianzas para 2 poblaciones

11.1. Datos

Consiste en dos muestras aleatorias

Sea $x_1, x_2, x_3, \dots, x_n$ una muestra aleatoria de tamaño n de la *población 1*.

Sea $y_1, y_2, y_3, \dots, y_m$ una muestra aleatoria de tamaño m de la *población 2*.

11.2. Supuestos

- 1) Ambas muestras son aleatorias de su respectiva población.
- 2) Además de la independencia dentro de cada muestra, existe una independencia mutua entre las dos muestras.
- 3) La escala de medida es al menos intervalo.

Asignación de Rangos

Modificaremos cada X_i y Y_j con el valor absoluto de la desviación de la media utilizando :

$$U_i = |X_i - \mu_1|, \quad i = 1, \dots, n$$

y

$$V_j = |Y_j - \mu_2|, \quad j = 1, \dots, m$$

donde μ_1 y μ_2 son las medias de las poblaciones 1 y 2 respectivamente.

Si μ_1 y μ_2 son desconocidas, se usará \bar{X} para μ_1 y \bar{Y} para μ_2 .

Asignamos los rangos 1 al $n + m$ a la muestra combinada de U y V de la manera habitual. Si varios valores de U y/o V son exactamente iguales entre sí (empatados), asigne a cada uno el promedio de los rangos que se les habrían asignado si no hubiera habido empates.

Sean $R(U_i)$ y $R(V_j)$ los rangos así asignados.

11.3. Estadístico de Prueba

■ Si no hay empates

Si no hay empates de la población U con la población V ocupamos:

$$T = \sum_{i=1}^n [R(U_i)]^2$$

La suma de los cuadrados de los rangos asignados a la *población* 1.

■ Si hay empates

$$T_1 = \frac{T - n\bar{R}^2}{\left[\frac{nm}{N(N-1)} \sum_{i=1}^N R_i^4 - \frac{nm}{N-1} (\bar{R}^2)^2 \right]^{\frac{1}{2}}}$$

Donde $N = n + m$, \bar{R}^2 representa el promedio de los cuadrados de los rangos de ambas muestras combinadas:

$$\bar{R}^2 = \frac{1}{N} \left(\sum_{i=1}^n [R(U_i)]^2 + \sum_{j=1}^m [R(V_j)]^2 \right)$$

y $\sum R_i^4$ representa la suma de los rangos elevados a la cuarta potencia:

$$\sum_{i=1}^N R_i^4 = \sum_{i=1}^n [R(U_i)]^4 + \sum_{j=1}^m [R(V_j)]^4$$

Los cuantiles exactos de la distribución nula de T se obtienen de las tablas correspondientes de la prueba para los casos de no tener empates y $n \leq 10$, $m \leq 10$.

Para muestras de tamaño mayor a 10 utilizamos la siguiente aproximación, basada en la estandarización normal de cuantiles z_p , obtenidas de la tabla de la distribución normal, puede ser usado la aproximación obtenida de cuantiles ω_p para T .

$$\omega_p = \frac{n(N+1)(2N+1)}{6} + z_p \sqrt{\frac{mn(N+1)(2N+1)(8N+11)}{180}}$$

Donde $N = n + m$

Dependiendo del planteamiento de nuestro problema a resolver se formulan las hipótesis:

11.4. Hipótesis

Caso A Prueba de dos colas

H_0 : X e Y son idénticamente distribuidas, excepto por medias diferentes.

vs

$$H_a : Var(x) \neq Var(Y)$$

Regla de decisión

Rechazamos H_0 al nivel de significancia α si:

$$T > t_{1-\frac{\alpha}{2}} \quad o \quad T < t_{\frac{\alpha}{2}}$$

Rechazamos H_0 al nivel de significancia α si T (o T_1 en caso de tener empates) es mayor que el cuantil $1 - \frac{\alpha}{2}$ o menor al cuantil $\frac{\alpha}{2}$ encontrados en las tablas correspondientes a la prueba (tablas de rangos al cuadrado). Y en el caso de T_1 en la tabla de distribución normal estándar.

Y calculamos el $p - value$

$$p - value = 2 * (\text{el menor } p - value \text{ de una cola})$$

donde

el $p - value$ de la cola inferior es aproximadamente:

$$p - value = \mathbf{P} \left[Z \leq \frac{T - n(N+1)(2N+1)/6}{\sqrt{mn(N+1)(2N+1)(8N+11)/180}} \right]$$

y el $p - value$ de la cola superior es aproximadamente:

$$p - value = \mathbf{P} \left[Z \geq \frac{T - n(N+1)(2N+1)/6}{\sqrt{mn(N+1)(2N+1)(8N+11)/180}} \right]$$

Caso B Prueba de cola inferior

H_0 : X e Y son idénticamente distribuidas, excepto por medias diferentes.

vs

$$H_a : Var(x) < Var(Y)$$

Regla de decisión

Rechazamos H_0 al nivel de significancia α si:

$T < t_\alpha$ (o T_1 en caso de tener empates) donde, t_α es el cuantil que se busca en tablas de la prueba. Y en el caso de T_1 en la tabla de distribución normal estándar.

y el $p - value$ se calcula:

$$p - value = \mathbf{P} \left[Z \leq \frac{T - n(N+1)(2N+1)/6}{\sqrt{mn(N+1)(2N+1)(8N+11)/180}} \right]$$

Caso C Prueba de cola superior

\mathbf{H}_0 : X e Y son idénticamente distribuidas, excepto por medias diferentes.

vs

\mathbf{H}_a : $Var(x) > Var(Y)$

Regla de decisión

Rechazamos H_0 al nivel de significancia α si:

$T > t_{1-\alpha}$ (o T_1 en caso de tener empates) donde, $t_{1-\alpha}$ es el cuantil que se busca en tablas de la prueba. Y en el caso de T_1 en la tabla de distribución normal estándar.

y el $p - value$ se calcula:

$$p - value = \mathbf{P} \left[Z \geq \frac{T - n(N+1)(2N+1)/6}{\sqrt{mn(N+1)(2N+1)(8N+11)/180}} \right]$$

Ahora vamos a resolver un ejemplo:

11.5. Ejemplo

Una cuenca hidrográfica particular se ha construido extensivamente en los últimos años, con desarrollos de vivienda, represas, etc. una muestra aleatoria de las tasas de flujo de la corriente (pies cúbicos por minuto) para una corriente en esa cuenca hidrográfica se compara con una muestra de las tasas de tiempos anteriores para ver si la variabilidad ha cambiado.

Tasas Actuales	Tasas Pasadas
32	39
36	21
41	58
27	46
35	30
48	22
31	17
28	19

Es significativa la diferencia en las varianzas? Utiliza $\alpha=0.05$

Paso 1 Prueba a utilizar **Prueba de Igualdad de Varianzas**

Paso 2 Formulamos las hipótesis

H_0 : Las tasas actuales y las tasas pasadas son idénticamente distribuidas.

vs

H_a : $Var(\text{tasas pasadas}) \neq Var(\text{tasas actuales})$

Paso 4 Procedimiento completo

Primero como μ_1 y μ_2 son desconocidas se usará \bar{X} para μ_1 y \bar{Y} para μ_2 .

Medidas Tasas Actuales (X)	Originales Tasas Pasadas (Y)	Desviación Actuales (U)	Absoluta Pasadas (V)	Rangos Actuales	Rangos Pasadas	Rangos Actuales	Al Cuadrado Pasadas
32	39	2.75	7.5	4	8	16	64
36	21	1.25	10.5	2	11	4	121
41	58	6.25	26.5	6	16	36	256
27	46	7.75	14.5	9	14.5(Empate)	81	210.25
35	30	0.25	1.5	1	3	1	9
48	22	13.25	9.5	13	10	169	100
31	17	3.75	14.5	5	14.5(Empate)	25	210.25
28	19	6.75	12.5	7	12	49	144
$\bar{X} = 34.75$	$\bar{Y} = 31.5$					$T = 381$	

$T =$ Sumas del cuadrado de los rangos (Actuales) = 381

$\bar{R}^2 = \frac{1}{16} (16 + 4 + 36 + \dots + 210.25 + 144) = 93.46$

$\sum_{i=1}^N R_i^4 = (16)^2 + (4)^2 + \dots + (210.25)^2 + (144)^2 = 243217.125$

Paso 3 Estadístico de Prueba

Debido a que encontramos empates utilizaremos el estadístico T_1 :

$$T_1 = \frac{T - n\bar{R}^2}{\left[\frac{nm}{N(N-1)} \sum_{i=1}^N R_i^4 - \frac{nm}{N-1} (\bar{R}^2)^2 \right]^{\frac{1}{2}}}$$

Entonces tenemos:

$$T_1 = \frac{381 - 8(93.46)}{\left[\frac{(8)(8)}{16(15)} (243217.125) - \frac{(8)(8)}{15} (93.46)^2 \right]^{\frac{1}{2}}}$$

$$T_1 = \frac{381 - 747.68}{\left[\frac{64}{240} (243217.125) - \frac{64}{15} (93.46)^2 \right]^{\frac{1}{2}}}$$

$$T_1 = \frac{-366.68}{[64857.9 - 37268.35]^{\frac{1}{2}}}$$

$$T_1 = \frac{-366.68}{\sqrt{27589.55}}$$

$$T_1 = \frac{-366.68}{166.10}$$

$$T_1 = -2.208282$$

Paso 5 Regla de Decisión

Rechazamos H_0 al nivel de significancia α si:

$$T_1 > t_{1-\frac{\alpha}{2}} \quad o \quad T_1 < t_{\frac{\alpha}{2}}$$

Rechazamos H_0 al nivel de significancia α si T_1 (ya que tuvimos empates) es mayor que el cuantil $1 - \frac{\alpha}{2}$ o menor al cuantil $\frac{\alpha}{2}$ encontrados en las tablas correspondientes a la distribución normal estándar.

Regresando a nuestro ejemplo:

$$T_1 > t_{0.975} \quad o \quad T_1 < t_{0.025}$$

$$-2.2082 \not> 1.96 \quad o \quad -2.2082 < -1.96$$

\therefore Entonces Rechazo H_0 .

Paso 6 Conclusión

Entonces podemos concluir que existe evidencia estadística suficiente para decir que existe diferencia significativa en las varianzas.

11.6. Ejemplo en R-Studio

Ahora haremos la réplica en R.

```
# Datos
TasasActuales = c(32,36,41,27,35,48,31,28)
TasasPasadas = c(39,21,58,46,30,22,17,19)

n=length(TasasActuales) #Tamaño de TasasActuales
m=length(TasasPasadas)  #Tamaño de TasasPasadas
N=n+m                    #Tamaño Total

# Matriz de TasasActuales y TasasPasadas
datos=matrix(c(TasasActuales,TasasPasadas), byrow=FALSE, ncol=2)
colnames(datos)=c("TasasActuales: X", "TasasPasadas: Y")
datos
```

	TasasActuales: X	TasasPasadas: Y
[1,]	32	39
[2,]	36	21
[3,]	41	58
[4,]	27	46
[5,]	35	30
[6,]	48	22
[7,]	31	17
[8,]	28	19

```
alpha=0.05 #Nivel de significancia

Xbarra=mean(TasasActuales)
cat("Media muestral de X: ", Xbarra, "\n")
```

Media muestral de X: 34.75

```
Ybarra=mean(TasasPasadas)
cat("Media muestral de Y: ", Ybarra, "\n")
```

Media muestral de Y: 31.5

```
# Valores absolutos de la desviacion de la media de X
desvX= c()
for(i in 1:n) {
  desvX[i]=abs(TasasActuales[i] - Xbarra)
}

# Valores absolutos de la desviacion de la media de Y
desvY = c()
for(i in 1:n) {
  desvY[i] = abs(TasasPasadas[i] - Ybarra)
}

# Creamos la matriz U, V
desviacion = matrix(c(desvX,desvY), byrow= FALSE, ncol=2)
colnames(desviacion) = c("U","V")
desviacion
```

	U	V
[1,]	2.75	7.5
[2,]	1.25	10.5
[3,]	6.25	26.5
[4,]	7.75	14.5
[5,]	0.25	1.5
[6,]	13.25	9.5
[7,]	3.75	14.5
[8,]	6.75	12.5

```
# Asignamos los rangos
rang = rank(desviacion)
#rang

# Los acomodamos en una nueva matriz U, V
rangos = matrix(rang, byrow=FALSE, ncol=2)
colnames(rangos) = c("Rangos U", "Rangos V")
rangos
```

	Rangos U	Rangos V
[1,]	4	8.0
[2,]	2	11.0
[3,]	6	16.0
[4,]	9	14.5
[5,]	1	3.0
[6,]	13	10.0
[7,]	5	14.5
[8,]	7	12.0

```
# Rangos de U al cuadrado
rangosU2 = c()
for(i in 1:n) {
```

```

    rangosU2[i] = (rangos[i,1])^2
}
#rangosU2

# Rangos de V al cuadrado
rangosV2 = c()
for(i in 1:n) {
    rangosV2[i] = (rangos[i,2])^2
}
#rangosV2

# Suma de los rangos al cuadrado de U
sumarangosU2 = sum(rangosU2)

# Suma de los rangos al cuadrado de V
sumarangosV2 = sum(rangosV2)

# Nueva matriz con los rangos^2
rangos2 = matrix(c(rangosU2,rangosV2), byrow=FALSE, ncol=2)
colnames(rangos2) = c("Rangos al cuadrado U", "Rangos al cuadrado V")
rangos2

```

	Rangos al cuadrado U	Rangos al cuadrado V
[1,]	16	64.00
[2,]	4	121.00
[3,]	36	256.00
[4,]	81	210.25
[5,]	1	9.00
[6,]	169	100.00
[7,]	25	210.25
[8,]	49	144.00

```

#Promedio de los rangos^2
promrangos2 = (sumarangosU2+ sumarangosV2)/N

#Rangos de U^4
rangosU4 = c()
for(i in 1:n) {
    rangosU4[i] = (rangosU2[i])^2
}
#rangosU4

#Rangos de V^4
rangosV4 = c()
for(i in 1:n) {
    rangosV4[i] = (rangosV2[i])^2
}
#rangosV4

#Nueva matriz con los rangos^4
rangos4 = matrix(c(rangosU4,rangosV4), byrow=FALSE, ncol=2)
colnames(rangos4) = c("Rangos a la cuarta U", "Rangos a la cuarta V")
rangos4

```

	Rangos a la cuarta U	Rangos a la cuarta V
[1,]	256	4096.00
[2,]	16	14641.00

```
[3,]          1296          65536.00
[4,]          6561          44205.06
[5,]           1           81.00
[6,]         28561          10000.00
[7,]           625          44205.06
[8,]         2401          20736.00
```

```
#Suma de los rangos a la cuarta^4
sumarangos4 = sum(rangosU4) + sum(rangosV4)

#Calculamos el estadístico de prueba T1 porque tenemos empates
T1=(sumarangosU2 - (n*promrangos2))/(sqrt(((n^2)/(N-1))/N)*sumarangos4)-(((n^2)/(N-1))*(pro
cat("T1 = ", T1, "\n")
```

```
T1 = -2.208273
```

```
#Regla de decision
#cat("Rechazamos H_0 a un nivel de significancia alpha=", alpha, "si T1>t1 o T1<t2, en donde t1 y t2

# Calculamos los cuantiles t1 y t2
t1 = qnorm(1-(alpha/2),mean=0,sd=1)
t2 = qnorm(alpha/2,mean=0,sd=1)

#cat("t1 = ", t1, "\n")
#cat("t2 = ", t2, "\n")

if((T1>t1) || (T1<t2)) {
  print("Rechazamos H_0")
} else {
  print("No rechazamos H_0")
}
```

```
[1] "Rechazamos H_0"
```

Ya que utilizamos la estadística T_1 , entonces el cuantil que hay que buscar es en una distribución normal. Con $\alpha = 0.05$, hace que $\frac{\alpha}{2} = 0.025$ y $1 - \frac{\alpha}{2} = 0.975$

```
t1=qnorm(.025,0,1)
t1
```

```
[1] -1.959964
```

```
t2=qnorm(.975,0,1)
t2
```

```
[1] 1.959964
```

Como $T_1 = -2.21 \leq -1.96 = t_1$ entonces se cumple la regla de decisión y rechazamos H_0 a un nivel $\alpha = 5\%$. Concluimos que existe evidencia suficiente para decir que las varianzas de las tasas de flujo de la corriente son diferentes.

```
pvalue=2*min(pnorm(T1,0,1),(1-pnorm(T1,0,1)))
pvalue
```

```
[1] 0.02722523
```

Capítulo 12

Prueba para más de dos Muestras

Si hay tres o más muestras, esta prueba se modifica fácilmente para probar la igualdad de varianzas. De cada observación se resta su media poblacional (o su media muestral (\bar{x}) cuando μ_i es desconocida) y convertimos el signo de la diferencia resultante a "+", como se acaba de describir para ambas muestras. Los rangos se asignan de menor a mayor y asignamos el promedio de los rangos correspondientes en caso de empate.

Calculamos la suma del cuadrado de los rangos asignados de cada muestra, denotamos por

S_1, S_2, \dots, S_k la suma de los rangos de cada muestra.

12.1. Datos

Se tienen k m.a que pueden tener distintos tamaños, la j –ésima m.a, de tamaño n_j es

$$x_{j1}, x_{j2}, x_{j3}, \dots, x_{jn_j}.$$

Así.

Muestra 1	Muestra 2	Muestra 3	...	Muestra k
x_{11}	x_{21}	x_{31}	...	x_{k1}
x_{12}	x_{22}	x_{32}	...	x_{k2}
\vdots	\vdots	\vdots	\ddots	\vdots
x_{1n_1}	x_{2n_2}	x_{3n_3}	...	x_{kn_k}

Sea N el número total de observaciones:

$$N = \sum_{j=1}^k n_j$$

12.2. Hipótesis

H_0 : Las k poblaciones son idénticas, excepto por diferencia en las medias.

vs

H_a : Algunas de las varianzas de las poblaciones no son idénticas entre sí.

12.3. Estadístico de Prueba

■ Cuando tenemos empates

$$T_2 = \frac{1}{D^2} \left[\sum_{j=1}^k \frac{S_j^2}{n_j} - N(\bar{S})^2 \right]$$

donde:

- n_j = Número de observaciones en la muestra j
- $N = n_1 + n_2 + \dots + n_k$
- S_j = la suma de los cuadrados de los rangos en la muestra j
- $\bar{S} = \frac{1}{N} \sum_{j=1}^k S_j$ Es el promedio de los cuadrados de todos los rangos.
- $D^2 = \frac{1}{N-1} \left[\sum_{i=1}^N R_i^4 - N(\bar{S})^2 \right]$ y $\sum R_i^4$ representa la suma de los rangos después de elevarlos a la cuarta potencia.
- **Cuando no tenemos empates**

D^2 y \bar{S} se simplifican:

$$D^2 = N(N+1)(2N+1)(8N+11)/180$$

$$\bar{S} = (N+1)(2N+1)/6$$

12.4. Regla de decisión

Rechazamos H_0 si:

$$T_2 > t_{1-\alpha}$$

donde $t_{1-\alpha}$ es el cuantil de la distribución χ^2 con $k-1$ grados de libertad donde el cuantil se localiza en las tablas de dicha distribución.

y calculamos el $p-value$

$$p-value = \mathbf{P} [\chi^2 > T_2]$$

es la probabilidad de una variable aleatoria χ^2 con $k-1$ grados de libertad sea mayor a el valor observado de T_2 .

12.5. Comparación múltiple

Si la hipótesis nula es rechazada, podríamos utilizar el siguiente procedimiento para determinar cuales pares de poblaciones tienden a ser diferentes.

Podemos decir que las poblaciones i y j parecen ser diferentes si se satisface la siguiente desigualdad:

$$\left| \frac{S_i}{n_i} - \frac{S_j}{n_j} \right| > t_{1-\alpha/2} \left(D^2 \frac{N-1-T_2}{N-k} \right)^{1/2} \left(\frac{1}{n_i} + \frac{1}{n_j} \right)^{1/2}$$

Donde $t_{1-\alpha/2}$ es el cuantil $1-\alpha/2$ de la distribución t obtenido en tablas de dicha distribución con $N-k$ grados de libertad.

Ahora resolveremos un ejercicio:

12.6. Ejemplo

Retomando en ejemplo de los 4 métodos de cultivo de maíz, probaremos ahora si las varianzas de la cantidad de cosecha por m^2 de tierra de cada terreno son diferentes.

Método 1	Método 2	Método 3	Método 4
83	91	101	78
91	90	100	82
94	81	91	81
89	83	93	77
89	84	96	79
96	83	95	81
91	88	94	80
92	91		81
90	89		
	84		

Paso 1 Prueba a utilizar **Prueba de Igualdad de Varianzas para mas de dos muestras**

Paso 2 Formulamos las hipótesis

H_0 : Los 4 métodos son idénticos, excepto por diferencias en las medias.

vs

H_a : Al menos alguno de los métodos tiene varianzas distintas

a al menos alguno de los otros métodos.

Paso 3 Estadístico de Prueba

Cuando tenemos empates

$$T_2 = \frac{1}{D^2} \left[\sum_{j=1}^k \frac{S_j^2}{n_j} - N(\bar{S})^2 \right]$$

donde:

- n_j = Número de observaciones en la muestra j
- $N = n_1 + n_2 + \dots + n_k$
- S_j = la suma de los cuadrados de los rangos en la muestra j
- $\bar{S} = \frac{1}{N} \sum_{j=1}^k S_j$ Es el promedio de los cuadrados de todos los rangos.
- $D^2 = \frac{1}{N-1} \left[\sum_{i=1}^N R_i^4 - N(\bar{S})^2 \right]$ y $\sum R_i^4$ representa la suma de los rangos después de elevarlos a la cuarta potencia.

$$T_2 = 6.006$$

Paso 5 Regla de Decisión

Rechazamos H_0 al nivel de significancia α si:

$$T_2 > t_{1-\alpha}$$

donde $t_{1-\alpha}$ es el cuantil de la distribución χ^2 con $k-1$ grados de libertad donde el cuantil se localiza en las tablas de dicha distribución.

Ya que utilizamos la estadística T_2 , entonces el cuantil que hay que buscar es en una distribución Ji-cuadrada con $k-1 = 3$ grados de libertad. Consideremos $\alpha = 0.01$.

$$t_{1-\alpha} = 11.334$$

Como $T_2 = 6.006 \not\geq 11.344 = t_{1-\alpha}$ entonces no se cumple la regla de decisión.

∴ Entonces no rechazo H_0 .

Paso 6 Conclusión

Concluimos que no hay evidencia suficiente para decir que las distribuciones de los rendimientos de los 4 métodos de cultivo no sean idénticos, excepto por diferencias en las medias.

12.7. Ejemplo en R-Studio

Ahora haremos la réplica en R.

```
M1=c(83,91,94,89,89,96,91,92,90)
M2=c(91,90,81,83,84,83,88,91,89,84)
M3=c(101,100,91,93,96,95,94)
M4=c(78,82,81,77,79,81,80,81)
n1=length(M1)
n2=length(M2)
n3=length(M3)
n4=length(M4)
N=n1+n2+n3+n4

U1=sort(abs(M1-mean(M1)))
U2=sort(abs(M2-mean(M2)))
U3=sort(abs(M3-mean(M3)))
U4=sort(abs(M4-mean(M4)))
#U=rbind(U1,U2,U3,U4)
#Asignación de rango
R1=c(3.5,3.5,5,11,12.5,12.5,25,33,34)
R2=c(14,18.5,18.5,20,23.5,23.5,26,28.5,28.5,32)
R3=c(2,6,15,21,27,30,31)
R4=c(1,7,9,9,9,16,17,22)
#rangos al cuadrado
R1s=R1^2
R2s=R2^2
R3s=R3^2
R4s=R4^2
#suma del cuadrado de los rangos
S1=sum(R1s)
S2=sum(R2s)
S3=sum(R3s)
S4=sum(R4s)

S=sum(S1,S2,S3,S4)/N
R4=sum(R1s^2)+sum(R2s^2)+sum(R3s^2)+sum(R4s^2)
D2=(R4-N*S^2)/(N-1)
T2=(S1^2/n1+S2^2/n2+S3^2/n3+S4^2/n4-N*S^2)/D2
T2
```

[1] 6.006228

Ya que utilizamos la estadística T_2 , entonces el cuantil que hay que buscar es en una distribución Ji-cuadrada con $k - 1 = 3$ grados de libertad. Consideremos $\alpha = 0.01$.

```
t=qchisq(.99,3)
t
```

```
[1] 11.34487
```

Como $T_2 = 6.006 \not\geq 11.344 = t_{1-\alpha}$ entonces no se cumple la regla de decisión y no rechazamos H_0 . Concluimos que no hay evidencia suficiente para decir que las distribuciones de los rendimientos de los 4 métodos de cultivo no sean idénticos, excepto por diferencias en las medias.

Capítulo 13

Ejercicios

1. Siete estudiantes aprendieron álgebra utilizando el método actual y seis estudiantes aprendieron álgebra según un nuevo método.

Método	Puntajes						
Actual	68	72	79	69	84	80	78
Nuevo	64	60	68	73	72	70	

¿Con una α del 5% podría decir que la efectividad (medida con los puntajes) de ambos métodos es similar?

2. En un laboratorio con entorno controlado, 10 hombres y 10 mujeres fueron evaluados para determinar la temperatura ambiente que encontraron más cómoda. los resultados fueron los siguientes:

Hombres	Mujeres
74	75
72	77
77	78
76	79
76	77
73	73
75	78
73	79
74	78
75	80

Suponiendo que estas temperaturas son una muestra aleatoria de la población, ¿la temperatura promedio es la misma para hombres y mujeres?.

3. Una muestra aleatoria de 5 diferentes marcas de focos son probados para medir la duración del foco, y los resultados fueron los siguientes:

A	B	C	D	E
73	84	82	80	85
64	80	79	85	82
67	81	71	82	80
62	77	75	86	
70		80		

Los datos observados indican una diferencia significativa entre las marcas? De haber diferencia, que marcas parecen ser las diferentes? Use $\alpha = 10\%$

4. Tres maestros desean comparar las claificaciones que pusieron en el semestre pasado para ver si alguno tiende a dar calificaciones diferentes que los otros.

Calificación	A	B	C
10	4	10	6
9	14	6	7
8	17	9	8
7	6	7	6
6	2	6	1

Con los datos observados se puede corroborar la sospecha de que alguno de los maestros da calificaciones diferentes a los otros? Use $\alpha = 5\%$

5. El Hospital “Los Ángeles” realizó un estudio en 4 Dietas a un grupo de 20 individuos; cada Dieta se aplicó aleatoriamente a cada individuo. Se registró la pérdida de peso en *kg*. Realiza la prueba correspondiente para poder concluir si, ¿Hay diferencia significativa entre la efectividad de las Dietas?

Dieta 1	Dieta 2	Dieta 3	Dieta 4
6.1	5	7.6	6.2
4.3	5.6	6.8	8
4.5	7.3	3.9	7.4
2.4	5.7	7.9	4.6
9.1	2.1	5.9	7

6. Un banco de sangre mantuvo un registro de la frecuencia cardíaca de varios donadores de sangre.

Hombres	Mujeres
58	66
76	74
82	69
74	76
79	72
65	73
74	75
86	67
	68

¿Es la variación entre los hombres significativamente mayor que la variación entre las mujeres? nivel de significancia 5%

7. Se desea probar que las variaciones de las temperaturas altas en Des Moines son mayores que las variaciones de las temperaturas altas en Spokane, para ello se tomó una muestra de las temperaturas altas diarias durante el verano. Use nivel de significancia 10%

Des Moines	Spokane
83	78
91	82
94	81
89	77
89	79
96	81
91	80
92	81
82	79
93	80
90	
93	

8. Una muestra aleatoria de 5 diferentes marcas de focos son probados para medir la duración del foco, y los resultados fueron los siguientes:

<i>A</i>	<i>B</i>	<i>C</i>	<i>D</i>	<i>E</i>
73	84	82	80	85
64	80	79	85	82
67	81	71	82	80
62	77	75	86	
70		80		

Los datos observados indican una diferencia significativa entre las varianzas de las duraciones por marca?
Use $\alpha = 10\%$

Parte IV

Tablas de Contingencia

Introducción

Estas son tablas en las que se muestran las frecuencias de observaciones medidas sobre variables que tienen diferentes clases. El objetivo es observar si dos variables son independientes entre sí. Para la realización de esta prueba las frecuencias son anotadas en tablas en las cuales cada observación es categorizada en solo una de las clases.

El uso de las tablas de contingencia fue mencionado previamente en la prueba de signos conocida como **McNemar**.

Una **tabla de contingencia** es un arreglo de números naturales en forma de matriz, donde cada número representa conteos o frecuencias.

Por ejemplo:

	Fuma	No Fuma	Total
Hombre	11	18	29
Mujer	22	13	35
Total	33	31	64

Queremos saber si influye el género en fumar.

Esta prueba nos ayuda para ver si hay alguna asociación entre 2 variables.

Algunos ejemplos:

- La asociación entre el estado nutricional de un estudiante con su desempeño académico.
- Si la preferencia por un refresco es independiente del sexo del consumidor.
- La asociación entre la región geográfica y la inversión financiera.

NOTA:

Si no hay asociación entre las variables decimos que son independientes.

En el caso de ser dependientes el valor de una variable nos ayudará a determinar el valor de la otra.

De forma general, una tabla de contingencia se construirá con r renglones y c columnas, y se llamará tabla de contingencia de $r \times c$. Éstas tablas de contingencia pueden usarse para presentar una tabulación de los datos contenidos en varias muestras, donde los datos representan al menos una escala de medición nominal.

Capítulo 14

Tablas de Contingencia de 2x2

En general una tabla de contingencia de $r \times c$ es un arreglo de números naturales que tiene r renglones y c columnas y por lo tanto tiene rc celdas o lugares para los números.

En este caso particular, $r=2$ y $c=2$, llamadas **Tablas de Contingencia de 2x2**.

Una aplicación de las tablas de contingencia de 2×2 surge cuando N objetos o personas, posiblemente seleccionadas aleatoriamente de alguna población, son clasificadas en una o dos categorías antes de aplicar un tratamiento o se produzca un evento.

Después de aplicar el tratamiento los mismos N objetos son nuevamente examinados y clasificados en las dos categorías. La pregunta a responder es ¿El tratamiento altera significativamente la proporción de objetos en cada una de las dos categorías?

El uso de las tablas de contingencia fue introducido anteriormente, y se vio que el procedimiento estadístico apropiado era una variación de la prueba de signos conocida como **McNemar**.

14.1. Datos

Una muestra aleatoria de n_1 observaciones se extrae de una población (o antes de aplicar un tratamiento) y cada observación se clasifica en la clase 1 o 2, los números totales en las dos clases son O_{11} y O_{12} respectivamente, donde $O_{11} + O_{12} = n_1$.

Una segunda muestra aleatoria de n_2 observaciones se extrae de una segunda población (o la primera población después de aplicar algún tratamiento) y el número de observaciones en la clase 1 o 2 es O_{21} y O_{22} respectivamente, donde $O_{21} + O_{22} = n_2$.

Los datos se organizan en la siguiente tabla de contingencia:

	Clase 1	Clase 2	Total
Población 1	O_{11}	O_{12}	n_1
Población 2	O_{21}	O_{22}	n_2
Total	C_1	C_2	$N = n_1 + n_2$

El número total de observaciones es denotado por N .

14.2. Supuestos

- 1) Cada muestra es una muestra aleatoria.
- 2) Las dos muestras son mutuamente independientes.
- 3) Cada observación puede clasificarse en *Clase 1* o *Clase 2*.

14.3. Estadístico de Prueba

Si alguna columna total es cero, el estadístico de prueba es definido así:

$$T_1 = 0$$

En otro caso:

$$T_1 = \frac{\sqrt{N}(O_{11}O_{22} - O_{12}O_{21})}{\sqrt{n_1 n_2 C_1 C_2}} \sim N(0, 1)$$

La distribución exacta de T_1 es difícil de array debido a todas las diferentes combinaciones de valores posibles para $O_{11}, O_{22}, O_{12}, O_{21}$.

Por lo tanto, se utiliza la aproximación de muestra grande, que es la distribución normal estándar cuyos cuantiles se dan en la tabla correspondiente.

14.4. Hipótesis

Caso A Prueba de dos colas

$$\mathbf{H}_0 : p_1 = p_2$$

vs

$$\mathbf{H}_a : p_1 \neq p_2$$

Donde p_1 es la probabilidad de elegir al azar un elemento de la *clase* 1 en la *población* 1 y p_2 es la probabilidad de elegir al azar un elemento de la *clase* 1 en la *población* 2.

Regla de Decisión

Rechazamos H_0 a un nivel de significancia α si:

$$T_1 < Z_{\frac{\alpha}{2}} \quad o \quad T_1 > Z_{1-\frac{\alpha}{2}}$$

Donde el cuantil $Z_{\frac{\alpha}{2}}$ y el cuantil $Z_{1-\frac{\alpha}{2}}$ se buscan en las tablas correspondientes de la distribución Normal.

Ahora calculamos el $p - value$ de la siguiente manera:

$$p - value = 2 * \min\{\mathbf{P}[Z < T_1], \mathbf{P}[Z > T_1]\}$$

donde Z es una normal estándar.

NOTA

Para la hipótesis anterior, también es usual usar T_1^2 en lugar de T_1 como estadístico de prueba. Entonces la región de rechazo es la cola superior de la distribución χ^2 con 1 grado de libertad, obtenidos en la tabla de la Distribución χ^2 .

Caso B Prueba de cola inferior

$$\mathbf{H}_0 : p_1 \geq p_2$$

vs

$$\mathbf{H}_a : p_1 < p_2$$

Donde p_1 es la probabilidad de elegir al azar un elemento de la *clase* 1 en la *población* 1 y p_2 es la probabilidad de elegir al azar un elemento de la *clase* 1 en la *población* 2.

Regla de Decisión

Rechazamos H_0 a un nivel de significancia α si:

$$T_1 < Z_\alpha$$

Donde el cuantil Z_α se buscan en las tablas correspondientes de la distribución Normal.

Ahora calcularemos el $p - value$ de la siguiente manera:

$$p - value = \mathbf{P}[Z < T_1]$$

donde Z es una normal estándar.

Caso C Prueba de cola superior

$$\mathbf{H}_0 : p_1 \leq p_2$$

vs

$$\mathbf{H}_a : p_1 > p_2$$

Donde p_1 es la probabilidad de elegir al azar un elemento de la *clase* 1 en la *población* 1 y p_2 es la probabilidad de elegir al azar un elemento de la *clase* 1 en la *población* 2.

Regla de Decisión

Rechazamos H_0 a un nivel de significancia α si:

$$T_1 > Z_{1-\alpha}$$

Donde el cuantil $Z_{1-\alpha}$ se buscan en las tablas correspondientes de la distribución Normal.

Ahora calcularemos el $p - value$ de la siguiente manera:

$$p - value = \mathbf{P}[Z > T_1]$$

donde Z es una normal estándar.

Ahora hagamos un ejercicio:

14.5. Ejemplo

En la academia naval se instaló un nuevo sistema de iluminación en las habitaciones guardamarinas. Se informó que el nuevo sistema de iluminación daba como resultado una vista deficiente debido a la tensión continua en los ojos de los guardias marinos. Se consideró un estudio (ficticio). Para probar la siguiente hipótesis nula. Utilizaremos $\alpha=0.05$

H_0 : La probabilidad de que un guardia marino tenga 20-20 (buena visión) es mayor o igual bajo el nuevo sistema de luces que con el viejo sistema de luces.

vs

H_a : La probabilidad de buena visión es menor con el nuevo sistema de luces que con el viejo sistema de luces.

Paso 1 Prueba a utilizar

Es una prueba **Tablas de contingencia de 2x2**

Paso 2 Planteamos las hipótesis

Es decir si planteamos nuestra hipótesis quedarían así:

Caso C Prueba de cola superior

$$H_0 : p_1 \leq p_2$$

vs

$$H_a : p_1 > p_2$$

Donde p_1 es la probabilidad de elegir al azar un guardia marino con buena visión bajo el viejo sistema de luces y p_2 es la probabilidad de elegir al azar un guardia marino con buena visión con el nuevo sistema de luces.

Paso 3 La tabla de contingencia, en este caso es:

	Buena Visión	Mala Visión	Total
Luces viejas	$O_{11} = 714$	$O_{12} = 111$	$n_1 = 825$
Luces nuevas	$O_{21} = 662$	$O_{22} = 154$	$n_2 = 816$
Total	$C_1 = 1376$	$C_2 = 265$	$N = n_1 + n_2 = 1641$

Paso 4 Estadístico de Prueba

$$T_1 = \frac{\sqrt{N}(O_{11}O_{22} - O_{12}O_{21})}{\sqrt{n_1 n_2 C_1 C_2}}$$

$$T_1 = \frac{\sqrt{1641}[(714)(154) - (111)(662)]}{\sqrt{(825)(816)(1376)(265)}}$$

$$T_1 = 2.9821$$

Paso 5 Regla de decisión

Rechazamos H_0 a un nivel de significancia α si:

$$T_1 > Z_{1-\alpha}$$

$$T_1 = 2.9821 > Z_{0.95} = 1.6449$$

Donde el cuantil $Z_{1-\alpha}$ se buscan en las tablas correspondientes de la distribución Normal.

\therefore Rechazo H_0 .

Ahora calculamos el p - *value* la probabilidad de que Z sea mayor que el valor observado de T_1 , de la tabla de la Distribución Normal.

$$p - value = \mathbf{P}[Z > T_1] = \mathbf{P}[Z > 2.9821] = 1 - P[Z \leq 2.9821] = 0.00143$$

$$\therefore p - value = 0.00143 < 0.05 = \alpha$$

\therefore Rechazo H_0 . Entonces podemos concluir que la probabilidad de buena visión es menor con el nuevo sistema de luces que con el viejo sistema de luces.

14.6. Ejemplo en R-Studio

Ahora haremos la réplica en R.

```
Guardias=c("LucesViejas","LucesNuevas")
BuenaVision=c(714,662)
MalaVision=c(111,154)
datos=data.frame(Guardias,BuenaVision,MalaVision)
datos
```

```
      Guardias BuenaVision MalaVision
1 LucesViejas         714         111
2 LucesNuevas         662         154
```

```
Guardias=c(Guardias,"Total")
BuenaVision=c(BuenaVision, sum(BuenaVision))
MalaVision=c(MalaVision, sum(MalaVision))
Totales=c(BuenaVision[1]+MalaVision[1],BuenaVision[2]+MalaVision[2],
          BuenaVision[1]+MalaVision[1]+BuenaVision[2]+MalaVision[2])
datos1=data.frame(Guardias,BuenaVision,MalaVision,Totales)
datos1
```

```
      Guardias BuenaVision MalaVision Totales
1 LucesViejas         714         111      825
2 LucesNuevas         662         154      816
3      Total        1376         265     1641
```

#Datos

```
011=714      #Guardias Buena visión/Luces viejas
012=111      #Guardias Mala vision/Luces viejas
021=662      #Guardias Buena vision/Luces nuevas
022=154      #Guardias Mala vision/Luces nuevas

n1=825      #Total luces viejas
```

```
n2=816      #Total luces nuevas
C1=1376     #Total buena vision
C2=265      #Total Mala vision
N=n1+n2     #Tamaño total
#N
```

Calcularemos el Estadístico de prueba:

```
T_1=(sqrt(N)*(O11*O22-O12*O21))/(sqrt(n1*n2*C1*C2))
T_1
```

```
[1] 2.982177
```

Calculamos el cuantil y el p -value:

```
cuantil=qnorm(0.95) #Cuantil a comparar
cuantil
```

```
[1] 1.644854
```

```
pvalue=1-pnorm(T_1) #Calculamos el p-value
pvalue
```

```
[1] 0.00143103
```

Como el estadístico $T_1 = 2.98$ es mayor a el cuantil $Z_{1-\alpha} = 1.64$ por lo tanto rechazamos H_0 y el p -value es menor a nuestra $\alpha = 5\%$; Entonces podemos concluir que la probabilidad de buena visión es menor con el nuevo sistema de luces que con el viejo sistema de luces.

Ahora podemos utilizar la prueba en R:

```
tabla<-matrix(c(714,111,825,662,154,816),ncol=3,byrow=TRUE)
tabla
```

```
      [,1] [,2] [,3]
[1,]  714  111  825
[2,]  662  154  816
```

```
chisq.test(tabla)
```

```
Pearson's Chi-squared test
```

```
data:  tabla
X-squared = 8.8934, df = 2, p-value = 0.01172
```

Capítulo 15

Prueba de Independencia

15.1. Datos

Una muestra aleatoria de tamaño N . Las observaciones en la muestra aleatoria son clasificados de acuerdo a dos criterios, usando el primer criterio cada observación es asociada con uno de los r renglones, y usando el segundo criterio cada observación es asociada con una de las C columnas. Sea O_{ij} el número de observaciones asociadas con el renglón i y la columna j simultáneamente; el número total de observaciones en el renglón i es designado por R_i , (en lugar de n_i como la prueba anterior, para enfatizar que los totales de las filas ahora son aleatorios en lugar de fijos), y en la columna j por C_j . La suma de los números en todas las celdas es N .

Los datos se organizan en la siguiente tabla de contingencia:

	Columna 1	Columna 2	...	Columna c	Totales
Renglón 1	O_{11}	O_{12}	...	O_{1c}	R_1
Renglón 2	O_{21}	O_{22}	...	O_{2c}	R_2
\vdots	\vdots	\vdots	\ddots	\vdots	\vdots
Renglón r	O_{r1}	O_{r2}	...	O_{rc}	R_r
Totales	C_1	C_2	...	C_c	N

15.2. Supuestos

- La muestra de N observaciones es una muestra aleatoria. (Cada observación tiene la misma probabilidad como cualquier otra observación de ser clasificada en el renglón i y la columna j independientemente de las otras observaciones).
- Cada observación puede ser clasificada en una de las r diferentes categorías de acuerdo al primer criterio y en una de las C diferentes categorías de acuerdo al segundo criterio.

15.3. Estadístico de Prueba

El estadístico de prueba T es obtenido de la siguiente manera:

$$T = \sum_{i=1}^r \sum_{j=1}^c \frac{(O_{ij} - E_{ij})^2}{E_{ij}}, \quad \text{Donde } E_{ij} = \frac{R_i C_j}{N}$$

Una expresión equivalente para T , mas adecuado para el uso de su calculadora es:

$$T = \sum_{i=1}^r \sum_{j=1}^c \frac{O_{ij}^2}{E_{ij}} - N$$

Distribución de T

La distribución nula de T es obtenida aproximadamente por la Distribución χ^2 con $(r-1) \times (c-1)$ grados de libertad, cuyos cuantiles se encuentran en las tablas de dicha distribución.

Es decir,

$$T \sim \chi_{(r-1)(c-1)}^2$$

15.4. Hipótesis

H_0 : El evento "una observación está en la fila i " es independiente del evento "esa misma observación está en la columna j " para todos i, j .

es decir,

$$H_0 : P[\text{ renglón } i, \text{ columna } j] = P[\text{ renglón } i] * P[\text{ columna } j] \quad \forall i, j.$$

vs

$$H_a : P[\text{ renglón } i, \text{ columna } j] \neq P[\text{ renglón } i] * P[\text{ columna } j] \quad ; \text{ para alguna } i, j.$$

Regla de decisión

Rechazo H_0 si $T > \chi^2(1-\alpha)$ con $(r-1) \times (c-1)$ grados de libertad, cuyos cuantiles se encuentran en las tablas de dicha distribución.

Ahora un ejercicio:

15.5. Ejercicio

Se especula que la preferencia del cereal ToastyOs está asociada con el nivel educativo de las personas. Si esto resulta cierto, Tabisco Food, la distribuidora del producto, siente que debería aprovechar este mercado dando un mayor empuje de su campaña de marketing a este segmento de la población. Sin embargo, antes de comprometerse a esta tarea se decidió realizar un análisis objetivo que verifiquen las especulaciones. Para esto se tomó una muestra aleatoria de 500 individuos que han probado el producto con los siguientes resultados:

Preferencia	Sin Universidad	Nivel de Estudios Truncados	Graduados
Gusta	75	90	135
Neutral/No les gusta	25	60	115

Paso 1 Prueba a utilizar Tablas de Contingencia rxc prueba de Independencia

Es una prueba de independencia ya que sólo tenemos una población.

Paso 2 Planteamiento de hipótesis

H_0 : El nivel educativo es independiente de la preferencia hacia el producto.

vs

H_a : El nivel educativo no es independiente de la preferencia hacia el producto.

Paso 3 Estadístico de Prueba

Para poder calcular el estadístico de prueba no ayudaremos de ciertos cálculos previos.

Preferencia	Sin Universidad	Nivel de Estudios Truncados	Graduados	Totales
Gusta	75	90	135	300
Neutral/No les gusta	25	60	115	200
Totales	100	150	250	500

Ahora:

$$P[\text{Guste el producto}] = \frac{300}{500} = .6 = 60\% \text{ de la población le gusta el cereal}$$

$$P[\text{No Guste el producto}] = \frac{200}{500} = .4 = 40\% \text{ de la población no le gusta el cereal}$$

$$P[\text{Sin universidad}] = \frac{100}{500} = .2 = 20\% \text{ de la población no tiene universidad.}$$

$$P[\text{Estudios truncados}] = \frac{150}{500} = .3 = 30\% \text{ de la población tiene estudios truncados.}$$

$$P[\text{Graduados}] = \frac{250}{500} = .5 = 50\% \text{ de la población están graduados.}$$

Bajo H_0 el nivel educativo y la preferencia por el cereal son independientes:

Ahora:

$$P[\text{Guste el producto y sin universidad}] = P[\text{Guste el producto}] * P[\text{sin universidad}] = .6 * .2 = .12$$

$$P[\text{Guste el producto y Estudios truncados}] = P[\text{Guste el producto}] * P[\text{Estudios truncados}] = .6 * .3 = .18$$

$$P[\text{Guste el producto y Graduados}] = P[\text{Guste el producto}] * P[\text{Graduados}] = .6 * .5 = .30$$

Análogo para no gustar el producto.

Ahora vamos a calcular los valores esperados:

$$E_{11} = \frac{300 * 100}{500} = 60 \quad E_{12} = \frac{300 * 150}{500} = 90 \quad E_{13} = \frac{300 * 250}{500} = 150$$

$$E_{21} = \frac{200 * 100}{500} = 40 \quad E_{22} = \frac{200 * 150}{500} = 60 \quad E_{23} = \frac{200 * 250}{500} = 100$$

Regresando al Estadístico de Prueba:

$$T = \sum_{i=1}^2 \sum_{j=1}^3 \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = \frac{(75 - 60)^2}{60} + \frac{(90 - 90)^2}{90} + \frac{(135 - 150)^2}{150} +$$

$$\frac{(25 - 40)^2}{40} + \frac{(60 - 60)^2}{60} + \frac{(115 - 100)^2}{100} = 13.125$$

Paso 4 Procedimiento completo

Utilizaremos un $\alpha = 0.05$

Entonces tenemos $\chi^2_{(.95, (2-1)*(3-1))} = \chi^2_{(.95, 2)} = 5.99$

Paso 5 Regla de decisión

Rechazo H_0 si $T > \chi^2(1 - \alpha)$ con $(r - 1) \times (c - 1)$ grados de libertad, cuyos cuantiles se encuentran en las tablas de dicha distribución.

como $\chi^2 = 5.99 < T = 13.125$

\therefore Rechazo H_0 .

Paso 6 Conclusión

Existe evidencia suficiente para suponer que hay una relación entre la preferencia del cereal y el nivel educativo.

15.6. Ejemplo en R-Studio

Ahora haremos la réplica en R.

```
Preferencia=c("Gusta","Neutral/No les gusta")
Sin_Universidad=c(75,25)
Nivel_Truncado=c(90,60)
Graduados=c(135,115)
datos=data.frame(Preferencia,Sin_Universidad,Nivel_Truncado,Graduados)
datos
```

	Preferencia	Sin_Universidad	Nivel_Truncado	Graduados
1	Gusta	75	90	135
2	Neutral/No les gusta	25	60	115

#ahora, vamos a calcular los totales y adjuntarlos a la tabla

```
Preferencia=c(Preferencia,"Totales")
Sin_Universidad=c(Sin_Universidad, sum(Sin_Universidad))
Nivel_Truncado=c(Nivel_Truncado, sum(Nivel_Truncado))
Graduados=c(Graduados, sum(Graduados))
Totales=c(Sin_Universidad[1]+Nivel_Truncado[1]+Graduados[1],Sin_Universidad[2]+Nivel_Truncado[2]+Gra
Sin_Universidad[1]+Nivel_Truncado[1]+Graduados[1]+Sin_Universidad[2]+Nivel_Truncado[2]+Gr
datos=data.frame(Preferencia,Sin_Universidad,Nivel_Truncado,Graduados,Totales)
datos
```

	Preferencia	Sin_Universidad	Nivel_Truncado	Graduados	Totales
1	Gusta	75	90	135	300
2	Neutral/No les gusta	25	60	115	200
3	Totales	100	150	250	500

#ESTADÍSTICO DE PRUEBA

#Primero calcularemos los valores esperados

```
Esperados=matrix(nrow = 2,ncol = 3)
for (i in 1:2) {
  for (j in 1:3) {
    Esperados[i,j]=(datos$Totales[i]*datos[3,j+1])/datos$Totales[3]
  }
}
#Ahora sí, vamos a calcular el estadístico completo
Totales=matrix(nrow = 2,ncol = 3)
for (i in 1:2) {
  for (j in 1:3) {
    Totales[i,j]=(datos[i,j+1]^2)/(Esperados[i,j])
  }
}
```



```

}
T1=sum(Totales)-datos$Totales[3]
print(c("Estadístico T1 = ", T1))

```

```
[1] "Estadístico T1 = " "13.125"
```

```

alpha=.05
#Ahora vamos a calcular nuestro cuantil
cuantil=qchisq(1-alpha,2)
print(c("Cuantil Ji-cuadrado con 2 grados de libertad = ", cuantil))

```

```
[1] "Cuantil Ji-cuadrado con 2 grados de libertad = " "5.99146454710798"
```

```

if(T1>cuantil){
  print("Se rechaza $H_0$")
  print(c("Estadístico T1 = ", T1))
  print(c("Cuantil Ji-cuadrado con 2 grados de libertad = ", cuantil))
}

```

```
[1] "Se rechaza $H_0$"
```

```
[1] "Estadístico T1 = " "13.125"
```

```
[1] "Cuantil Ji-cuadrado con 2 grados de libertad = " "5.99146454710798"
```

Observamos que la estadística de prueba tiene un valor de 13.125 y su correspondiente p-value es menor a 0.05, por lo tanto con $\alpha = 5\%$ rechazaremos H_0 y concluimos existe evidencia suficiente para suponer que hay una relación entre la preferencia del cereal y el nivel educativo.

Ahora podemos utilizar la prueba en R:

```

tabla<-matrix(c(75,90,135,25,60,115),ncol=3,byrow=TRUE)
dimnames(tabla)<- list(Preferencia=c("Gustar","Neutral o no les gusta"), Nivel=c("Sin universidad","Graduados"))

```

```
tabla
```

	Nivel		
Preferencia	Sin universidad	Estudios truncados	Graduados
Gustar	75	90	135
Neutral o no les gusta	25	60	115

```
chisq.test(tabla)
```

Pearson's Chi-squared test

```
data: tabla
```

```
X-squared = 13.125, df = 2, p-value = 0.001412
```

Capítulo 16

Tablas de Contingencia de $r \times c$

Como una generalización inmediata de la tabla de contingencia 2×2 mencionadas anteriormente, tenemos la tabla de contingencia con r renglones y c columnas, llamada **tablas de contingencia de $r \times c$** . Éstas tablas de contingencia pueden usarse, como en la sección anterior, para presentar una tabulación de los datos contenidos en varias muestras, donde los datos representan al menos una escala de medición nominal, y para probar la hipótesis de que las probabilidades no difieren de muestra en muestra.

Prueba de χ^2 para Tablas de Contingencia (Proporciones)

16.1. Datos

Hay r poblaciones en total, y se extrae una muestra aleatoria de cada población. Supongamos que n_i representa el número de observaciones en la muestra i -ésima (de la población i -ésima) para $1 \leq i \leq r$. Cada observación en cada muestra es clasificada en una de las C diferentes categorías.

Sea O_{ij} el número de observaciones de la i -ésima muestra de la categoría j :

$$n_i = O_{i1} + O_{i2} + \cdots + O_{ic} \quad \forall \ i.$$

Los datos se organizan en la siguiente tabla de contingencia de $r \times c$:

	Clase 1	Clase 2	...	Clase c	Total
Población 1	O_{11}	O_{12}	...	O_{1c}	n_1
Población 2	O_{21}	O_{22}	...	O_{2c}	n_2
...
Población r	O_{r1}	O_{r2}	...	O_{rc}	n_r
Total	C_1	C_2	...	C_c	N

El número total de observaciones de todas las muestras es denotado por N :

$$N = n_1 + n_2 + \cdots + n_r.$$

El número de observaciones en la j -ésima columna denotada como C_j . Esto es, C_j es el número total de observaciones en la j -ésima categoría, o clase, de todas las muestras combinadas.

$$C_j = O_{1j} + O_{2j} + \cdots + O_{rj}, \quad \text{para } j = 1, 2, \dots, c.$$

16.2. Supuestos

- Cada muestra es una muestra aleatoria.
- Los resultados de las diversas muestras son mutuamente independientes (particularmente entre las muestras, porque la independencia dentro de las muestras es parte del primer supuesto).
- Cada observación puede clasificarse en exactamente una de las categorías o clases C .

16.3. Estadístico de Prueba

El estadístico de prueba T es obtenido de la siguiente manera:

$$T = \sum_{i=1}^r \sum_{j=1}^c \frac{(O_{ij} - E_{ij})^2}{E_{ij}}, \quad \text{Donde } E_{ij} = \frac{n_i C_j}{N}$$

Mientras el término O_{ij} representa el número de observaciones en la celda (i, j) , el término E_{ij} representa el número de observaciones esperadas en la celda (i, j) . Si H_0 es realmente verdadera, es decir, si H_0 es cierta el número de observaciones en la celda (i, j) podrían estar cerca a la i -ésima muestra de tamaño n_i multiplicado por la proporción $\frac{C_j}{N}$ de todas las observaciones en la categoría j .

NOTA:

En el caso de 2×2 el estadístico T es equivalente a T^2 visto anteriormente, porque solo se considera la hipótesis alternativa de dos colas.

Una expresión equivalente para T , mas adecuado para el uso de su calculadora es:

$$T = \sum_{i=1}^r \sum_{j=1}^c \frac{O_{ij}^2}{E_{ij}} - N$$

Distribución de T

La distribución nula de T es obtenida aproximadamente por la Distribución χ^2 con $(r - 1) \times (c - 1)$ grados de libertad, cuyos cuantiles se encuentran en las tablas de dicha distribución.

Es decir,

$$T \sim \chi_{(r-1)(c-1)}^2$$

16.4. Hipótesis

Sea la probabilidad de que un valor seleccionado aleatoriamente de la i -ésima población se clasifique en la j -ésima clase, denotado por p_{ij} para $i = 1, 2, \dots, r$ y $j = 1, 2, \dots, c$.

H_0 : Todas las probabilidades en la misma columna son iguales entre sí.

vs

H_a : Al menos dos de las probabilidades en la misma columna no son iguales entre sí.
en otros términos:

$$H_0 : p_{1j} = p_{2j} = \dots = p_{rj} \quad \forall j.$$

vs

$$H_a : p_{ij} \neq p_{kj} \text{ para algún par } i \text{ y } k.$$

Regla de decisión

Rechazo H_0 si $T > \chi^2(1 - \alpha)$ con $(r - 1) \times (c - 1)$ grados de libertad, cuyos cuantiles se encuentran en las tablas de dicha distribución.

Ahora aplicaremos los conocimientos en el próximo ejemplo:

16.5. Ejercicio

Una muestra de estudiantes seleccionados aleatoriamente de escuelas secundarias privadas recibió pruebas de rendimiento estandarizadas con los siguientes resultados:

	Puntaje de Prueba				
	0-275	276-350	351-425	426-500	Totales
Escuela Privada	6	14	17	9	46
Escuela Pública	30	32	17	3	82
Totales	36	46	34	12	128

Paso 1 Prueba a utilizar: **Tablas de Contingencia de $r \times c$**

Paso 2 Plantear hipótesis:

H_0 : La distribución del puntaje de los estudiantes en la prueba es la misma para la escuela privada como para la escuela pública.

vs

H_a : La distribución del puntaje de los estudiantes en la prueba es distinta para la escuela privada como para la escuela pública.

Para poder calcular nuestro estadístico de prueba podemos auxiliarnos de una pequeña tablita para

■ E_{ij} :

	Columnas			
	1	2	3	4
Renglón 1	12.9	16.5	12.2	4.3
Renglón 2	23.1	29.5	21.8	7.7

Paso 3 Estadístico de Prueba

Para la celda en el *renglón 1, columna 1* tenemos:

$$\frac{(O_{ij} - E_{ij})^2}{E_{ij}} = \frac{(O_{11} - E_{11})^2}{E_{11}} = \frac{(6 - 12.9)^2}{12.9} = \frac{47.61}{12.9} = 3.69$$

Si hacemos los calculos celda por celda el resultado es:

$$T = 3.69 + 0.38 + 1.89 + 5.14 + 2.06 + 0.21 + 1.06 + 2.87 = 17.3$$

Paso 4 Procedimiento completo

Ahora buscaremos el cuantil: $\chi^2(1-\alpha)_{(r-1)(c-1)}$ donde $(r-1) \times (c-1) = (2-1) \times (4-1) = 3$ entonces buscamos $\chi^2_{(3)} = 7.815$

Paso 5 Regla de desición :

Rechazo H_0 si $T > \chi^2(1-\alpha)$ y $17.3 > 7.815$ entonces Rechazo H_0 .

Paso 6 Conclusión:

Entonces la distribución del puntaje de los estudiantes en la prueba es distinta para la escuela privada como para la escuela pública.

16.6. Ejemplo en R-Studio

Ahora haremos la réplica en R.

```
data <- matrix(c(6,14,17,9,46,30,32,17,3,82,36,46,34,12,128),nrow=3,byrow=T,
               dimnames=list("Escuelas"=c("EscuelaPrivada","EscuelaPublica","Totales"),
                             "Puntajes"=c("cat1","cat2","cat3","cat4","Totales")))
data
```

Escuelas	Puntajes				Totales
	cat1	cat2	cat3	cat4	
EscuelaPrivada	6	14	17	9	46
EscuelaPublica	30	32	17	3	82
Totales	36	46	34	12	128

#Datos

```
011=6
012=14
013=17
014=9
021=30
022=32
023=17
024=3
n1=46
n2=82
C1=36
C2=46
C3=34
C4=12
N=n1+n2
```

#Ahora calcularemos los Eij

```
E11=n1*C1/N
E12=n1*C2/N
E13=n1*C3/N
E14=n1*C4/N
E21=n2*C1/N
```

```
E22=n2*C2/N
E23=n2*C3/N
E24=n2*C4/N
```

```
#Ahora calcularemos los  $(O_{ij}-E_{ij})^2/E_{ij}$ 
```

```
j1=(O11-E11)^2/E11
j2=(O12-E12)^2/E12
j3=(O13-E13)^2/E13
j4=(O14-E14)^2/E14
j5=(O21-E21)^2/E21
j6=(O22-E22)^2/E22
j7=(O23-E23)^2/E23
j8=(O24-E24)^2/E24
```

```
#Ahora calculamos el estadístico de prueba
```

```
T=sum(j1,j2,j3,j4,j5,j6,j7,j8)
T
```

```
[1] 17.28581
```

```
#El cuantil es de una distribución ji.cuadrada con  $(r-1)(c-1)$  grados de libertad en este caso serian
```

```
qchisq(0.95,3)
```

```
[1] 7.814728
```

Observamos que la estadística de prueba tiene un valor de 17.28 y su correspondiente p -value es menor a 0.05, por lo tanto con $\alpha = 5\%$ rechazaremos H_0 y concluimos existe evidencia suficiente para suponer la distribución de los puntajes en la prueba es distinta entre las escuelas privadas y las escuelas públicas.

Ahora podemos utilizar la prueba en R:

```
tabla<-matrix(c(6,14,17,9,46,30,32,17,3,82),ncol=5,byrow=TRUE)
tabla
```

```
      [,1] [,2] [,3] [,4] [,5]
[1,]    6   14   17    9   46
[2,]   30   32   17    3   82
```

```
chisq.test(tabla)
```

```
Pearson's Chi-squared test
```

```
data:  tabla
X-squared = 17.286, df = 4, p-value = 0.001701
```

Capítulo 17

Prueba de la Mediana

La prueba mediana está diseñada para examinar si varias muestras provienen de poblaciones que tienen la misma mediana. En realidad, la prueba de la mediana no es nueva, es simplemente una aplicación especial de la prueba de ji cuadrado con totales marginales fijos. Sin embargo, es una aplicación muy útil y consideramos que vale la pena un trato especial.

Para probar si varias c poblaciones tienen la misma mediana, se extrae una muestra aleatoria de cada población (la escala de medición es al menos ordinal, o el término “mediana” no tendría sentido). Se construye una tabla de contingencia $2 \times c$ y las dos entidades en la i –ésima columna son los números de observaciones en la i –ésima muestra que están por encima y por debajo de la gran mediana (la mediana de todas las observaciones combinadas). La prueba de χ^2 habitual se aplica luego a la tabla de contingencia.

17.1. Datos

Para cada una de las poblaciones c se obtiene una muestra aleatoria de tamaño n_i , $i = 1, 2, \dots, c$. Se determina la mediana de la muestra combinada; es decir, se determina el número que excedió aproximadamente la mitad de las observaciones en toda la muestra $N = n_1 + n_2 + \dots + n_c$. Esto se llama mediana. Sea O_{1i} , el número de observaciones en la i –ésima muestra que excede la mediana y, sea O_{2i} el número en la i –ésima muestra que sea menor o igual a la mediana, organice los conteos de frecuencia en una tabla de contingencia $2 \times c$ de la siguiente manera:

	Muestra 1	Muestra 2	\dots	Muestra c	Totales
>Mediana	O_{11}	O_{12}	\dots	O_{1c}	a
\leq Mediana	O_{21}	O_{22}	\dots	O_{2c}	b
Totales	n_1	n_2	\dots	n_c	N

Sea a el número total de observaciones mas grandes a la mediana en todas las muestras, sea b el número total de observaciones menores o iguales a la mediana. Entonces $a + b = N$ es el número total de observaciones.

17.2. Supuestos

- Cada muestra es una muestra aleatoria.
- Las muestras son independientes entre si.
- La escala de medida es al menos ordinal.
- Si todas las poblaciones tienen la misma mediana, todas las poblaciones tienen la misma probabilidad p de que una observación exceda la mediana.

17.3. Estadístico de Prueba

El estadístico de prueba es obtenido por el reordenamiento del estadístico utilizado en la prueba anterior, notando que $O_{2i} = n_i - O_{1i}$ en el caso especial con 2 renglones.

$$T = \frac{N^2}{ab} * \sum_{i=1}^c \frac{(O_{1i} - \frac{n_i a}{N})^2}{n_i}$$

Para ahorrarnos algunos cálculos:

$$T = \frac{N^2}{ab} * \sum_{i=1}^c \frac{O_{1i}^2}{n_i} - \frac{Na}{b}$$

NOTA:

Si a es exactamente igual a b nuestro estadístico de prueba se simplifica en:

$$T = \sum_{i=1}^c \frac{(O_{1i} - O_{2i})^2}{n_i}$$

Distribución de T

La distribución nula de T es obtenida aproximadamente por la Distribución χ^2 con $(c - 1)$ grados de libertad, cuyos cuantiles se encuentran en las tablas de dicha distribución.

Es decir,

$$T \sim \chi_{(c-1)}^2$$

17.4. Hipótesis

H_0 : Todas las c poblaciones tienen la misma mediana.

vs

H_a : Al menos 2 de las poblaciones tienen diferente mediana.

Regla de decisión

Rechazo H_0 si $T > \chi^2(1 - \alpha)$ con $(c - 1)$ grados de libertad, cuyos cuantiles se encuentran en las tablas de dicha distribución.

17.5. Comparación Múltiple

Si la hipótesis nula es rechazada, se pueden hacer comparaciones múltiples por parejas entre poblaciones utilizando la prueba de mediana repetidamente en tablas de contingencia de 2×2 . En cada comparación se encuentra la mediana de las dos muestras, y el número por encima o por debajo de esa mediana se utiliza en la tabla de contingencia 2×2 . Calculamos el estadístico de prueba T para la prueba y si T es más grande que el cuantil $1 - \alpha$ de la distribución χ^2 con 1 grado de libertad encontrado en tablas de dicha distribución, entonces decimos que las medianas de esas 2 poblaciones son iguales.

Aplicaremos nuestros conocimientos en el siguiente ejercicio:

17.5.1. Ejemplo

Se asignaron al azar cuatro métodos diferentes de cultivo de maíz a un gran número de parcelas diferentes y se calculó el rendimiento por acre para cada parcela. Los datos son los siguientes:

Método 1	Método 2	Método 3	Método 4
83	91	101	78
91	90	100	82
94	81	91	81
89	83	93	77
89	84	96	79
96	83	95	81
91	88	94	80
92	91		81
90	89		
	84		

Para determinar si existe una diferencia en los rendimientos como resultado del método utilizado, se empleó la prueba mediana porque se consideró que una diferencia en las medianas de la población podría interpretarse como una diferencia en el valor del método utilizado.

Paso 1 Prueba a utilizar **Tablas de contingencia, Prueba de la mediana**

Paso 2 Planteamiento de Hipótesis

H_0 : Todos los métodos tienen el mismo rendimiento medio (mediana) por acre.

vs

H_a : Al menos dos de los métodos difieren con respecto al rendimiento medio(mediana) por acre.

En el conteo rápido revela que hay 34 observaciones en total, por lo que el promedio de las observaciones más pequeñas decimoséptima y decimoctava es la mediana, y se ve que es 89. Luego, para cada método (muestra), el número de valores que exceden 89 y el número que es menor o igual a 89 se registra en la siguiente forma.

	Método 1	Método 2	Método 3	Método 4	Totales
>89	6	3	7	0	16
≤ 89	3	7	0	8	18
Totales	9	10	7	8	34

Paso 3 Estadístico de Prueba

$$T = \frac{N^2}{ab} * \sum_{i=1}^c \frac{(O_{1i} - \frac{n_i a}{N})^2}{n_i}$$

$$T = \frac{(34)^2}{(16)(18)} * \left(\frac{\left[6 - \frac{(9)(16)}{34}\right]^2}{9} + \frac{\left[3 - \frac{(10)(16)}{34}\right]^2}{10} + \frac{\left[7 - \frac{(7)(16)}{34}\right]^2}{7} + \frac{\left[0 - \frac{(8)(16)}{34}\right]^2}{8} \right)$$

$$T = 4.01(0.34 + 0.29 + 1.97 + 1.78) = 17.6$$

Paso 4 Regla de decisión

Rechazo H_0 si $T > \chi^2(1 - \alpha)$ con $(c - 1)$ grados de libertad, cuyos cuantiles se encuentran en las tablas de dicha distribución.

Tomaremos un $\alpha=0.05$, entonces $\chi^2(.95)$ con 3 grados de libertad es 7.815, obtenido en las tablas correspondientes a la Distribución χ^2 .

Como $T = 17.6 > 7.815 = \chi^2(.95)$ rechazamos H_0 .

Paso 5 Conclusión

Como rechazo H_0 , entonces al menos 2 de los métodos difieren con respecto al rendimiento medio (mediana) por acre.

Como se rechazó H_0 , se le deja al lector hacer la comparación múltiple.

17.6. Ejemplo en R-Studio

Ahora haremos la réplica en R.

```
#Datos
Rendimiento=c(83,91,94,89,89,96,91,92,90,91,90,81,83,84,83,88,91,89,84,101,100,91,
              93,96,95,94,78,82,81,77,79,81,80,81)
Metodo=c(1,1,1,1,1,1,1,1,2,2,2,2,2,2,2,2,3,3,3,3,3,3,4,4,4,4,4,4,4)

#Calcular la mediana
mediana=median(Rendimiento)
mediana
```

```
[1] 89
```

Después de calcular la mediana, contamos cuantas observaciones hay por método por arriba y por abajo de la mediana y con dichas frecuencias construimos la tabla de contingencia.

```
#Construir la tabla de contingencia
table(Metodo[which(Rendimiento>mediana)])
```

```
1 2 3
6 3 7
```

```
table(Metodo[which(Rendimiento<=mediana)])
```

```
1 2 4
3 7 8
```

```
Observados=matrix(c(6,3,7,0,3,7,0,8), nrow = 2, ncol = 4, byrow = T)
rownames(Observados)=c('Mayor_med','MenorIgual_med')
colnames(Observados)=c('Metodo 1','Metodo 2','Metodo 3','Metodo 4')
Observados
```

	Metodo 1	Metodo 2	Metodo 3	Metodo 4
Mayor_med	6	3	7	0
MenorIgual_med	3	7	0	8

Ya que se tiene la tabla se le puede aplicar la prueba. En este caso observaremos que la prueba nos advierte sobre el uso de la estadística Ji-Cuadrada, lo anterior se debe a que las frecuencias en la tabla de contingencia son pequeñas e incluso cero. La prueba alternativa en estos casos es la prueba exacta de Fisher.

```
#Prueba
```

```
T1 <- chisq.test(Observados)
T1
```

Pearson's Chi-squared test

```
data: Observados
X-squared = 17.543, df = 3, p-value = 0.0005464
```

```
T2 <- fisher.test(Observados)
T2
```

Fisher's Exact Test for Count Data

```
data: Observados
p-value = 0.0001631
alternative hypothesis: two.sided
```

Con ambas pruebas se llega a la conclusión de rechazar H_0 con $\alpha = 0.01$ y por lo tanto al menos dos de los métodos difieren con respecto al rendimiento medio (mediana) por m^2 .

Podemos ver los valores esperados con:

```
#Valores de las frecuencias esperadas
T1$expected
```

	Metodo 1	Metodo 2	Metodo 3	Metodo 4
Mayor_med	4.235294	4.705882	3.294118	3.764706
MenorIgual_med	4.764706	5.294118	3.705882	4.235294

Podemos realizar las comparaciones múltiples. A continuación se presenta la comparación de los métodos 2 y 3.

```
#Datos
```

```
Rendimiento=c(91,90,81,83,84,83,88,91,89,84,101,100,91,
               93,96,95,94)
Metodo=c(2,2,2,2,2,2,2,2,2,3,3,3,3,3,3,3)
```

```
#Calcular la mediana
```

```
mediana=median(Rendimiento)
mediana
```

```
[1] 91
```

```
table(Metodo[which(Rendimiento>mediana)])
```

```
3
6
```

```
table(Metodo[which(Rendimiento<=mediana)])
```

```
2 3
10 1
```

```
Observados=matrix(c(0,6,10,1), nrow = 2, ncol = 2, byrow = T)
rownames(Observados)=c('Mayor_med','MenorIgual_med')
colnames(Observados)=c('Metodo 2','Metodo 3')
Observados
```

```
      Metodo 2 Metodo 3
Mayor_med      0      6
MenorIgual_med 10      1
```

```
#Prueba
T2 <- fisher.test(Observados)
T2
```

Fisher's Exact Test for Count Data

```
data: Observados
p-value = 0.0005656
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.0000000 0.2669263
sample estimates:
odds ratio
      0
```

Los métodos 2 y 3 difieren con respecto al rendimiento medio (mediana) por m^2 .

Capítulo 18

Ejercicios

1. Queremos probar si la selección de cierto deporte es independiente del género. Para ellos se les preguntó a 100 hombres y 100 mujeres que deporte entre arquería, boxeo y ciclismo preferían practicar y en la siguiente tabla se resume las respuestas que dieron:

Género	Arquería	Boxeo	Ciclismo	Total
Mujer	35	15	50	100
Varon	10	30	60	100
Total	45	45	110	200

2. En un estudio, llevado a cabo por el INE, se tomó una muestra aleatoria de ciudadanos registrados en el Padrón Electoral. Se obtuvo información sobre el partido por el que votaron para Presidente y si tenían o no estudios universitarios.

X_2	X_1						
SI	PRI	PRI	PRI	PRI	PRI	PRI	PRI
NO	PRI	PRI	PRI	PRI	PRI	PRI	PRI
SI	PRI	PAN	PAN	PAN	PAN	PAN	PAN
NO	PRI	PAN	PAN	PAN	PAN	PAN	PAN
SI	PAN	PAN	PAN	PAN	PAN	PAN	PAN
NO	PAN	PAN	PAN	PAN	PAN	PAN	PAN

Considere X_1 = Partido por el cual votó, X_2 = tiene o no estudios universitarios.

Se desea probar si existe asociación entre el partido por el cual votó y si tienen o no estudios universitarios. Plantee, en el contexto del problema, las hipótesis, estadística de prueba y obtenga su conclusión con un nivel de significancia del 5 %

3. En una encuesta telefónica se preguntó a los participantes hasta que grado estaban de acuerdo con la proposición: “se debe prohibir fumar en lugares públicos”. Con base en los datos recabados se desea saber si existen diferencias significativas en el grado en el que están de acuerdo hombres y mujeres con respecto a prohibir fumar en lugares públicos.

Sexo	Muy de acuerdo	De acuerdo	Neutral	En desacuerdo	En total desacuerdo
Mujer	41	16	28	27	31
Varon	22	40	14	39	41

4. ¿Qué tan bueno es el servicio que dan las líneas aéreas a sus clientes? En un estudio las evaluaciones dadas por los clientes fueron las siguientes: 3 excelente, 28 bueno, 45 aceptable y 24 malo (BusinessWeek, 11 de septiembre de 2000). En otro estudio sobre las empresas de servicio telefónico, en una muestra de 400 adultos las evaluaciones fueron las siguientes: 24 excelente, 124 bueno, 172 aceptable y 80 malo. ¿La distribución de las evaluaciones a las empresas telefónicas difiere de la distribución de las evaluaciones a las líneas aéreas? Emplee $\alpha = 5\%$ ¿Cuál es su conclusión?

5. Estamos interesados en estudiar la fiabilidad de cierto componente informático con relación al distribuidor que nos lo suministra. Para realizar esto, tomamos una muestra de 100 componentes de cada uno de los tres distribuidores que nos sirven el producto comprobando el número de defectuosos en cada lote. La siguiente tabla muestra el número de defectuosos para cada uno de los distribuidores.

	Componentes Defectuosos	Componentes Correctos
Distribuidor 1	16	94
Distribuidor 2	24	76
Distribuidor 3	9	81

Plantea la hipótesis correspondiente y concluye de acuerdo al contexto.

6. Durante las primeras 13 semanas, se registraron las proporciones siguientes de televidentes los sábados de 8 a 9 de la noche: ABC 29 %, CBS 28 %, NBC 25 % e independientes 18 %. Dos semanas después en una muestra de 300 hogares se obtuvieron las audiencias siguientes en sábado por la noche: ABC 95 hogares, CBS 70 hogares, NBC 89 hogares e independientes 46 hogares. Use $\alpha = 5\%$ para determinar si han variado las proporciones en la audiencia de televidentes.

7. En el campo de entrenamiento 100 reclutas son asignados aleatoriamente en cuatro regimientos con 4 diferentes sargentos. Al final del entrenamiento sólo quedarán 84 reclutas y sus tiempos en el ejercicio de obstáculos fue medido para todos ellos. Los resultados fueron, para el sargento Adams 11 de sus 20 reclutas tuvieron tiempos por arriba de la mediana, para el sargento Baker 8 de sus 22 reclutas tuvieron tiempos por arriba de la mediana, el sargento Callahan 8 de sus 20 reclutas tuvieron tiempos por encima de la mediana y del sargento Davis 15 de 22 tuvieron tiempos arriba de la mediana. Se puede decir con un nivel de significancia del 5 % que existe una diferencia significativa en los tiempos en los ejercicios de obstáculos entre cada uno de los regimientos?

8. Se subastaron varios contratos de explotación de petróleo al mejor postor. Para cada contrato se recibieron una o más ofertas selladas. Pruebe la hipótesis de que los contratos que eventualmente se convirtieron en productores de petróleo tuvieron la misma mediana de ofertas que los contratos que no produjeron petróleo. A continuación se muestra una muestra aleatoria de cada tipo de contrato.

	Número de ofertas en cada contrato de arrendamiento
Productores	6, 3, 1, 14, 8, 9, 12, 1, 3, 2, 1, 7
No productores	6, 2, 1, 1, 3, 1, 2, 4, 8, 1, 2

Parte V

Bondad de Ajuste

Introducción

Las pruebas de bondad de ajuste se utilizan para decidir si un conjunto de datos (una muestra aleatoria) se ajusta a una función de distribución dada.

Estas pruebas son importantes por que existen métodos estadísticos que se basan en algún supuesto de la distribución de los datos y si tal supuesto no se cumple, el método no es válido. Por ejemplo, el modelo de regresión lineal supone que los errores tienen distribución normal y de no validarse este supuesto entonces la inferencia hecha sobre los parámetros del modelo de regresión carece de sustento estadístico.

La hipótesis nula en este tipo de pruebas es que los datos tienen la distribución requerida. Algunas veces la hipótesis nula específica totalmente a la distribución (es decir también especifica el valor de los parámetros), otras veces sólo especifica de que distribución se trata (sin importar los parámetros).

Las pruebas de bondad de ajuste desarrolladas dependen de la variable aleatoria que se esta modelando. Para distribuciones discretas, se comparan las frecuencias esperadas con las observadas. (Prueba Ji-Cuadrada). Para distribuciones continuas, se compara la función de distribución empírica con la de distribución requerida. (Prueba Kolmogorov-Smirnov, Lilliefors, Anderson-Darling)

Capítulo 19

Prueba de la Ji-cuadrada

19.1. Datos

X_1, \dots, X_n m.a. de tamaño "n" que proviene de una distribución $F(x)$ desconocida.

Cada una de las variables se pueden acomodar en alguna clase "k" (o categoría)

19.2. Hipótesis

\mathbf{H}_0 : Los datos siguen una distribución $F_0(x)$.

vs

\mathbf{H}_a : Los datos no siguen una distribución $F_0(x)$.

- Donde $F_0(x)$ es la distribución que se propone.

Es decir:

\mathbf{H}_0 : $\mathbf{P}[X \text{ pertenezca a la categoría } j] = \mathbf{P}_j$, para toda $j = 1, \dots, k$.

vs

\mathbf{H}_a : $\mathbf{P}[X \text{ pertenezca a la categoría } j] \neq \mathbf{P}_j$, para alguna $j = 1, \dots, k$.

Procedimiento

Vamos a buscar las probabilidades de ocurrencia en cada categoría.

Calcular los valores esperados e_j ; Donde $e_j = n \times P_j$

19.3. Estadístico de Prueba

$$Q = \sum_{j=1}^k \frac{(f_j - e_j)^2}{e_j}, \quad \text{Donde } Q \sim \chi^2_{(k-1)}$$

Observaciones

- Q es estable cuando el número de observaciones en cada categoría debe ser mayor a 5.
- En caso de que alguna categoría tenga menos de 5 observaciones colapsamos las categorías.
- Si desconocemos los parámetros los estimamos con la muestra (*EMV*, *Momentos*), con esto se van perdiendo grados de libertad.
- $Q \sim \chi^2_{(k-1-r)}$ Donde k =Número de clases o intervalos, r =Número de parámetros estimados.

Regla de decisión

Rechazamos H_0 si $Q > q_{teórica}$

- Cuando $Q = \sum_{j=1}^k \frac{(f_j - e_j)^2}{e_j} > \chi^2_{(k-1)}(1 - \alpha)$

19.4. Ejemplo

Un gobierno local tiene registros del número de niños y el número de hogares en el área. Se sabe que el número promedio de niños por hogar es 1.40. Se sugiere que el número de niños por hogar se pueda modelar por una distribución Poisson con parámetro 1.40. Para probar esta hipótesis se toma una muestra de 1000 hogares; los resultados se muestran en la siguiente tabla:

Número de niños	0	1	2	3	4	5+
Número de hogares	273	361	263	78	21	4

NOTA: Como tenemos una categoría con observaciones menores a 5 colapsamos la categoría.

Número de niños	0	1	2	3	4+
Número de hogares	273	361	263	78	25

Paso 1 Prueba a utilizar **Prueba Ji-cuadrada**

Paso 2 Planteamiento de hipótesis:

H_0 : El número de niños por hogar sigue una distribución Poisson (1.40)

vs

H_a : El número de niños por hogar no sigue una distribución Poisson (1.40)

Necesitamos buscar la Distribución de una variable aleatoria Poisson, con parámetro $\lambda = 1.40$

Nota En R-Studio tenemos `dpois(c(0,1,2,3,4),1.40)` para la primera categoría.

Nos ayudaremos de la siguiente tablita:

Número de niños	Número de hogares	P_i	$e_i = n \times P_i$
0	273	0.2465	247
1	361	0.3452	345
2	263	0.2416	242
3	78	0.1127	113
4	25	0.0394	39
$n = 1000$		$.9854 \approx 1$	$9854 \approx 1000$

Paso 3 Estadístico de prueba

$$Q = \sum_{j=0}^k \frac{(f_j - e_j)^2}{e_j} = \sum_{j=0}^4 \frac{(f_j - e_j)^2}{e_j} =$$

$$= \frac{(273 - 247)^2}{247} + \frac{(361 - 345)^2}{345} + \frac{(263 - 242)^2}{242} + \frac{(78 - 113)^2}{113} + \frac{(25 - 39)^2}{39}$$

$$T = 21.4605$$

Paso 4 Regla de decisión

Rechazamos H_0 si $Q > q_{teórica}$

■ Cuando

$$Q = \sum_{j=1}^k \frac{(f_j - e_j)^2}{e_j} > \chi_{(k-1)}^2(1 - \alpha)$$

Ocuparemos $\alpha = 0.05$

$$\chi_{(k-1)}^2(1 - \alpha) = \chi_{(5-1)}^2(1 - 0.05) = \chi_{(4)}^2(.95)$$

$$Q = 21.4605 > 9.48 = \chi_{(4)}^2(.95)$$

\therefore Rechazamos H_0 .

Paso 5 Conclusión

\therefore Existe evidencia estadística suficiente para decir que el número de niños por hogar no sigue una Distribución Poisson (1.40).

19.5. Ejemplo en R-Studio

Ahora haremos la réplica en R.

```
#Número de hogares
observados =c(273,361,263,78,21,4)
#Matriz de frecuencias observadas
tabla=matrix(observados,nrow=1)
#Número de niños
num_ninos=c("0","1","2","3","4","5+")
#Asigna nombres a la tabla de cada categoría
dimnames(tabla)=list(NULL,num_ninos)
tabla
```

```
      0   1   2   3   4 5+
[1,] 273 361 263 78 21  4
```

```
#Una de las categorías es menor a 5, así que colapsamos la categoría 5+
```

```
#Número de hogares
observados =c(273,361,263,78,25)
#Matriz de frecuencias observadas
tabla=matrix(observados,nrow=1)
#Número de niños
num_ninos=c("0","1","2","3","4+")
#Asigna nombres a la tabla de cada categoría
dimnames(tabla)=list(NULL,num_ninos)
tabla
```

```
      0    1    2    3  4+
[1,] 273 361 263  78  25
```

```
#Total
N=sum(observados)

#Calculamos las probabilidades
p=dpois(0:4,1.40)
p
```

```
[1] 0.24659696 0.34523575 0.24166502 0.11277701 0.03947195
```

```
#Calculamos los valores esperados
esperados=N*p
round(esperados,0) #Hacemos esto por que estamos hablando de niños
```

```
[1] 247 345 242 113  39
```

```
residuos=round((observados-esperados)^2/esperados,4)
ji_cal=sum(residuos)
ji_cal
```

```
[1] 21.4605
```

```
#En un principio teníamos 6 categorías pero colapsamos una, entonces
#tenemos 5 categorías menos una que corresponden a nuestros grados de libertad.
```

```
ji_teo=qchisq(0.95,df=4)
ji_teo
```

```
[1] 9.487729
```

```
#Como Ji-Cal > Ji_teo rechazamos H0.
```

```
p_value=pchisq(ji_cal,df=4,lower.tail=FALSE)
p_value
```

```
[1] 0.0002565777
```

Observamos que la estadística de prueba tiene un valor de 21.46 y su correspondiente p-value es menor a 0.05, por lo tanto con $\alpha = 5\%$ rechazaremos H_0 y concluimos existe evidencia suficiente para suponer que el número de niños por hogar no sigue una distribución Poisson(1.40).

Ahora podemos utilizar la prueba de R:

```
chisq.test(c(273,361,263,78,25), p = c(0.2466,0.3452,0.2417,  
                                         0.1128,0.03947), rescale.p=TRUE)
```

Chi-squared test for given probabilities

data: c(273, 361, 263, 78, 25)

X-squared = 20.96, df = 4, p-value = 0.0003226

Capítulo 20

Prueba Kolmogorov

Comenzaremos con una prueba de bondad de ajuste que fue presentada por Kolmogorov (1933). Esta prueba es quizás la más útil, en parte porque nos proporciona una alternativa, diseñada para datos ordinales, a la prueba χ^2 para bondad de ajuste, que fue diseñada para datos de tipo nominal, y en parte porque la estadística de prueba de Kolmogorov nos permite formar una “banda de confianza” para la función de distribución desconocida.

Una prueba de bondad de ajuste generalmente involucra una muestra aleatoria de alguna distribución desconocida para probar la hipótesis nula de que la función de distribución desconocida es de hecho una función conocida y especificada. Esto es, la hipótesis nula especifica alguna función de distribución $F^*(x)$, tal vez gráficamente como en la figura 1, o tal vez como una función matemática que puede ser graficada. Luego se toma una muestra aleatoria, X_1, X_2, \dots, X_n de alguna población y se compara con $F^*(x)$ de alguna manera para ver si es razonable decir que, $F^*(x)$ es la verdadera función de distribución de la muestra aleatoria.

Una forma lógica de comparar la muestra aleatoria con $F^*(x)$ es mediante la función de distribución empírica $S(x)$, definida como la fracción de X_i s que son menores o iguales a x , para cada x , $-\infty < x < +\infty$. La función de distribución empírica $S(x)$ es útil como estimador de $F(x)$, la función de distribución desconocida de la X_i s. Entonces podemos comparar la función de distribución empírica $S(x)$ con la función de distribución hipotética $F^*(x)$ para ver si hay un buen ajuste. Si no hay un buen ajuste, entonces podemos rechazar la hipótesis nula y concluir que la función de distribución verdadera pero desconocida, $F(x)$, en realidad no está dada por la función $F^*(x)$ en la hipótesis nula.

Figura 1

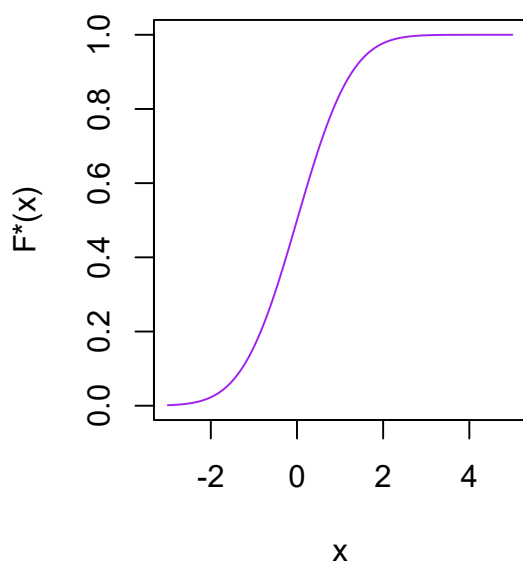
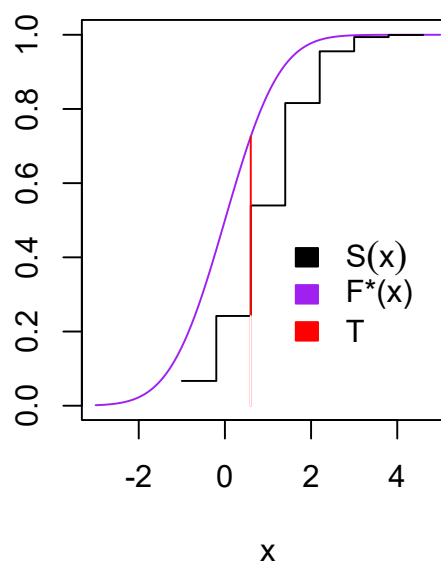


Figura 2



Pero, ¿qué tipo de estadística de prueba podemos usar como medida de la diferencia entre $S(x)$ y $F^*(x)$? Una de las medidas más simples imaginables es la mayor distancia entre los dos gráficos $S(x)$ y $F^*(x)$, medidos en dirección vertical. Ésta es la estadística sugerida por Kolmogorov (1993). Es decir, si la **figura 1** proporciona $F^*(x)$ y se extrae una muestra aleatoria de tamaño 5 de la población, la función de distribución empírica $S(x)$ se puede dibujar en el mismo gráfico junto con $F^*(x)$, como se muestra en la **figura 2**. Si $F^*(x)$ y $S(x)$ son como se indica, la distancia vertical máxima entre los dos gráficos se produce justo antes del tercer paso de $S(x)$. Esta distancia es de aproximadamente 0.5 en la **figura 2**; por lo tanto, el estadístico Kolmogorov T es igual a 0.5 en este caso. Los valores grandes de T según lo determinado por la correspondiente tabla (Tabla de Cuantiles Kolmogorov) conducen a rechazar $F^*(x)$ como una aproximación razonable a la función de distribución verdadera desconocida $F(x)$.

La prueba de Kolmogorov puede preferirse a la prueba de Ji-Cuadrada para bondad del ajuste si el tamaño de la muestra es pequeño; la prueba de Kolmogorov es exacta incluso para muestras pequeñas, mientras que la prueba de Ji-cuadrada supone que el número de observaciones es lo suficientemente grande como para que la distribución de χ^2 proporcione una buena aproximación como la distribución del estadístico de prueba. Existe controversia sobre ¿qué prueba es la más “poderosa”? , pero la sensación general parece ser que la prueba de Kolmogorov es probablemente más poderosa que la prueba de Ji-cuadrada en la mayoría de las situaciones que involucran datos ordinales.

20.1. Datos

Los datos consisten en una muestra aleatoria X_1, X_2, \dots, X_n de tamaño n asociada con alguna función de distribución desconocida, denotada por $F(x)$.

20.2. Supuestos

- 1) La muestra es una muestra aleatoria.

20.3. Estadístico de Prueba

Sea $S(x)$ la función de distribución empírica basada en la muestra aleatoria X_1, X_2, \dots, X_n . El estadístico de prueba es definido diferente para los 3 casos para las hipótesis correspondientes. Sea $F^*(x)$ una función de distribución hipotética completamente especificada.

Caso A (Prueba de 2 colas)

Sea el estadístico de prueba T la mayor distancia vertical entre $S(x)$ y $F^*(x)$ (denotado por sup o el supremo).

$$T = \sup_x |F^*(x) - S(x)|$$

Esto se lee “ T es igual al supremo para todas las x , del valor absoluto de la diferencia de $F^*(x) - S(x)$ ”

Caso B (Prueba de 1 cola)

Denotamos el estadístico de prueba T^+ la mayor distancia vertical alcanzada por $F^*(x)$ sobre $S(x)$.

$$T^+ = \sup_x [F^*(x) - S(x)]$$

Que es similar a T , a excepción que solo vamos a considerar la mayor diferencia alcanzada por $F^*(x)$ sobre la función $S(x)$.

Caso C (Prueba de 1 cola)

Denotamos el estadístico de prueba T^- la mayor distancia vertical alcanzada por $S(x)$ sobre $F^*(x)$.

$$T^- = \sup_x [S(x) - F^*(x)]$$

Que es similar a T , a excepción que solo vamos a considerar la mayor diferencia alcanzada por $S(x)$ sobre la función $F^*(x)$.

20.4. Hipótesis

Caso A (Prueba de 2 colas)

$$\mathbf{H}_0 : F(x) = F^*(x) \quad \forall x \text{ de } -\infty \text{ a } +\infty$$

vs

$$\mathbf{H}_a : F(x) \neq F^*(x) \quad \text{para al menos un valor de } x.$$

Regla de decisión

Rechazo H_0 a un nivel de significancia α si $T > W$. Donde W es el cuantil $(1 - \alpha)$, obtenido en la tabla correspondiente a nuestra prueba, para la prueba de 2 colas.

Caso B (Prueba de 1 cola)

$$\mathbf{H}_0 : F(x) \geq F^*(x) \quad \forall x \text{ de } -\infty \text{ a } +\infty$$

vs

$$\mathbf{H}_a : F(x) < F^*(x) \quad \text{para al menos un valor de } x.$$

Regla de decisión

Rechazo H_0 a un nivel de significancia α si $T^+ > W$. Donde W es el cuantil $(1 - \alpha)$, obtenido en la tabla correspondiente a nuestra prueba, para la prueba de 1 cola.

Caso C (Prueba de 1 cola)

$$\mathbf{H}_0 : F(x) \leq F^*(x) \quad \forall x \text{ de } -\infty \text{ a } +\infty$$

vs

$$\mathbf{H}_a : F(x) > F^*(x) \quad \text{para al menos un valor de } x.$$

Regla de decisión

Rechazo H_0 a un nivel de significancia α si $T^- > W$. Donde W es el cuantil $(1 - \alpha)$, obtenido en la tabla correspondiente a nuestra prueba, para la prueba de 1 cola.

Vamos a aplicar este conocimiento en un ejemplo:

20.5. Ejemplo

Una muestra aleatoria de tamaño 10, es obtenida:

$$X_1 = 0.621, \quad X_2 = 0.503, \quad X_3 = 0.203, \quad X_4 = 0.477, \quad X_5 = 0.710, \quad X_6 = 0.581, \\ X_7 = 0.329, \quad X_8 = 0.480, \quad X_9 = 0.554, \quad X_{10} = 0.382$$

La hipótesis nula es que la función de distribución es una función de distribución uniforme. La expresión matemática de función de distribución hipotética es:

$$\mathbf{F}^*(\mathbf{x}) = \begin{cases} 0 & \text{si } x < 0 \\ x & \text{si } 0 \leq x < 1 \\ 1 & \text{si } 1 \leq x \end{cases}$$

Paso 1 Prueba a utilizar **Prueba de Bondad de Ajuste Kolmogorov**

Paso 2 Planteamiento de Hipótesis

$$\mathbf{H}_0 : F(x) = F^*(x) \quad \forall x \text{ de } -\infty \text{ a } +\infty$$

vs

$$\mathbf{H}_a : F(x) \neq F^*(x) \quad \text{para al menos un valor de } x.$$

- Donde $\mathbf{F}(\mathbf{x})$ es la función de distribución desconocida común a las X_i s y $F^*(x)$ se da por la expresión matemática.

$$\mathbf{F}^*(\mathbf{x}) = \begin{cases} 0 & \text{si } x < 0 \\ x & \text{si } 0 \leq x < 1 \\ 1 & \text{si } 1 \leq x \end{cases}$$

Paso 3 Estadístico de Prueba

Calculamos el Estadístico de Prueba:

$$T = \sup_x |F^*(x) - S(x)|$$

$$T = 0.290$$

Paso 4 Procedimiento completo para el cálculo del Estadístico de Prueba: La siguiente tabla representa los cálculos para encontrar nuestro Estadístico de Prueba T :

i

$X(i) = x$

$F^*(x)$

S_n

$F^*(x) - S_n(x)$

$|F^*(x) - S_n(x)|$

1

0.203

0.203

0.1

0.103

0.103

2

0.329

0.329

0.2

0.129

0.129

3

0.382

0.382

0.3

0.082

0.082

4

0.477

0.477

0.4
0.077
0.077
5
0.480
0.480
0.5
-0.020
0.020
6
0.503
0.503
0.6
-0.097
0.097
7
0.554
0.554
0.7
-0.146
0.146
8
0.581
0.581
0.8
-0.219
0.219
9
0.621
0.621
0.9
-0.279
0.279
10
0.710
0.710
1.0
-0.290
0.290

Paso 5 Regla de Decisión

El cuantil W que acumula $1 - \alpha$ de probabilidad, usando $\alpha=0.05$ es $W = 0.409$, encontrado en las tablas correspondientes.

Tenemos que $T < W$, entonces no rechazamos la hipótesis nula.

Paso 6 Conclusión

Entonces podemos concluir que los datos siguen una distribución uniforme.

20.6. Ejemplo en R-Studio

Ahora haremos la réplica en R.

```
i = c(1:10) #Representa el numero de nuestra muestra
x = c(0.203,0.329,0.382,0.477,0.480,0.503,0.554,0.581,0.621, 0.710) #Los datos de la muestra
X_i = sort(x) #ordena nuestros datos
F_ = c(0.203,0.329,0.382,0.477,0.480,0.503,0.554,0.581,0.621, 0.710)
Sn = c(1/10,2/10,3/10,4/10,5/10,6/10,7/10,8/10,9/10,1)
Tabla = cbind(i,X_i=X_i,F_,Sn,"|F_-Sn|"=abs(F_-Sn))
Tabla
```

	i	X_i	F_	Sn	F_-Sn
[1,]	1	0.203	0.203	0.1	0.103
[2,]	2	0.329	0.329	0.2	0.129
[3,]	3	0.382	0.382	0.3	0.082
[4,]	4	0.477	0.477	0.4	0.077
[5,]	5	0.480	0.480	0.5	0.020
[6,]	6	0.503	0.503	0.6	0.097
[7,]	7	0.554	0.554	0.7	0.146
[8,]	8	0.581	0.581	0.8	0.219
[9,]	9	0.621	0.621	0.9	0.279
[10,]	10	0.710	0.710	1.0	0.290

```
EstdPrueba = max(Tabla [,5])
EstdPrueba
```

```
[1] 0.29
```

Observamos que la estadística de prueba tiene un valor de 0.29. El cuantil W que acumula $1 - \alpha$ de probabilidad, usando $\alpha = 0.05$ es $W = 0.409$, encontrado en las tablas correspondientes. Por lo tanto tenemos que $T_1 < W$, entonces no rechazamos la hipótesis nula y su correspondiente p-value es mucho mayor a 0.05, por lo tanto con $\alpha = 5\%$ no rechazaremos H_0 y concluimos no existe evidencia suficiente para suponer que la distribución de la muestra no es uniforme en el intervalo $(0,1)$.

Ahora haremos la prueba usando la función “**ks.test**” de R.

```
#Prueba
ks.test(X_i,Sn)
```

Two-sample Kolmogorov-Smirnov test

```
data: X_i and Sn
D = 0.3, p-value = 0.7869
alternative hypothesis: two-sided
```

20.7. Otro ejemplo en R

Se mencionó que uno de los usos de estas pruebas es para validar el supuesto de normalidad en los modelos de regresión lineal. Veremos ahora un ejemplo en donde los datos a los que se aplica la prueba de bondad de ajuste son los residuales de una regresión lineal simple.

La base de datos “**Loblolly**” en R, contiene información sobre tres características de árboles de pino originarios del sudeste de Estados Unidos. Al ajustar un modelo de regresión lineal simple entre $X = \text{"edad"}$ y $Y = \text{"altura"}$, deseamos probar con un nivel de significancia del 1 % que los residuales estandarizados se distribuyen normal estándar.

```
m1=lm(Loblolly$height~Loblolly$age)
x=rstandard(m1)

ks.test(x,"pnorm",0,1)
```

One-sample Kolmogorov-Smirnov test

```
data: x
D = 0.10804, p-value = 0.2613
alternative hypothesis: two-sided
```

Observamos que la estadística de prueba tiene un valor de 0.10804 y su correspondiente $p - value$ es 26.13 %, por lo tanto con $\alpha = 5\%$ no rechazaremos H_0 y concluimos no existe evidencia suficiente para suponer que la distribución de la muestra no es normal(0,1).

Capítulo 21

Prueba Kolmogorov-Smirnov

En ejercicios prácticos es muy difícil conocer la distribución de una muestra aleatoria, generalmente sólo se tiene la información; ésta hay que procesarla para averiguar si sigue una determinada distribución probabilística, en un primer intento se ajustó mediante la prueba de la Ji-cuadrada, sin embargo, al ser una de las pruebas más sencillas su “**potencia**” al estimar una determinada distribución es baja, es por ello, que se idearon otros métodos y uno de ellos es la Prueba de Kolmogorov-Smirnov.

La prueba de Kolmogorov presenta la ventaja de que los datos no deben ser categorizadas para poder realizar estimaciones en su distribución. Al igual que en la prueba de la Ji-Cuadrada, Kolmogorov-Smirnov trabaja con una distribución $F^*(x)$ totalmente especificada, es decir, se debe de tener sospecha de que la muestra aleatoria siga una determinada distribución. De esta manera el objeto de estudio es una muestra X_1, \dots, X_n de variables aleatorias idénticamente distribuidas, las cuales siguen una distribución desconocida $F(X)$ y se tiene la sospecha de que la muestra sigue una distribución conocida $F^*(x)$.

Para probar la suposición de la distribución $F^*(x)$ se realiza la siguiente contraste:

21.1. Hipótesis

21.1.1. Caso A (Prueba de 2 colas)

Solo será este caso

$$H_0 : F(x) = F^*(x) \quad \forall x \text{ de } -\infty \text{ a } +\infty$$

vs

$$H_a : F(x) \neq F^*(x) \quad \text{para al menos un valor de } x.$$

Regla de decisión

Rechazo H_0 a un nivel de significancia α si $D_n > W$. Donde W es el cuantil $(1 - \alpha)$, obtenido en la tabla correspondiente a nuestra prueba O para la prueba de 2 colas en las tablas de Kolmogorov.

Donde $F^*(x)$ es una distribución completamente conocida, es decir además de conocer a la familia que pertenece también se conocen sus parámetros.

Lo que se busca es poder medir las distancia entre $F(x)$, la distribución desconocida, con los datos que siguen la función de distribución propuesta y completamente conocida $F^*(x)$. Sin embargo, $F(x)$ al ser desconocida se recurre a la construcción de una distribución empírica la cual se define como:

$$S_n(x) = \frac{\sum_{i=1}^n \text{número de valores muestrales} \leq x}{n}$$

Es decir, la función empírica mide el número de elementos menores o iguales a la observación X , puede observarse que en el caso continuo, al no haber “empates” la función empírica puede ser vista como:

$$S_n(x) = \frac{i}{n} \quad i = 1, \dots, n$$

Al tener una distribución desconocida $F(x)$, la función empírica $S_n(x)$ puede ser usada como un estimador insesgado de $F(x)$ pues:

$$\mathbb{E}(S_n(x)) = F(x)$$

La función empírica es de gran importancia ya que gracias al teorema de **Glivenko-Cantelli** se sabe que cuando el tamaño de la muestra tiende a infinito cualquier distribución empírica se aproxima a la distribución real de los datos, la cual, es una distribución completamente especificada. El teorema de **Glivenko-Cantelli**, menciona que al calcular las diferencias de la distribución real y la empírica éstas son cero en cada observación dada, el teorema que se enuncia como:

Sea X_1, \dots, X_n una muestra aleatoria de distribución $F(x)$ desconocida y sea $S_n(x)$ la función empírica entonces:

$$\sup_x |S_n(x) - F(x)| \longrightarrow 0$$

Es decir, conforme mayor sea el tamaño de la muestra, $S_n(x)$ reproduce la verdadera distribución. De esta manera se establece el estadístico de Prueba D_n , el cual no depende de ningún parámetro desconocido, ya que engloba a la distribución empírica y a la distribución propuesta:

$$D_n = \sup_x |S_n(x) - F^*(x)| = \max \{ \max \{ S_n(X_{i-1}) - F^*(X_i) \}, \max \{ S_n(X_i) - F^*(X_i) \} \} \quad \forall i$$

y

$$D_n = \sup_x |S_n(x) - F^*(x)| = \max \{ D^+, D^- \}$$

Donde:

$$D^+ = \max \{ S_n(X_i) - F^*(X_i) \}$$

$$D^- = \max \{ S_n(X_{i-1}) - F^*(X_i) \}$$

Finalmente, se observa que si H_0 es cierta si $D_n \longrightarrow 0$ ya que las diferencias entre la función empírica y la propuestas son mínimas, lo que cumple con el **Teorema de Glivenko-Cantelli**; por lo que hay evidencia para rechazar H_0 cuando $D_n > W$ Donde W es el cuantil que acumula el $1 - \alpha$ de probabilidad de la distribución asociada a D_n la cual puede obtenerse de la tablas correspondientes la cual muestra los cuantiles de la distribución Kolmogorov-Smirnov o la Tabla Kolmogorov para 2 colas.

21.2. Ejemplo

Dada la siguiente muestra

0.6379 1.5299 0.35005 2.0505 2.1906 0.3459 2.3214 0.3128
0.65482.4373 1.803 2.3674 1.2716 0.2566 0.2513

Se desea hacer el siguiente contraste:

$$\mathbf{H}_0 : \text{Los datos} \sim \text{LogN}(0, 1)$$

vs

$$\mathbf{H}_a : \text{Los datos} \approx \text{LogN}(0, 1)$$

Realice la prueba de Kolmogorov-Smirnov al 5 % de significancia.

Paso 1 Prueba a utilizar **Prueba de Bondad de Ajuste Kolmogorov-Smirnov**

Paso 2 Planteamiento de Hipótesis

$$\mathbf{H}_0 : \text{Los datos} \sim \text{LogN}(0, 1)$$

vs

$$\mathbf{H}_a : \text{Los datos} \approx \text{LogN}(0, 1)$$

Paso 3 Estadístico de Prueba:

$$D_n = \sup_x | S_n(x) - F^*(x) | = \max \{ D^+, D^- \}$$

Paso 4 Procedimiento completo para el cálculo del Estadístico de Prueba:

- 1) Se procede a ordenar nuestras observaciones de menor a mayor.
- 2) Se calcula la función empírica, como no tenemos ningún valor repetido:

$$S_n = \frac{i}{n} = \frac{1}{15}, \frac{2}{15}, \dots, 1.$$

- 3) Se calcula la función empírica menos un valor, es decir,}

$$S_n = \frac{i-1}{n} = \frac{0}{15}, \frac{1}{15}, \dots, \frac{14}{15}.$$

- 4) Se calcula la distribución conocida, es decir, $F^*(x) \sim \text{LogN}(0, 1)$
- 5) Se calcula D^+ que es el resultado de la resta de la distribución conocida menos la distribución empírica, es decir:

$$D^+ = \max \{ S_n(X_i) - F^*(X_i) \}$$

- 6) Se calcula D^- que es el resultado de la resta de la distribución empírica menos uno menos la distribución conocida, es decir:

$$D^- = \max \{ S_n(X_{i-1}) - F^*(X_i) \}$$

- 7) Finalmente realizada la tabla, se calcula el máximo de las columnas D^+ y D^- de ésta manera, se tiene la siguiente tabla:

 i X X_i $S_n(X_i)$ $S_n(X_{i-1})$ $F^*(X_i)$ $D^+ = S_n(X_i) - F^*(X_i)$ $D^- = S_n(X_{i-1}) - F^*(X_i)$

1

0.63790

0.25130

0.0666667

0.0000000

0.0836

-0.0169

-0.0836

2

1.52990

0.25660

0.1333333

0.0666667

0.0869

0.0464

-0.0202

3

0.35005

0.31280

0.2000000

0.1333333

0.1226

0.0774

0.0107

4

2.05050

0.34590

0.2666667

0.2000000

0.1442
0.1224
0.0558
5
2.19060
0.35005
0.3333333
0.2666667
0.1472
0.1861
0.1194
6
0.34590
0.63790
0.4000000
0.3333333
0.3265
0.0735
0.0068
7
2.32140
0.65480
0.4666667
0.4000000
0.3360
0.1306
0.0640
8
0.31280
1.27160
0.5333333
0.4666667
0.5949
-0.0615
-0.1282
9
0.65480
1.52990
0.6000000

0.5333333

0.6647

-0.0647

-0.1313

10

2.43730

1.80300

0.6666667

0.6000000

0.7222

-0.0555

-0.1222

11

1.80300

2.05050

0.7333333

0.6666667

0.7636

-0.0302

-0.0969

12

2.36740

2.19060

0.8000000

0.7333333

0.7835

0.0165

-0.0501

13

1.27160

2.32140

0.8666667

0.8000000

0.8002

0.0664

-0.0002

14

0.25660

2.36740

0.9333333

0.8666667

0.8056

0.1277

0.0610

15

0.25130

2.43730

1.0000000

0.9333333

0.8135

0.1865

0.1198

8) Entonces

$$D^+ = \max \{ S_n(X_i) - F^*(X_i) \} = 0.1865 \quad y \quad D^- = \max \{ S_n(X_{i-1}) - F^*(X_i) \} = 0.1198$$

Por lo tanto:

$$D_n = \sup_x | S_n(x) - F^*(x) | = \max \{ D^+, D^- \} = \max \{ 0.1865, 0.1198 \} = 0.1865$$

Paso 5 Regla de Decisión

Este último resultado se compara con la tabla de valores críticos de la Tabla Kolmogorov-Smirnov, para un nivel de significancia $\alpha = 0.05$ $W_{0.05}=0.338$, de esta manera se tiene que $0.338 = W_{0.05} > D_n = 0.1865$, como el estadístico $W_{0.05}$ es mayor a comparación de $D_n=0.1865$. No Rechazamos H_0 .

Paso 6 Conclusión

Se acepta la prueba de lognormalidad con media 1 y varianza 0, con un nivel de significancia $\alpha = 0.05$. Es decir, Los datos se distribuyen $LogN(0, 1)$.

```
i= 1:15
X= c(0.6379,1.5299,0.35005,2.0505,2.1906,0.3459,2.3214,0.3128,0.6548,2.4373,
      1.803,2.3674,1.2716,0.2566,0.2513)
X_i=c(0.25130,0.25660,0.31280, 0.34590,0.35005,0.63790, 0.65480, 1.27160,
      1.52990,1.80300, 2.05050, 2.19060, 2.32140, 2.36740,2.43730)

Sn=c(1/15,2/15,3/15,4/15,5/15,6/15,7/15,8/15,9/15,10/15,11/15,12/15,13/15,14/15,1)
Sn_1=c(0/15,1/15,2/15,3/15,4/15,5/15,6/15,7/15,8/15,9/15,10/15,11/15,12/15,13/15,14/15)
F_=c(0.0836,0.0869,0.1226,0.1442,0.1472,0.3265,0.3360,0.5949,
      0.6647,0.7222,0.7636,0.7835,0.8002,0.8056,0.8135)
D_mas=c(-0.0169,0.0464,0.0774,0.1224,0.1861,0.0735,0.1306,-0.0615,-0.0647,-0.0555,
      -0.0302,0.0165,0.0664,0.1277,0.1865)
D_menos=c(-0.0836,-0.0202,0.0107,0.0558,0.1194,0.0068,0.0640,-0.1282,
      -0.1313,-0.1222,-0.0969,-0.0501,-0.0002,0.0610,0.1198)

Tabla=cbind(i,X_i,Sn,Sn_1,F_,D_mas,D_menos)
Tabla
```

	i	X_i	Sn	Sn_1	F_	D_mas	D_menos
[1,]	1	0.25130	0.06666667	0.00000000	0.0836	-0.0169	-0.0836
[2,]	2	0.25660	0.13333333	0.06666667	0.0869	0.0464	-0.0202
[3,]	3	0.31280	0.20000000	0.13333333	0.1226	0.0774	0.0107
[4,]	4	0.34590	0.26666667	0.20000000	0.1442	0.1224	0.0558
[5,]	5	0.35005	0.33333333	0.26666667	0.1472	0.1861	0.1194
[6,]	6	0.63790	0.40000000	0.33333333	0.3265	0.0735	0.0068
[7,]	7	0.65480	0.46666667	0.40000000	0.3360	0.1306	0.0640
[8,]	8	1.27160	0.53333333	0.46666667	0.5949	-0.0615	-0.1282
[9,]	9	1.52990	0.60000000	0.53333333	0.6647	-0.0647	-0.1313
[10,]	10	1.80300	0.66666667	0.60000000	0.7222	-0.0555	-0.1222
[11,]	11	2.05050	0.73333333	0.66666667	0.7636	-0.0302	-0.0969
[12,]	12	2.19060	0.80000000	0.73333333	0.7835	0.0165	-0.0501
[13,]	13	2.32140	0.86666667	0.80000000	0.8002	0.0664	-0.0002
[14,]	14	2.36740	0.93333333	0.86666667	0.8056	0.1277	0.0610
[15,]	15	2.43730	1.00000000	0.93333333	0.8135	0.1865	0.1198

```

EstdPrueba = max(D_mas,D_menos)
EstdPrueba

```

```
[1] 0.1865
```

Por lo tanto:

$$D_n = \sup_x |S_n(x) - F^*(x)| = \max \{ D^+, D^- \} = \max \{ 0.1865, 0.1198 \} = 0.1865$$

Lo comparamos con la tabla de valores críticos de la **Tabla Kolmogorov-Smirnov**, para un nivel de significancia $\alpha = 0.05$ $W_{0.05}=0.338$, de esta manera se tiene que $0.338 = W_{0.05} > D_n = 0.1865$, como el estadístico $W_{0.05}$ es mayor a comparación de $D_n=0.1865$. No Rechazamos H_0 .

Podemos concluir aceptando la prueba de lognormalidad con media 1 y varianza 0, con un nivel de significancia $\alpha = 0.05$. Es decir, Los datos se distribuyen $LogN(0,1)$.

```

#Ahora podemos utilizar la prueba en R
ks.test(F_,Sn,alternative = "two.sided")

```

Two-sample Kolmogorov-Smirnov test

```

data: F_ and Sn
D = 0.2, p-value = 0.9383
alternative hypothesis: two-sided

```

Capítulo 22

Prueba Lilliefors para Normalidad

La prueba de bondad de ajuste de Kolmogorov presentada anteriormente es una buena prueba para ver si una muestra aleatoria tiene alguna función de distribución especificada. La prueba de Kolmogorov está diseñada para usarse solo cuando la función de distribución hipotética está completamente especificada, es decir, cuando no hay parámetros desconocidos que deben estimarse a partir de la muestra. La prueba de bondad de ajuste Ji-cuadrada es lo suficientemente flexible como para permitir que algunos parámetros se estimen a partir de los datos. Simplemente se resta un grado de libertad para cada parámetro estimado descrita anteriormente.

Sin embargo, la prueba de Ji-cuadrada requiere que los datos se agrupen, y dicha agrupación de datos suele ser arbitraria. Además, la distribución del estadístico de prueba se conoce solo aproximadamente, y a veces el poder de la prueba de Ji-cuadrada no es muy bueno. Por estas razones, se buscan otras pruebas de bondad de ajuste, especialmente para distribuciones probadas con frecuencia.

La prueba de Kolmogorov se ha modificado para permitir su uso en varias situaciones en las que los parámetros se estiman a partir de los datos. En realidad, el estadístico de prueba permanece sin cambios, pero se utilizan diferentes tablas de valores críticos. Estas tablas ya no son las mismas para todas las distribuciones; cambian de una distribución hipotética a otra. La prueba sigue siendo una prueba no paramétrica porque la validez de la prueba (el nivel de α) no depende de supuestos no probados con respecto a la distribución de la población; en cambio, la forma de distribución de la población es la hipótesis que se está probando.

La primera modificación a la prueba de Kolmogorov es para probar la hipótesis compuesta de la normalidad. Es decir, la hipótesis nula establece que la población es una de la familia de distribuciones normales sin especificar la media o la varianza de la distribución normal. Ésta prueba fue presentada por primera vez por Hubert Lilliefors (1967). Una característica interesante de esta prueba es que este es uno de los primeros casos en que el compilador se utilizó para generar números aleatorios con el fin de obtener estimaciones precisas de los verdaderos cuantiles de la distribución exacta del estadístico de la prueba y además, aproximar a los parámetros a través del uso de los estimadores puntuales.

22.1. Datos

Los datos consisten en una muestra aleatoria X_1, X_2, \dots, X_n de tamaño n asociada con alguna función de distribución desconocida, denotada por $F(x)$.

Calculando la media muestral

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

para usarla como una estimación puntual de μ

y calculando como una estimación de σ

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2}$$

Después calcularemos los valores de muestra “normalizados” Z_i definidos por:

$$Z_i = \frac{X_i - \bar{X}}{s} \quad i = 1, 2, \dots, n$$

El estadístico de prueba se calcula a partir de Z_i s en lugar de a partir de la muestra aleatoria. original

22.2. Supuestos

- 1) La muestra es una muestra aleatoria.

22.3. Hipótesis

H_0 : La muestra aleatoria proviene de una población con distribución normal,
con media y desviación estándar desconocidas.

vs

H_a : La función de distribución de las X_i s no es normal.

22.4. Estadístico de Prueba

Normalmente, el estadístico de prueba es el mismo para la prueba de Kolmogorov de dos colas, definida como la distancia vertical máxima entre la función de distribución empírica de X_i s y la función de distribución normal con media $\mu = \bar{X}$ y desviación estándar $\sigma = s$. Sin embargo, el siguiente método para calcular el estadístico de prueba es un poco más fácil, ya que es equivalente al método indicado. Es decir, el cálculo del estadístico T_1 será en función de la Z_i s.

$$T_1 = \sup_x |F^*(x) - S(x)|$$

Regla de Decisión

Rechazo H_0 a un nivel de significancia α si $T_1 > W_{1-\alpha}$ donde $W_{1-\alpha}$ es el cuantil obtenido en las tablas correspondientes a nuestra prueba.

22.5. Ejemplo

Los siguientes datos, corresponden a una muestra aleatoria en la que mide la pérdida y ganancia de peso en KG de un grupo después de vacaciones.

0.6822, 3.994, -0.9705, -0.5575, -2.1532, 0.0829, 2.9224, 0.2425
-0.4962, -0.1621, 0.449, -0.8827, -0.8368, -1.5805, 0.386.

Se desea probar si los datos provienen de una distribución normal con μ y σ desconocidas. Realizar la prueba a un nivel de significancia del 95 %.

Paso 1 Prueba a utilizar **Prueba de Bondad de Ajuste Lilliefors para Normalidad**

Paso 2 Planteamiento de Hipótesis

$$\mathbf{H}_0 : \text{La muestra} \underset{vs}{\sim} N(\mu, \sigma^2).$$

$$\mathbf{H}_a : \text{La muestra} \approx N(\mu, \sigma^2).$$

Paso 3 Estadístico de Prueba

$$T_1 = \sup_z | F^*(z) - S(z) |$$

Paso 4 Procedimiento completo para el cálculo del Estadístico de Prueba:

- 1) Se procede a ordenar nuestras observaciones de menor a mayor.
- 2) Se obtienen los estimadores puntuales de distribución normal con los datos de la muestra,

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i = 0.07463333 \text{ y } s = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2} = 1.590808$$

- 3) Después calcularemos los valores de muestra “normalizados” Z_i definidos por:

$$Z_i = \frac{X_i - \bar{X}}{s} \quad i = 1, 2, \dots, 15$$

- 4) Se calcula la función empírica, como no tenemos ningún valor repetido:

$$S_n = \frac{i}{n} = \frac{1}{15}, \frac{2}{15}, \dots, 1.$$

- 5) Se calcula la función empírica menos un valor, es decir,

$$S_n = \frac{i-1}{n} = \frac{0}{15}, \frac{1}{15}, \dots, \frac{14}{15}.$$

- 6) Se calcula D^+ que es el resultado de la resta de la distribución conocida menos la distribución empírica, es decir:

$$D^+ = \max \{ S_n(Z_i) - F^*(Z_i) \}$$

- 7) Se calcula D^- que es el resultado de la resta de la distribución empírica menos uno menos la distribución conocida, es decir:

$$D^- = \max \{ S_n(Z_{i-1}) - F^*(Z_i) \}$$

- 8) Finalmente realizada la tabla, se calcula el máximo de las columnas D^+ y D^- de ésta manera, se tiene la siguiente tabla:

i X X_i $Z_i = \frac{x_i - \bar{x}}{s}$ $S_n(Z_i)$ $S_n(Z_{i-1})$ $F^*(Z_i)$ $D^+ = S_n(Z_i) - F^*(Z_i)$ $D^- = S_n(Z_{i-1}) - F^*(Z_i)$

1

0.6822

-2.1532

-1.4004

0.0666667

0.0000000

0.0806

-0.0139

-0.0806

2

3.9940

-1.5805

-1.0404

0.1333333

0.0666667

0.1490

-0.0156

-0.0823

3

-0.9705

-0.9705

-0.6569

0.2000000

0.1333333

0.2555

-0.0555

-0.1221

4

-0.5575

-0.8827

-0.6017
0.2666667
0.2000000
0.2736
-0.0069
-0.0736
5
-2.1532
-0.8368
-0.5729
0.3333333
0.2666667
0.2833
0.0500
-0.0166
6
0.0829
-0.5575
-0.3973
0.4000000
0.3333333
0.3455
0.0545
-0.0121
7
2.9224
-0.4962
-0.3588
0.4666667
0.4000000
0.3598
0.1068
0.0402
8
0.2425
-0.1621
-0.1488
0.5333333
0.4666667

0.4408
0.0925
0.0258
9
-0.4962
0.0829
0.0051
0.6000000
0.5333333
0.5020
0.0980
0.0313
10
-0.1621
0.2425
0.1055
0.6666667
0.6000000
0.5420
0.1246
0.0580
11
0.4490
0.3860
0.1957
0.7333333
0.6666667
0.5775
0.1558
0.0891
12
-0.8827
0.4490
0.2353
0.8000000
0.7333333
0.5930
0.2070
0.1403

13
 -0.8368
 0.6822
 0.3819
 0.8666667
 0.8000000
 0.6487
 0.2179
 0.1513
 14
 -1.5805
 2.9224
 1.7901
 0.9333333
 0.8666667
 0.9632
 -0.0298
 -0.0965
 15
 0.3860
 3.9940
 2.4637
 1.0000000
 0.9333333
 0.9931
 0.0069
 -0.0597

9) Entonces

$$D^+ = \max \{ S_n(Z_i) - F^*(Z_i) \} = 0.2179 \quad y \quad D^- = \max \{ S_n(Z_{i-1}) - F^*(Z_i) \} = 0.1513$$

Por lo tanto:

$$T_1 = \sup_x | S_n(z) - F^*(z) | = \max \{ D^+, D^- \} = \max \{ 0.2179, 0.1513 \} = 0.2179$$

Paso 5 Regla de Decisión

Este último resultado se compara con la tabla de valores críticos de la Tabla Lilliefors para Normalidad, para un nivel de significancia $\alpha = 0.05$ $W_{0.05}=0.2190$, de esta manera se tiene que $0.2190 = W_{0.05} > T_1 = 0.2179$, como el cuantil $W_{0.05}$ es mayor a comparación de $T_1=0.2179$.

No Rechazamos H_0 .

Paso 6 Conclusión

Podemos concluir que a un nivel de significancia $\alpha = 0.05$, no existe evidencia estadística suficiente para decir que la muestra no tiene distribución normal.

22.6. Ejemplo en R-Studio

Ahora haremos la réplica en R.

```
#Datos

i= 1:15
X=c(0.6822, 3.994, -0.9705, -0.5575, -2.1532, 0.0829, 2.9224,
    0.2425,-0.4962, -0.1621, 0.449, -0.8827, -0.8368, -1.5805, 0.386)
X_i=c(-2.1532,-1.5805,-0.9705,-0.8827,-0.8368,-0.5575,-0.4962, -0.1621, 0.0829, 0.2425,
    0.3860, 0.4490, 0.6822,2.9224,3.9940)
Z_i=c(-1.4004, -1.0404,-0.6569,-0.6017,-0.5729,-0.3973,-0.3588,-0.1488,0.0051,
    0.1055,0.1957,0.2353,0.3819,1.7901, 2.4637) #Calculado xi-xbarra/s
Sn_Zi=c(1/15,2/15,3/15,4/15,5/15,6/15,7/15,8/15,9/15,10/15,11/15,12/15,13/15,14/15,1)
Sn_Zi_1=c(0/15,1/15,2/15,3/15,4/15,5/15,6/15,7/15,8/15,9/15,10/15,11/15,12/15,13/15,14/15)
F_Zi=c(0.0806, 0.1490,0.2555,0.2736,0.2833,0.3455,0.3598,0.4408,0.5020,0.5420,0.5775,0.5930,
    0.6487,0.9632, 0.9931)
D_mas=c(-0.0139,-0.0156,-0.0555,-0.0069,0.0500,0.0545,0.1068,0.0925,0.0980,0.1246, 0.1558,
    0.2070,0.2179,-0.0298,0.0069) #Sn(Zi)-F*(Zi)
D_menos=c(-0.0806,-0.0823,-0.1221,-0.0736,-0.0166,-0.0121,0.0402,0.0258,0.0313,
    0.0580,0.0891,0.1403,0.1513,-0.0965,-0.0597) #Sn(Zi-1)-F*(Zi)

Tabla=cbind(i,X_i,Z_i,Sn_Zi,Sn_Zi_1,F_Zi,D_mas,D_menos)
Tabla
```

	i	X_i	Z_i	Sn_Zi	Sn_Zi_1	F_Zi	D_mas	D_menos
[1,]	1	-2.1532	-1.4004	0.06666667	0.00000000	0.0806	-0.0139	-0.0806
[2,]	2	-1.5805	-1.0404	0.13333333	0.06666667	0.1490	-0.0156	-0.0823
[3,]	3	-0.9705	-0.6569	0.20000000	0.13333333	0.2555	-0.0555	-0.1221
[4,]	4	-0.8827	-0.6017	0.26666667	0.20000000	0.2736	-0.0069	-0.0736
[5,]	5	-0.8368	-0.5729	0.33333333	0.26666667	0.2833	0.0500	-0.0166
[6,]	6	-0.5575	-0.3973	0.40000000	0.33333333	0.3455	0.0545	-0.0121
[7,]	7	-0.4962	-0.3588	0.46666667	0.40000000	0.3598	0.1068	0.0402
[8,]	8	-0.1621	-0.1488	0.53333333	0.46666667	0.4408	0.0925	0.0258
[9,]	9	0.0829	0.0051	0.60000000	0.53333333	0.5020	0.0980	0.0313
[10,]	10	0.2425	0.1055	0.66666667	0.60000000	0.5420	0.1246	0.0580
[11,]	11	0.3860	0.1957	0.73333333	0.66666667	0.5775	0.1558	0.0891
[12,]	12	0.4490	0.2353	0.80000000	0.73333333	0.5930	0.2070	0.1403
[13,]	13	0.6822	0.3819	0.86666667	0.80000000	0.6487	0.2179	0.1513
[14,]	14	2.9224	1.7901	0.93333333	0.86666667	0.9632	-0.0298	-0.0965
[15,]	15	3.9940	2.4637	1.00000000	0.93333333	0.9931	0.0069	-0.0597

```
EstadPrueba=max(D_mas,D_menos)
EstadPrueba
```

```
[1] 0.2179
```

Por lo tanto:

$$T_1 = \sup_x |S_n(z) - F^*(z)| = \max \{ D^+, D^- \} = \max \{ 0.2179, 0.1513 \} = 0.2179$$

Lo vamos a comparar con la tabla de valores críticos de la Tabla Lilliefors para Normalidad, para un nivel de significancia $\alpha = 0.05$ $W_{0.05}=0.2190$, de esta manera se tiene que $0.2190 = W_{0.05} > T_1 = 0.2179$, como el cuantil $W_{0.05}$ es mayor a comparación de $T_1=0.2179$. Entonces no Rechazamos H_0 y podemos concluir que a un nivel de significancia $\alpha =0.05$, no existe evidencia estadística suficiente para decir que la muestra no tiene distribución normal.

Ahora podemos utilizar la prueba en R

```
library(nortest) #prueba lilliefors
lillie.test(X)
```

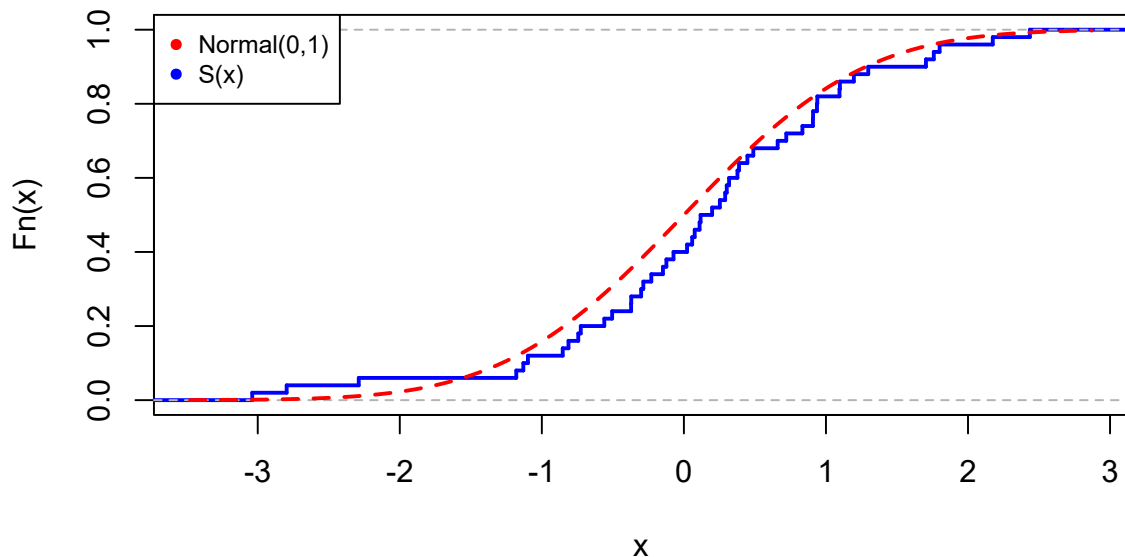
Lilliefors (Kolmogorov-Smirnov) normality test

```
data: X
D = 0.21793, p-value = 0.05356
```

Otro ejemplo en R En R fije la semilla 2020, y genere 250 observaciones distribuidas como una $N(0,1)$ y con ella realice:

1. Grafique la función de distribución empírica de las observaciones generadas.
2. Agregar sobre esa misma gráfica, la curva de la distribución verdadera $N(0,1)$.
3. Realizar la prueba Lilliefors de bondad de ajuste para probar que la muestra proviene de una distribución $N(0,1)$.

```
#Semilla y simulación
set.seed(2020)
x=rnorm(50,0,1)
#Gráfico de la función de distribución empírica
plot.ecdf(x,col="blue",verticals = TRUE, do.points = FALSE, lwd=2, main="")
#Gráfico de la distribución verdadera $N(0,1)$
curve(pnorm,add=TRUE, col="red", lty="dashed", lwd=2)
legend("topleft", c("Normal(0,1)", "S(x)"), cex=0.8, col=c("red", "blue"), pch=c(16,16))
```



```
#Prueba
library(nortest)
T1<-lillie.test(x)
T1
```

```
Lilliefors (Kolmogorov-Smirnov) normality test  
  
data:  x  
D = 0.08739, p-value = 0.4436
```

Observamos que la estadística de prueba tiene un valor de 0.08739 y su correspondiente p-value es 44.36 %, por lo tanto no rechazaremos H_0 y concluimos no existe evidencia suficiente para suponer que la distribución de la muestra no es $N(0, 1)$. Esperabamos que la prueba diera ese resultado ya que por construcción la muestra proviene de una distribución $N(0, 1)$.

Capítulo 23

Pueba de Lilliefors Exponencial

Otro problema importante de bondad de ajuste es la prueba para la distribución exponencial con media no especificada. Este problema es importante porque el supuesto de una distribución exponencial con media desconocida tiene muchas aplicaciones, particularmente donde las variables aleatorias bajo estudio representan el tiempo de espera o el tiempo en que ocurre cierto evento. Lilliefors en 1969 desarrolló una prueba análoga a la prueba de Kolmogorov-Smirnov y dió una tabla de valores críticos basados en simulaciones Monte Carlo.

23.1. Datos

Los datos consisten en una muestra aleatoria X_1, X_2, \dots, X_n de tamaño n que sigue una distribución exponencial con media $\hat{\lambda} = \frac{1}{\bar{X}}$, el cual corresponde al estimador puntual de la media.

Después calcularemos los valores de muestra Z_i definidos por:

$$Z_i = \frac{X_i}{\bar{X}} \quad i = 1, 2, \dots, n$$

El estadístico de prueba se calcula a partir de Z_i s en lugar de a partir de la muestra aleatoria original.

23.2. Supuestos

- 1) La muestra es una muestra aleatoria.

23.3. Hipótesis

H_0 : La muestra aleatoria proviene de una población con distribución exponencial:

$$F(x) = \begin{cases} 1 - e^{-\frac{x}{\lambda}} & \text{si } x > 0 \\ 0 & \text{si } x < 0 \end{cases}$$

Donde λ es un parámetro desconocido.

vs

H_a : La función de distribución de las X_i s no es exponencial.

23.4. Estadístico de Prueba

El estadístico de prueba está dada por la máxima distancia vertical:

$$T_2 = \sup_x | F^*(x) - S(x) |$$

Regla de Decisión

Rechazo H_0 a un nivel de significancia α si $T_2 > W_{1-\alpha}$ donde $W_{1-\alpha}$ es el cuantil obtenido en las tablas correspondientes a nuestra prueba.

23.5. Ejemplo

Dada la siguiente muestra

$$0.4976, 1.2514, 0.6619, 0.561, 1.0026, 0.3529, 0.8595, 1.6254, \\ 1.1514, 1.5181, 0.8642, 0.5206, 0.4229, 0.9825, 1.0183.$$

Se desea probar si la muestra sigue una distribución exponencial con parámetro λ desconocido.

Realizar la prueba con un nivel de significancia del 95 %.

Paso 1 Prueba a utilizar **Prueba de Bondad de Ajuste Lilliefors Exponencial**

Paso 2 Planteamiento de Hipótesis

H_0 : La muestra aleatoria proviene de una población con distribución exponencial

con λ parámetro desconocido.

vs

H_a : La función de distribución de las X_i s no es exponencial.

Paso 3 Estadístico de Prueba

$$T_2 = \sup_z | F^*(z) - S(z) |$$

Paso 4 Procedimiento completo para el cálculo del Estadístico de Prueba:

- 1) Se procede a ordenar nuestras observaciones de menor a mayor.
- 2) Se obtiene el estimador puntual de la distribución exponencial con los datos de la muestra, por lo que el parámetro λ es calculado como:

$$\hat{\lambda} = \frac{1}{\bar{X}} = \frac{1}{0.88602}; \quad \bar{X} = \frac{1}{n} \sum_{i=1}^n X_i = \frac{1}{15} \sum_{i=1}^{15} X_i = 0.88602$$

- 3) Después calcularemos los valores de muestra Z_i definidos por:

$$Z_i = \frac{X_i}{\bar{X}} \quad i = 1, 2, \dots, 15$$

- 4) Se calcula la función empírica, como no tenemos ningún valor repetido:

$$S_n = \frac{i}{n} = \frac{1}{15}, \frac{2}{15}, \dots, 1.$$

- 5) Se calcula la función empírica menos un valor, es decir,

$$S_n = \frac{i-1}{n} = \frac{0}{15}, \frac{1}{15}, \dots, \frac{14}{15}.$$

- 6) Se calcula D^+ que es el resultado de la resta de la distribución conocida menos la distribución empírica, es decir:

$$D^+ = \max \{ S_n(Z_i) - F^*(Z_i) \}$$

- 7) Se calcula D^- que es el resultado de la resta de la distribución empírica menos uno menos la distribución conocida, es decir:

$$D^- = \max \{ S_n(Z_{i-1}) - F^*(Z_i) \}$$

- 8) Finalmente realizada la tabla, se calcula el máximo de las columnas D^+ y D^- de ésta manera, se tiene la siguiente tabla:

i

X

X_i

$Z_i = \frac{X_i}{x}$

$S_n(Z_i)$

$S_n(Z_{i-1})$

$F^*(Z_i)$

$D^+ = S_n(Z_i) - F^*(Z_i)$

$D^- = S_n(Z_{i-1}) - F^*(Z_i)$

1

0.4976

0.3529

0.3982

0.0666667

0.0000000

0.3285

-0.2618

-0.3285

2

1.2514

0.4229
0.4773
0.1333333
0.0666667
0.3795
-0.2462
-0.3128
3
0.6619
0.4976
0.5616
0.2000000
0.1333333
0.4297
-0.2297
-0.2963
4
0.5610
0.5206
0.5875
0.2666667
0.2000000
0.4443
-0.1776
-0.2443
5
1.0026
0.5610
0.6331
0.3333333
0.2666667
0.4690
-0.1357
-0.2024
6
0.3529
0.6619
0.7470
0.4000000

0.3333333

0.5262

-0.1262

-0.1929

7

0.8595

0.8595

0.9700

0.4666667

0.4000000

0.6209

-0.1542

-0.2209

8

1.6254

0.8642

0.9753

0.5333333

0.4666667

0.6229

-0.0896

-0.1562

9

1.1514

0.9825

1.1088

0.6000000

0.5333333

0.6700

-0.0700

-0.1367

10

1.5181

1.0026

1.1315

0.6666667

0.6000000

0.6774

-0.0108

-0.0774

11

0.8642

1.0183

1.1492

0.7333333

0.6666667

0.6831

0.0501

-0.0164

12

0.5206

1.1514

1.2995

0.8000000

0.7333333

0.7273

0.0726

0.0059

13

0.4229

1.2514

1.4123

0.8666667

0.8000000

0.7564

0.1102

0.0435

14

0.9825

1.5181

1.7133

0.9333333

0.8666667

0.8197

0.1135

0.0469

15

1.0183

1.6254
 1.8344
 1.0000000
 0.9333333
 0.8403
 0.1596
 0.0930

9) Entonces

$$D^+ = \max \{ S_n(Z_i) - F^*(Z_i) \} = 0.1596 \quad y \quad D^- = \max \{ S_n(Z_{i-1}) - F^*(Z_i) \} = 0.0930$$

Por lo tanto:

$$T_2 = \sup_z | S_n(z) - F^*(z) | = \max \{ D^+, D^- \} = \max \{ 0.1596, 0.0930 \} = 0.1596$$

Paso 5 Regla de Decisión

Este último resultado se compara con la tabla de valores críticos de la Tabla Lilliefors Exponencial, para un nivel de significancia $\alpha = 0.05$ $W_{0.05}=0.2776$, de esta manera se tiene que $0.2776 = W_{0.05} > T_2 = 0.1596$, como el cuantil $W_{0.05}=0.2776$ es mayor a comparación de $T_2=0.2776$.

No Rechazamos H_0 .

Paso 6 Conclusión

Podemos concluir que a un nivel de significancia $\alpha=0.05$, no existe evidencia estadística suficiente para decir que la muestra no tiene distribución exponencial.

23.6. Ejemplo en R-Studio

Ahora haremos la réplica en R.

```
#Datos
i=1:15
X=c(0.4976, 1.2514, 0.6619, 0.561, 1.0026, 0.3529, 0.8595, 1.6254,
    1.1514, 1.5181, 0.8642, 0.5206, 0.4229, 0.9825, 1.0183)
X_i=c(0.3529,0.4229,0.4976,0.5206,0.5610, 0.6619,0.8595,0.8642,0.9825,1.0026,1.0183,
    1.1514,1.2514,1.5181,1.6254)
Z_i=c(0.3982, 0.4773, 0.5616,0.5875,0.6331,0.7470,0.9700,0.9753,1.1088,1.1315,1.1492,
    1.2995,1.4123,1.7133,1.8344)
Sn_Zi=c(1/15,2/15,3/15,4/15,5/15,6/15,7/15,8/15,9/15,10/15,11/15,12/15,13/15,14/15,1)
Sn_Zi_1=c(0/15,1/15,2/15,3/15,4/15,5/15,6/15,7/15,8/15,9/15,10/15,11/15,12/15,13/15,14/15)
F_Zi=c(0.3285,0.3795,0.4297,0.4443,0.4690,0.5262,0.6209,0.6229,0.6700,0.6774,0.6831,
    0.7273,0.7564,0.8197, 0.8403)
D_mas=c(-0.2618,-0.2462,-0.2297,-0.1776,-0.1357,-0.1262,-0.1542,-0.0896,-0.0700,
    -0.0108,0.0501,0.0726,0.1102,0.1135,0.1596)
D_menos=c(-0.3285,-0.3128,-0.2963,-0.2443, -0.2024,-0.1929,-0.2209,-0.1562,
    -0.1367,-0.0774,-0.0164,0.0059,0.0435,0.0469,0.0930)

Tabla=cbind(i,X_i,Z_i,Sn_Zi,Sn_Zi_1,F_Zi,D_mas,D_menos)
Tabla
```

	i	X_i	Z_i	Sn_Zi	Sn_Zi_1	F_Zi	D_mas	D_menos
[1,]	1	0.3529	0.3982	0.06666667	0.00000000	0.3285	-0.2618	-0.3285
[2,]	2	0.4229	0.4773	0.13333333	0.06666667	0.3795	-0.2462	-0.3128
[3,]	3	0.4976	0.5616	0.20000000	0.13333333	0.4297	-0.2297	-0.2963
[4,]	4	0.5206	0.5875	0.26666667	0.20000000	0.4443	-0.1776	-0.2443
[5,]	5	0.5610	0.6331	0.33333333	0.26666667	0.4690	-0.1357	-0.2024
[6,]	6	0.6619	0.7470	0.40000000	0.33333333	0.5262	-0.1262	-0.1929
[7,]	7	0.8595	0.9700	0.46666667	0.40000000	0.6209	-0.1542	-0.2209
[8,]	8	0.8642	0.9753	0.53333333	0.46666667	0.6229	-0.0896	-0.1562
[9,]	9	0.9825	1.1088	0.60000000	0.53333333	0.6700	-0.0700	-0.1367
[10,]	10	1.0026	1.1315	0.66666667	0.60000000	0.6774	-0.0108	-0.0774
[11,]	11	1.0183	1.1492	0.73333333	0.66666667	0.6831	0.0501	-0.0164
[12,]	12	1.1514	1.2995	0.80000000	0.73333333	0.7273	0.0726	0.0059
[13,]	13	1.2514	1.4123	0.86666667	0.80000000	0.7564	0.1102	0.0435
[14,]	14	1.5181	1.7133	0.93333333	0.86666667	0.8197	0.1135	0.0469
[15,]	15	1.6254	1.8344	1.00000000	0.93333333	0.8403	0.1596	0.0930

```
EstPrueba=max(D_mas,D_menos)
EstPrueba
```

```
[1] 0.1596
```

```
EstDPrueba
```

```
[1] 0.2179
```

El cuantil W que acumula $1 - \alpha$ de probabilidad, usando $\alpha = 0.05$ es $W = 0.2776$, encontrado en las tablas correspondientes. Por lo tanto tenemos que $T1 = 0.1596 < W = 0.2776$, entonces rechazamos la hipótesis nula. Y concluimos que hay evidencia suficiente para decir que la muestra no proviene de una distribución exponencial.

```
library(nortest) #prueba lilliefors
lillie.test(F_Zi) #Ocupamos estos datos ya que debemos recordar que
```

```
Lilliefors (Kolmogorov-Smirnov) normality test
```

```
data: F_Zi
D = 0.15238, p-value = 0.458
```

```
#es una distribución exponencial y F_Zi son datos estandarizados
```

Capítulo 24

Prueba Anderson-Darling

La prueba de Anderson Darling, al igual que la prueba de Lilliefors sirve para probar la hipótesis de que una muestra aleatoria sigue una cierta distribución especificada.

24.1. Datos

Los datos consisten en una muestra aleatoria X_1, X_2, \dots, X_n de tamaño n asociada con alguna función de distribución desconocida, denotada por $F(x)$.

24.2. Supuestos

- La muestra es una muestra aleatoria.

24.3. Hipótesis

Caso A (Prueba de 2 colas) Solo será este caso

$$H_0 : F(x) = F^*(x) \quad \forall \ x \text{ de } -\infty \text{ a } +\infty$$

vs

$$H_a : F(x) \neq F^*(x) \quad \text{para al menos un valor de } x.$$

Donde $F^*(x)$ es la distribución teórica que se quiere probar con un nivel de significancia α .

Para probar dicha hipótesis Anderson propone examinar las diferencias al cuadrados entre la distribución empírica de los datos $S_n(X)$ y la distribución teórica propuesta y completamente especificada $F^*(X)$ y luego integrar respecto a la distribución propuesta. A este tipo de pruebas se les conoce como funciones de distribución empíricas cuadráticas (**QEDF**) por sus siglas en inglés. De esta manera la estadística de la prueba **Anderson-Darling** se obtiene de integrar la siguiente función **QEDF**:

$$A_n^2 = n \int_{-\infty}^{\infty} (S_n(X) - F^*(X))^2 \frac{1}{F^*(X)(1 - F^*(X))}$$

Una característica importante es que se usa la expresión $\frac{1}{F^*(X)(1 - F^*(X))}$ debido a que se busca que las colas de distribución tengan un peso cuantificablemente mayor, con la finalidad de detectar diferencias en las colas de la distribución.

24.4. Estadístico de Prueba

Resolviendo la integral se obtiene la estadístico de la forma:

$$A_n^2 = -n - \frac{1}{n} \sum_{i=1}^n (2i-1) [\ln(F^*(X_{(i)})) + \ln(1 - F^*(X_{(n-i+1)}))]$$

Dado que el estadístico no depende de $F(X)$ y sólo depende de n entonces la distribución asintótica de **Anderson-Darling** es la que se muestra a continuación, asimismo se mostrarán algunos ajustes a la estadística con la finalidad de que la prueba sea más potente para determinados casos:

Caso	Ajuste en el estadístico	0.90	0.95	0.975	0.99
Todos los parámetros conocidos	A_n^2 para $n \geq 5$	1.933	2.492	3.070	3.857
Normal con $N(\bar{X}, S^2)$	$(1 + \frac{4}{n} + \frac{25}{n^2})A_n^2$	0.632	0.751	0.870	1.029
Exponencial con $\exp(\bar{X})$	$(1 + \frac{0.6}{n})A_n^2$	1.070	1.326	1.587	1.943
Weibull con $Weibull(\hat{\alpha}, \hat{\beta})$	$(1 + \frac{0.2}{\sqrt{n}})A_n^2$	0.637	0.757	0.877	1.038
Log-logística con $Log - log(\hat{\alpha}, \hat{\beta})$	$(1 + \frac{0.25}{\sqrt{n}})A_n^2$	0.563	0.660	0.769	0.906

24.5. Regla de Decisión

Rechazo H_0 a un nivel de significancia α si $A_n^2 > A_{1-\alpha}$ donde A_n^2 ya tiene el ajuste para cada caso mencionado anteriormente y $A_{1-\alpha}$ es el cuantil obtenido en las tablas correspondientes a nuestra prueba.

Para ejemplificar la prueba de *Anderson – Darling* veamos el siguiente ejemplo:

24.5.1. Ejemplo

Se desea probar si la siguiente muestra:

$$\begin{aligned} &-4.1302, 9.315, 3.9757, 8.49, 5.6204, -6.9098, -0.1426, -2.3838, \\ &-2.0039, 1.7349, 5.7442, 2.7931, 6.2938, 11.7337, -0.1318. \end{aligned}$$

Sigue una distribución Normal, para ello se realizará la prueba de Anderson Darling con un nivel de significancia del 5%.

Paso 1 Prueba a utilizar **Prueba de Bondad de Ajuste Anderson-Darling**

Paso 2 Planteamiento de Hipótesis

$$\mathbf{H}_0 : \text{La muestra} \sim N(\mu, \sigma^2).$$

vs

$$\mathbf{H}_a : \text{La muestra} \not\sim N(\mu, \sigma^2).$$

Paso 3 Estadístico de Prueba

$$A_n^2 = -n - \frac{1}{n} \sum_{i=1}^n (2i-1) [\ln(F^*(X_{(i)})) + \ln(1 - F^*(X_{(n-i+1)}))]$$

Paso 4 Procedimiento completo para el cálculo del Estadístico de Prueba:

- 1) Se procede a ordenar nuestras observaciones de menor a mayor.
- 2) Se obtienen los estimadores puntuales de distribución normal con los datos de la muestra, $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i = 2.66658$ y $s = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2} = 5.313212$
- 3) Se calcula la distribución propuesta, en este caso una distribución normal con media \bar{X} y varianza s^2 , es decir, $F^*(X)$, para ello se usa la aproximación a una normal estándar
- 4) Se calcula el primer sumando $\ln(F^*(X_{(i)}))$ el cual se denotará como $L1$, después se calculará el segundo sumando $\ln(1 - F^*(X_{(n-i+1)}))$ el cual se denotará como $L2$
- 5) Se realiza el sumando de manera puntal, es decir, calcular:

$$Q_i = \left(\frac{2i-1}{n} \right) [\ln(F^*(X_{(i)})) + \ln(1 - F^*(X_{(n-i+1)}))] \quad \text{para } i = 1, \dots, 15.$$

- 6) Y se realiza la siguiente tabla:

 i X X_i $F^*(X_i)$ $L_1 = \ln(F^*(X_i))$ $L_2 = \ln(1 - F^*(X_{n-i+1}))$ Q_i

1

-4.1302

-6.9098

0.0357

-3.3313

-3.1245

-0.4303

2

9.3150

-4.1302

0.1004

-2.2984

-2.2498

-0.9096

3

3.9757

-2.3838

0.1709

-1.7665

-1.9911

-1.2525

4

8.4900

-2.0039

0.1896

-1.6623

-1.3967

-1.4275

5

5.6204

-0.1426

0.2985

-1.2089

-1.2686

-1.4865

6

-6.9098

-0.1318

0.2992

-1.2066

-1.2408

-1.7948

7

-0.1426

1.7349

0.4304

-0.8430

-0.9095

-1.5189

8

-2.3838

2.7931

0.5094

-0.6743

-0.7123

-1.3866

9

-2.0039

3.9757

0.5973
-0.5153
-0.5628
-1.2218
10
1.7349
5.6204
0.7108
-0.3412
-0.3555
-0.8826
11
5.7442
5.7442
0.7187
-0.3301
-0.3545
-0.9586
12
2.7931
6.2938
0.7525
-0.2842
-0.2103
-0.7583
13
6.2938
8.4900
0.8634
-0.1468
-0.1874
-0.5570
14
11.7337
9.3150
0.8945
-0.1113
-0.1058
-0.3909

15
 -0.1318
 11.7337
 0.9560
 -0.0449
 -0.0363
 -0.1572

7) Finalmente se suma todos los Q_i s y se construye el estadístico:

$$A_n^2 = -n - \sum_{i=1}^n Q_i = -n - \frac{1}{n} \sum_{i=1}^n (2i-1) [\ln(F^*(X_{(i)})) + \ln(1 - F^*(X_{(n-i+1)}))]$$

$$A_n^2 = -n - \sum_{i=1}^n Q_i = -15 - (-15.1340) = 15.1340 - 15 = 0.1340$$

$$A_n^2 = 0.1340$$

8) Aplicando el ajuste de la tabla anterior:

Caso	Ajuste en el estadístico	0.90	0.95	0.975	0.99
Normal con $N(\bar{X}, S^2)$	$(1 + \frac{4}{n} + \frac{25}{n^2})A_n^2$	0.632	0.751	0.870	1.029

$$(1 + \frac{4}{n} + \frac{25}{n^2})A_n^2 = (1 + \frac{4}{15} + \frac{25}{15^2})A_n^2 = 0.1846$$

Paso 5 Regla de Decisión

Este último resultado se compara con la tabla de valores críticos de la Tabla Anderson-Darling para el caso de Normalidad, para un nivel de significancia $\alpha = 0.05$.

Por lo anterior $A_{0.95}=0.751$, de esta manera se tiene que $0.751 = A_{0.95} > A_n^2 = 0.1846$, como el valor crítico $A_{0.95}$ es mayor a comparación de $A_n^2=0.1846$.

\therefore No Rechazamos H_0 .

Paso 6 Conclusión

Podemos concluir que a un nivel de significancia $\alpha = 0.05$, no existe evidencia estadística suficiente para decir que la muestra no tiene distribución normal.

24.6. Ejemplo en R-Studio

Ahora haremos la réplica en R.

```
#Datos
i=1:15
n=15
X=c(-4.1302, 9.315, 3.9757, 8.49, 5.6204, -6.9098, -0.1426, -2.3838,
    -2.0039, 1.7349, 5.7442, 2.7931, 6.2938, 11.7337, -0.1318)
X_i=c(-6.9098, -4.1302, -2.3838, -2.0039, -0.1426, -0.1318, 1.7349, 2.7931, 3.9757, 5.6204,
    5.7442, 6.2938, 8.4900, 9.3150, 11.7337)
F_Xi=c( 0.0357, 0.1004, 0.1709, 0.1896, 0.2985, 0.2992, 0.4304, 0.5094, 0.5973, 0.7108,
    0.7187, 0.7525, 0.8634, 0.8945, 0.9560)
```

```
L1=c( -3.3313,-2.2984,-1.7665,-1.6623,-1.2089,-1.2066,-0.8430,-0.6743,-0.5153,-0.3412,-0.3301,
      -0.2842,-0.1468,-0.1113,-0.0449)      #ln(F*(Xi))
L2=c(-3.1245,-2.2498,-1.9911,-1.3967,-1.2686,-1.2408,-0.9095,
      -0.7123,-0.5628,-0.3555,-0.3545,-0.2103,-0.1874,-0.1058,-0.0363)  #ln(1-F*(Xn-i+1))
Qi=c(-0.4303,-0.9096,-1.2525,-1.4275,-1.4865,-1.7948,-1.5189,-1.3866,-1.2218,-0.8826,
      -0.9586,-0.7583,-0.5570,-0.3909,-0.1572)
```

```
#Ahora calcularemos el estadístico de prueba
```

```
A_2=-n-(sum(Qi))
```

```
A_2
```

```
[1] 0.1331
```

```
#Ahora aplicaremos el ajuste correspondiente
```

```
A_2ajust=(1+4/n+25/n^2)*A_2
```

```
A_2ajust
```

```
[1] 0.1833822
```

Por la tabla auxiliar tenemos $A_{0.95} = 0.751$, de esta manera se tiene que $0.751 = A_{0.95} > A_n^2 = 0.1846$, como el valor crítico $A_{0.95}$ es mayor a comparación de $A_n^2 = 0.1846$, entonces no rechazamos H_0 . Podemos concluir que a un nivel de significancia $\alpha=0.05$, no existe evidencia estadística suficiente para decir que la muestra no tiene distribución normal.

Ahora podemos utilizar la prueba en R

```
ad.test(X)      #Recordemos que la función no hace el ajuste
```

```
Anderson-Darling test of goodness-of-fit
Null hypothesis: uniform distribution
Parameters assumed to be fixed
```

```
data: X
```

```
An = Inf, p-value = 4e-05
```

Capítulo 25

Otras estadísticas

La prueba de Kuiper esta muy relacionada con la prueba Kolmogorov–Smirnov (o prueba K-S). Como la prueba $K-S$, esta prueba utiliza las estadísticas D^+ y D^- que representan las diferencias mas positivas y mas negativas entre las distribuciones que se estan comparando. La estadística de prueba de Kuiper es:

$$V = D^+ + D^-$$

Con este pequeño cambio, la prueba de Kuiper es tan sensible en las colas como lo es en la mediana de la distribución.

Las pruebas **Anderson–Darling** y **Cramér–von Mises** pertenecen a un grupo llamado **estadísticas EDF cuadráticas**, en donde el término EDF se refiere a que se basan en la función de distribución empírica.

Este grupo de estadísticas esta definido de la siguiente manera:

$$n \int_{-\infty}^{\infty} (S(x) - F_0(x))^2 w(x) dF_0(x)$$

En donde la diferencia entre la distribución empírica y la hipotética está calculada con el término cuadrático y el término $w(x)$ es un peso que se da esas diferencias.

Cuando $w(x) = 1$ entonces se tiene la estadística de Cramér–von Mises; cuando $w(x) = [F_0(x)(1 - F_0(x))]^{-1}$ entonces se tiene la estadística de Anderson–Darling, la cual por construcción asigna mayores pesos a observaciones en las colas de la distribución.

En R, la libreria “goftest” contiene las pruebas Anderson–Darling y Cramér–von Mises entre otras.

Retomando el ejemplo de las alturas de los pinos, probaremos ahora que las alturas tienen distribución normal y exponencial.

```
X=Loblolly$height
meanx=mean(X)
meanx
```

```
[1] 32.3644
```

```
sdx=sd(X)
sdx
```

```
[1] 20.6736
```

```
library(goftest)
ad.test(X,null = "pnorm",mean=meanx,sd=sdx)
```

Anderson-Darling test of goodness-of-fit
 Null hypothesis: Normal distribution
 with parameters mean = 32.3644047619048, sd = 20.6736047504145
 Parameters assumed to be fixed

data: X
 An = 2.7319, p-value = 0.03765

```
cvm.test(X,null = "pnorm",mean=meanx,sd=sdx)
```

Cramer-von Mises test of goodness-of-fit
 Null hypothesis: Normal distribution
 with parameters mean = 32.3644047619048, sd = 20.6736047504145
 Parameters assumed to be fixed

data: X
 omega2 = 0.38648, p-value = 0.07825

```
ad.test(X,null = "pexp",rate=1/meanx)
```

Anderson-Darling test of goodness-of-fit
 Null hypothesis: exponential distribution
 with parameter rate = 0.0308981428009166
 Parameters assumed to be fixed

data: X
 An = 4.5223, p-value = 0.004895

```
cvm.test(X,null = "pexp",rate=1/meanx)
```

Cramer-von Mises test of goodness-of-fit
 Null hypothesis: exponential distribution
 with parameter rate = 0.0308981428009166
 Parameters assumed to be fixed

data: X
 omega2 = 0.79946, p-value = 0.007174

Capítulo 26

Ejercicios

1. Se lanza un dado 600 veces se obtuvieron los siguientes resultados.

Número del dado	1	2	3	4	5	6
Observaciones	87	96	108	89	122	98

¿El dado está balanceado (es decir, los datos tienen distribución uniforme con proba $1/6$)? Use $\alpha = 0.10$

2. Cierta banco otorga crédito a las personas con una tasa preferencial, de tal manera que los acreditados pueden pagar en cualquier momento desde que piden el préstamo hasta 8 semanas posteriores para que les sea respetada la tasa preferencial. Se seleccionaron aleatoriamente a 1,000 personas y observó su comportamiento de pago, generando de esta manera la siguiente tabla de frecuencia:

Semana	Créditos Pagados
Menos de 1 semana	64
$1 \leq x < 2$	191
$2 \leq x < 3$	283
$3 \leq x < 4$	241
$4 \leq x < 5$	140
$5 \leq x < 6$	51
$6 \leq x < 7$	25
$7 \leq x < 8$	4
8 semanas o más	1

Probar que el pago de estos créditos, sigue una distribución binomial con parámetros $n = 10$ y $p = 0.25$.

3. Cinco niños de cuarto grado de primaria fueron seleccionados al azar de todos los niños de ese grado en la escuela “15 de septiembre”, para participar en una carrera. Sus tiempos en segundos fueron: 6.3, 4.2, 4.7, 6.0, y 5.7. Probar con la hipótesis de que los tiempos provienen de una distribución uniforme en el intervalo de 4 a 8 segundos.

4. La siguiente muestra aleatoria hace referencia a los rendimientos positivos de cierta acción a lo largo del tiempo.

0.2513, 0.2566, 0.3459, 0.6379, 2.0505, 1.803, 2.1906, 1.5299, 0.35005, 0.3128, 1.2726, 2.3674, 2.3214, 2.4373, 0.6548.

¿Usted piensa que la anterior muestra sigue una distribución normal?

Realizar la prueba correspondiente para verificar que su suposición es cierta con un nivel de confianza del 90 %.

5. Se obtuvieron sesenta y dos observaciones de un experimento, y se plantea la pregunta de si dichas observaciones provienen de una distribución normal con media 12 y desviación estándar 3. Ninguna observación se encontró por debajo del cuartil inferior de la distribución, 35 estuvieron por arriba de

cuartil superior, 22 tomaron valores menores a la mediana, y 5 estuvieron entre la mediana y el cuartil superior. ¿Es posible concluir que las observaciones provienen de la distribución mencionada?

6. Dada la siguiente muestra

0.6379, 1.5299, 0.35005, 2.0505, 2.1906, 0.3459, 2.3214, 0.3128 0.6548, 2.4373, 1.803, 2.3674, 1.2716, 0.2566, 0.2513.

Se desea hacer el siguiente contraste:

$$\mathbf{H}_0 : \text{Los datos} \sim \text{LogN}(0,1) \text{ vs } \mathbf{H}_a : \text{Los datos} \not\sim \text{LogN}(0,1)$$

Realice la prueba de Kolmogorov al 5 % de significancia.

7. Se desea probar la hipótesis de que los tiempos entre las llegadas de los pacientes a un hospital con una emergencia se distribuyen exponencial con media \bar{x} . Para ello se registró el tiempo transcurrido entre las llegadas sucesivas de pacientes en una mañana. El tiempo en minutos es el siguiente:

14.3,	38.0,	3.8,
10.8,	6.1,	10.1,
3.6,	6.2,	12.8,
22.1,	4.2,	4.6,
1.5,	3.3,	1.2,
20.0,	7.1,	8.1.

Pruebe la hipótesis con un nivel de significancia del 5 %.

8. Se desea probar la hipótesis de que las tasas de interés de un determinado producto financiero tiene el comportamiento de una variable aleatoria como función de distribución normal.

9.1, 5, 7.3, 7.4, 5.5, 8.6, 7, 4.3, 4.7, 8, 4, 8.5, 6.4, 6.1, 5.8, 9.5, 5.2, 6.7, 8.3, 9.2.

9. El gerente de una tienda quiere probar la hipótesis de que los clientes llegan aleatoriamente a su tienda, para ello registro el tiempo transcurrido entre las llegadas sucesivas de clientes en una mañana. El tiempo en minutos es el siguiente:

3.6, 6.2, 12.9, 14.2, 38.0, 3.8,
10.8, 6.1, 10.1, 22.1, 4.2, 4.6,
1.4, 3.3, 8.2.

Pruebe la hipótesis nula de que el tiempo entre las llegadas de los clientes se distribuyen con función de distribución exponencial.

Bibliografía

Conover, W. J. (1998). *Practical nonparametric statistics*, volume 350. John Wiley & Sons.