

CONTRASTES NO PARAMETRICOS

Conceptos básicos

En los “contrastes paramétricos”, se requiere una serie de supuestos estadísticos para asegurar la validez de las pruebas.

Particularmente, las "pruebas paramétricas" *sobre la media o sobre la varianza* exigen que la variable se mida (como mínimo) en escalas de intervalo. Ello hace que no sea posible efectuarlas cuando la escala sea ordinal.

Las pruebas no paramétricas pueden ser efectuadas cuando el nivel de medida sea ordinal, así como las condiciones de los supuestos estadísticos (como es el caso de homogeneidad de varianzas, normalidad de las puntuaciones) son menos estrictas.

Contraste de Aleatoriedad

En una determinada secuencia de sucesos observables la interacción entre elementos iguales lo denominaremos como racha. El n° de elementos de una racha se llama Longitud.

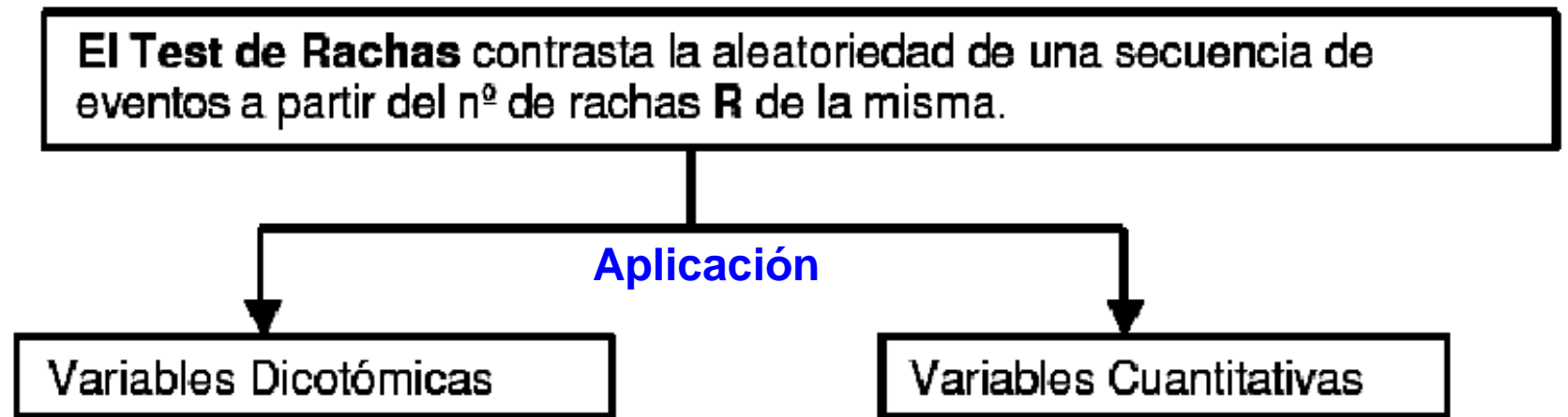
Ejemplo, la secuencia:

aaa bb a bb aa



aquí existen 5 rachas, de dos sucesos **a** y **b** de longitudes (3) ; (2) ; (1) ; (2); (2).

En muchas ocasiones se está interesado en la **aleatoriedad de una secuencia de sucesos** donde en ocasiones la aparición, en una muestra, de un elemento condiciona la aparición de otro. Si ocurre esto la muestra no es aleatoria, incumpliendo así una de las hipótesis básicas de los procedimientos de inferencia.



Ejemplo

Dada la sucesión de eventos: **aaa bb a bb aa**, tenemos:

R1: n° de rachas del elemento **1 (a)**.

N1: n° de veces que aparece el elemento **1 (a)** en la muestra.

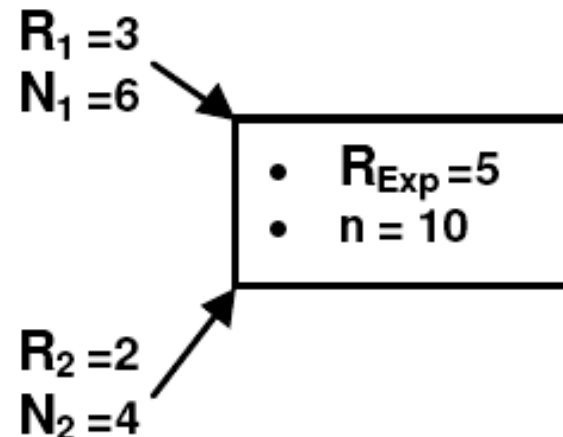
R2: n° de rachas del elemento **2 (b)**.

N2: n° de veces que aparece el elemento **2 (b)** en la muestra.

El total de la muestra es : **$n = N1 + N2$** .

El total de rachas es : **$R_{exp} = R1 + R2$**

Según lo visto, en nuestro ejemplo:



Procedimiento:

Hipótesis

Ho: La muestra es aleatoria

Ha: La muestra no es aleatoria

Test (Si $N_i > 20$).

Bajo Ho, el estadístico **R** tomara valores 2 , 3 , ... $N_1 + N_2$ tiene una probabilidad **$P[R=r]$** y la distribución asintótica de **R** es una NORMAL con:

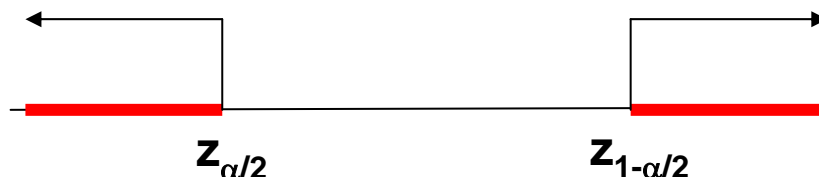
$$E[R] = \frac{2 \cdot N_1 \cdot N_2 + n}{n}$$

$$Var[R] = \frac{2 \cdot N_1 \cdot N_2 (2 \cdot N_1 \cdot N_2 - n)}{n^2 \cdot (n - 1)}$$

$$Z_{\text{exp}} = \frac{|R_{\text{exp}} - E[R]| - 0.5}{\sigma_R}$$

Regla de Decisión

Rechazaremos **Ho** si $|Z_{\text{exp}}| > Z_{1-\alpha/2}$ donde Z sigue una $N(0,1)$.

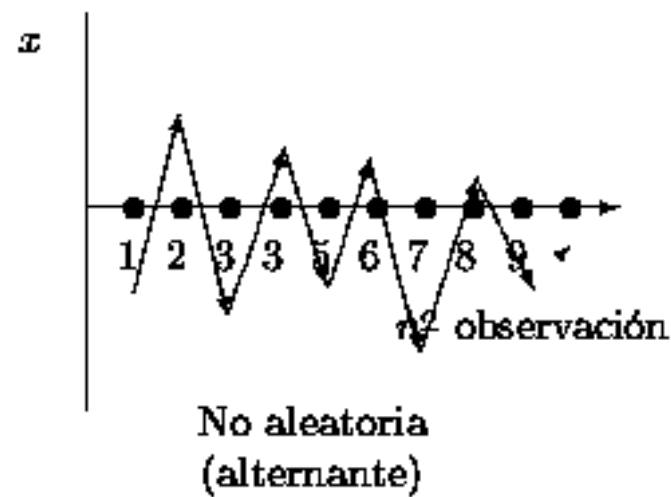
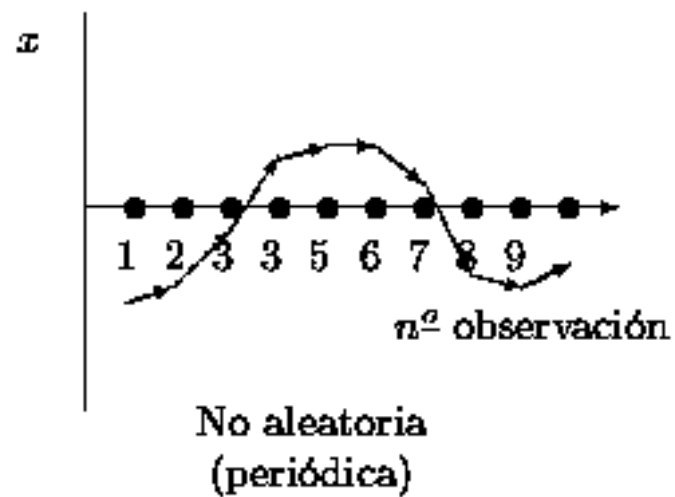
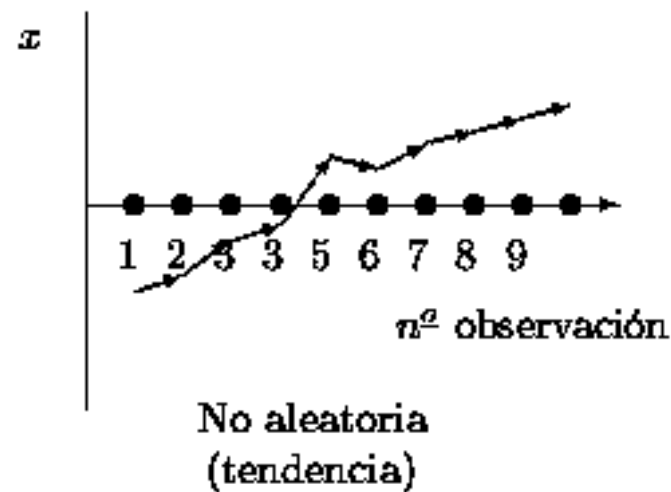
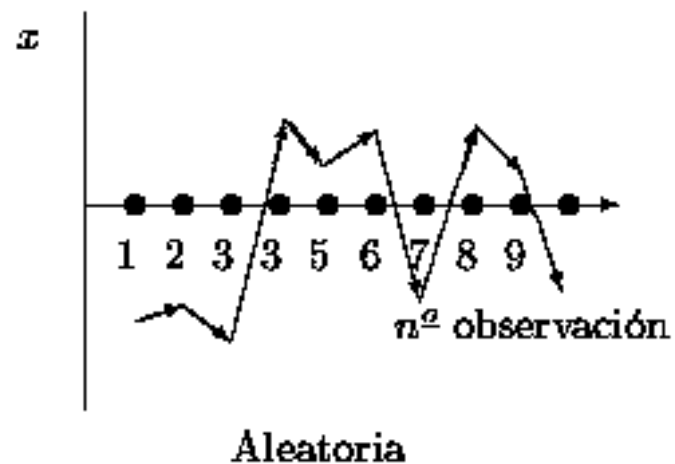


Para valores pequeños de $Ni \leq 20$ utilizar la tabla de aleatoriedad, donde para cada nivel de significación vienen dados los valores críticos $RI = r_{a/2}$ ó $RS = r_{1-a/2}$.

Variables Cuantitativas

Cuando los datos son cuantitativos realizar el siguiente proceso:

1. Se calcula la mediana muestral.
2. Se obtiene la diferencia entre cada valor y la mediana, asignándole el signo correspondiente.
3. Se eliminan los valores 0.
4. Se procede igual que en el apartado anterior tomando la sucesión de signos + ó -



Ejemplo

Una empresa pretende enviar a sus empleados a realizar unos cursos de formación en el extranjero y para ello decide escoger entre sus trabajadores a 50 empleados entre ambos sexos (a=masculino y b=femenino). Siendo la sucesión de sexos la siguiente:

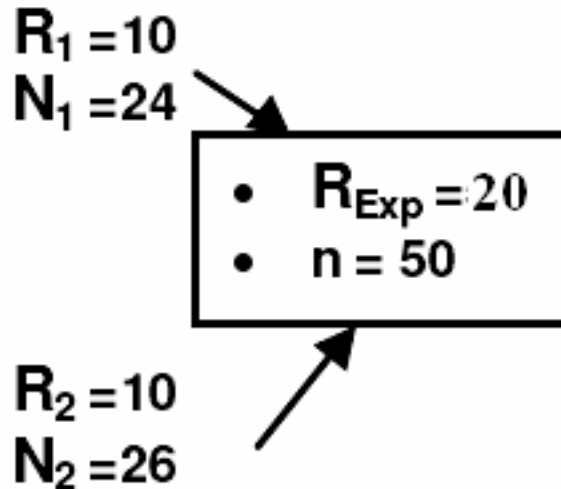
**aaa bb a b aaaaa bbbbbb aa bbb aaaa bbbbbb a b a b a
bb aaa bbbb aaa b**

Contrastar si el procedimiento seguido ha sido aleatorio para un nivel del 5%

Solución

Ho: La sucesión por género es aleatoria

Ha: La sucesión por género no es aleatoria



Test

$$Z_{exp} = \frac{|20 - 25.96| - 0.5}{3.4935} = 1.56$$

Donde;

$$E[R] = \frac{2 \cdot 24 \cdot 26 + 50}{50} = 25.96$$

$$Var[R] = \frac{2 \cdot 24 \cdot 26(2 \cdot 24 \cdot 26 - 50)}{50^2 \cdot (50 - 1)} = 1220493 \Rightarrow \sigma_{R_{exp}} = 3'4935$$

Como $N_i > 20$ y $|Z_{exp}| = 1.56 < Z_{0.025} = 1.96 \Rightarrow$ no rechazamos H_0 .

Por tanto la selección por género es aleatoria.

Test de Kolmogorov – Smirnov de Bondad de ajuste

Es un Test alternativo a la Chi-cuadrado. Tiene el mismo objetivo: contrastar la **Hipótesis Nula** de que X sigue una distribución específica $F(x)$ frente a la Hipótesis Alternativa de que no la siga.

Debe utilizarse si:

- **Modelo propuesto es de tipo continuo**
- **Tamaño muestral es pequeño**

El fundamento del contraste radica en comparar la distribución de las Frecuencias observadas (F_o) con la distribución propuesta bajo la H_o : $F(x)$ y si esta comparación revela diferencias significativas se rechazara la H_o

F_o	F. Acum. Relat.	N_i / N
F_E	D. NORMAL	F_E=p(Z) → Z=(x_i-μ)/σ
	D.UNIFORME	F_E=(x-min) / (max - min)
	D.POISSON	F_E=e^{-λ} λⁱ/i!

Estadístico

$$D = \max \{D_1 ; D_2\}$$

donde:

$$D_1 = |F_o(x_i) - F_e(x_i)|$$

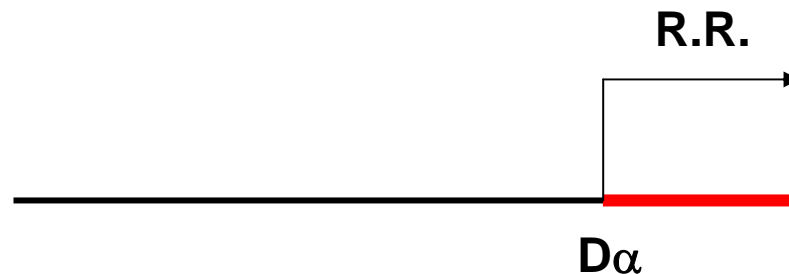
$$D_2 = |F_o(x_{i-1}) - F_e(x_i)|$$

Regla de Decisión

Fijado el nivel de significación buscamos en la tabla los valores críticos (del Test K-S) el valor D_α que depende de α y de N de manera que:

Rechazamos H_0 (las desviaciones entre la muestra y la población son significativas) $D_i > D_\alpha$.

No rechazamos H_0 (las desviaciones entre la muestra y la población no son significativas) $D_i < D_\alpha$.



Ejemplo

Realizar un test de K-S para un $\alpha = 0.1$ para saber si los datos procedentes de una muestra de tamaño 10 siguen una distribución Normal de media 10.84 y desviación 3.5.

{10.5; 8; 15; 12.1; 4.1; 12.1; 8; 10.5; 16; 12.1}

Solución

Ho: Los datos siguen un modelo normal

Ha: Los datos no siguen un modelo normal

X_i	n_i	N_i	F_o	F_e
4'1	1	1	0'1	0'027
8	2	3	0'3	0'209
10'5	2	5	0'5	0'464
12'1	3	8	0'8	0'641
15	1	9	0'9	0'882
16	1	10	1	0'929
	10			

Las frecuencias esperadas se calculan como:

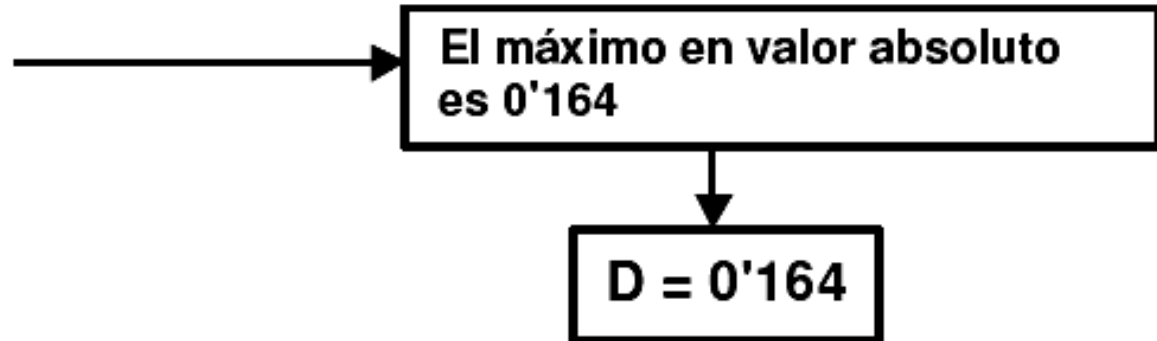
$$F_e(x_1) \rightarrow Z_1 = \frac{4'1 - 10'84}{3'5} = -1'93 \rightarrow P(z) \rightarrow 0'027$$

$$F_e(x_6) \rightarrow Z_6 = \frac{16 - 10'84}{3'5} = 1'47 \rightarrow P(z) \rightarrow 0'929$$

Test

$F_o(x_i) - F_e(x_i)$	D_1	$F_o(x_{i-1}) - F_e(x_i)$	D_2
0'1 - 0'027	0'073	0 - 0'027	-0'027
0'3 - 0'209	0'091	0'1 - 0'209	-0'109
0'5 - 0'464	0'036	0'3 - 0'464	-0'164
0'8 - 0'641	0'159	0'5 - 0'641	-0'141
0'9 - 0'882	0'018	0'8 - 0'882	-0'082
1 - 0'929	0'07	0'9 - 0'929	-0'029

$$\begin{aligned}\text{Max } D_1 &= 0'159 \\ \text{Max } D_2 &= -0'164\end{aligned}$$



Para un nivel de $\alpha = 0.1$ y $N = 10$, en la tabla K-S se tiene $D_{0.10} = 0.368$.

Decisión

Como $D = 0.164 < D_{\alpha} = 0.368$, entonces no rechazamos H_0 .

Gráfica de Probabilidad

La gráfica de probabilidad es un método gráfico que permite determinar si una muestra de datos se ajusta a una distribución propuesta en base a un análisis visual subjetivo.

Originalmente esta gráfica se realizaba sobre un papel especial llamado papel de probabilidad diseñado con las escalas adecuadas para las diferentes distribuciones.

Procedimiento:

1. Se ordena la muestra de menor a mayor: x_1, x_2, \dots, x_N
2. Se grafica la frecuencia acumulada observada $(i-0.5)/N$ contra el valor de los datos ordenados
3. Si los puntos obtenidos se desvían significativamente de una línea recta, el modelo propuesto no será el apropiado.

Ejemplo:

Las siguientes son diez observaciones sobre la duración en minutos de las baterías de computadoras portátiles:

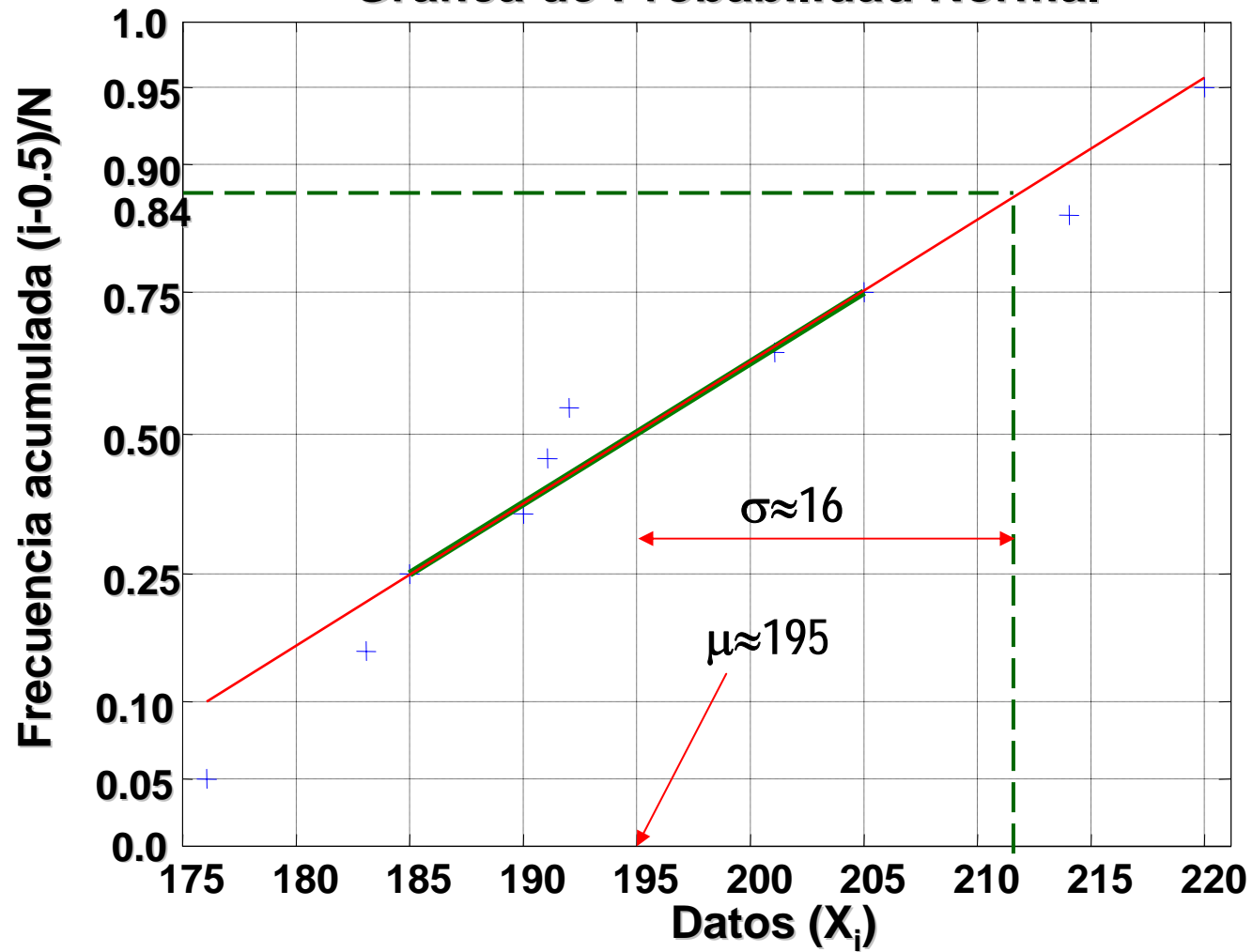
176, 183, 185, 190, 191, 192, 201, 205, 214, 220

Utilizar la gráfica de probabilidad para determinar si la muestra corresponde a una distribución Normal.

Procedimiento: Formamos la tabla de los datos ordenados y las frecuencias acumuladas $(i-0.5)/N$ siguiente:

i	1	2	3	4	5	6	7	8	9	10
x_i	176	183	185	190	191	192	201	205	214	220
$(i-0.5)/10$	0.05	0.15	0.25	0.35	0.45	0.55	0.65	0.75	0.85	0.95

Gráfica de Probabilidad Normal

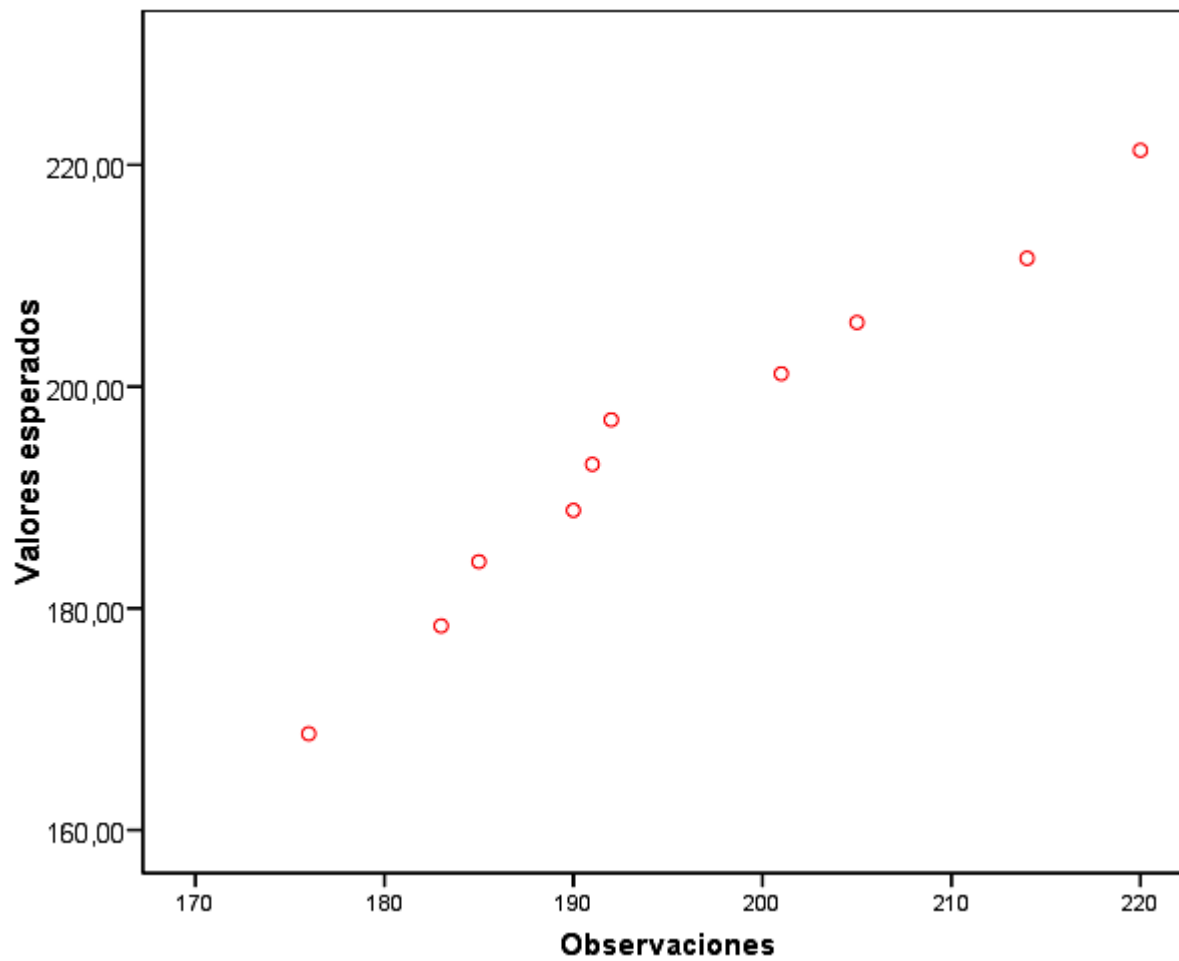


Observaciones:

- Al analizar la gráfica debe recordarse que el eje vertical está graduado en percentiles, por ello la media se encuentra en el percentil 50.
- Los puntos más confiables son los que están entre el percentil 25 y el 75, de hecho, la línea trazada debe unir estos percentiles
- Se puede obtener una gráfica sobre papel normal ajustando la escala vertical de acuerdo a z_i , donde $F(z_i) = (i-0.5)/N$, para el ejemplo:

i	1	2	3	4	5	6	7	8	9	10
x_i	176	183	185	190	191	192	201	205	214	220
$(i-0.5)/10$	0.05	0.15	0.25	0.35	0.45	0.55	0.65	0.75	0.85	0.95
z_i	-1.64	-1.04	-0.67	-0.39	-0.13	0.13	0.39	0.67	1.04	1.64
Ei	168.68	178.42	184.21	188.83	192.99	197.01	201.17	205.79	211.58	221.32

Gráfica de Probabilidad Normal



La Prueba de Signos

La prueba t supone que los datos se distribuyen normalmente. La prueba del signo prescinde de tal hipótesis y es mucho más fácil de realizar. Se puede utilizar de diferentes formas, la más simple se describe a continuación:

Hipótesis

Ho: $\mu = \mu_0$ vs. Ha: una alternativa apropiada

Equivalente

Ho: $Me = Me_0$

Ho: $p = 0.5$

Basándose en una muestra aleatoria de tamaño n , se reemplaza cada valor muestral mayor que μ_0 por un signo más y cada valor muestral menor que μ_0 por un signo menos. Se ignoran por completo aquellos valores que son iguales a μ_0 . Para contrastar si la preponderancia de signos menos, es significativa se utiliza la ley de la binomial acumulada.

Test

Esta ley establece que la probabilidad de que aparezcan r signos menos entre n signos.

$$p(r) = nC_r p^r q^{n-r}$$

Si la probabilidad experimental es menor que un nivel de significación α , la hipótesis nula debe rechazarse.

Regla de Decisión

Si $H_a: \mu < \mu_0$ debemos hallar $p = P(X \leq r)$, en este caso r describe el signo menos frecuente. Entonces si $p = P(X \leq r) < \alpha$, rechazamos H_0 .

Particularmente en el caso unilateral hacia la izquierda y derecha respectivamente:

a) Rechazaremos H_0 , solo si la proporción de signos (+) es suficientemente menor que 0.5, entonces P se calcula:

$$P\left(X \leq x, p = \frac{1}{2}\right)$$

Es menor o igual al nivel de significancia (alfa) preestablecido.

b) Rechazamos H_0 , si la proporción de signos (+) es bastante mayor que 0.5 , cuando x es grande.

$$P\left(X \geq x, p = \frac{1}{2}\right)$$

c) Para el caso bilateral, rechazamos H_0 , si la proporción de signos (+) es significativamente menor o igual a 0.5, depende que x sea bastante grande o pequeña. Si:

$$x < \frac{n}{2}$$

$$P = 2P\left(X \leq x, p = \frac{1}{2}\right)$$

$$x > \frac{n}{2}$$

$$P = 2P\left(X \geq x, p = \frac{1}{2}\right)$$

Ejemplo

Se informa acerca de un estudio en el que se modela las utilidades diarias de un empresa como se muestran a continuación. Se desea probar la hipótesis de que la mediana de las utilidades es de 2000 u.m., utilizando un nivel de significancia de 0.05.

Día	Utilidad x
1	2158.7
2	1678.15
3	2316.00
4	2061.30
5	2207.50
6	1708.30
7	1784.70
8	2575.10
9	2357.90
10	2256.70
11	2165.20
12	2399.55
13	1779.8
14	2336.75
15	1765.30
16	2053.50
17	2414.40
18	2200.50
19	2654.20
20	1753.70

Solución

El parámetro de interés es la mediana de las utilidades.

Hipótesis

$$\alpha = 0.05$$

$H_0: Me = 2000$

$H_a: Me \neq 2000$.

Test

El estadístico de prueba de las diferencias es:

X: Nro. de signos positivos, $X=14$

Día	x	DIFERENCIAS X - 2000	SIGNO
1	2158.7	+158.7	+
2	1678.15	-321.85	-
3	2316.00	+316	+
4	2061.30	+61.30	+
5	2207.50	+207.5	+
6	1708.30	-291.7	-
7	1784.70	-215.3	-
8	2575.10	+575.1	+
9	2357.90	+357.9	+
10	2256.70	+256.7	+
11	2165.20	+165.2	+
12	2399.55	+399.55	+
13	1779.8	-220.20	-
14	2336.75	+336.75	+
15	1765.30	-234.7	-
16	2053.50	+53.5	+
17	2414.40	+414.4	+
18	2200.50	+200.5	+
19	2654.20	+654.2	+
20	1753.70	-246.3	-

Decisión

Se rechaza H_0 si el valor p-value correspondiente a $X=14$ es menor o igual a 0.05. Como $X=14$ es mayor que $n/2=10$, el valor p-value se calcula:

$$P = 2P\left(X \geq 14, p = \frac{1}{2}\right)$$
$$P = 2 \sum_{14}^{20} \binom{20}{x} (0.5)^x (0.5)^{20-x}$$
$$P = 0.1153$$

Como $P = 0.1153$ no es menor que 0.05, no es posible rechazar la hipótesis nula.

Ejemplo

Una compañía de taxis trata de decidir si el uso de llantas radiales en lugar de llantas regulares con cinturón mejora la economía de combustible. Se equipan 16 automóviles con llantas radiales y se manejan por un recorrido de prueba establecido. Sin cambiar de conductores, se equipan los mismos autos con llantas regulares con cinturón y se manejan una vez más por el recorrido de prueba. Se registra el consumo de gasolina, en kilómetros por litro, como se muestra en la tabla. ¿Se puede concluir en el nivel de significancia de 0.05 que los autos equipados con llantas radiales obtienen mejores economías de combustible que los equipados con llantas regulares con cinturón?

Automóvil	Llantas radiales	Llantas con cinturón
1	4.2	4.1
2	4.7	4.9
3	6.6	6.2
4	7.0	6.9
5	6.7	6.8
6	4.5	4.4
7	5.7	5.7
8	6.0	5.8
9	7.4	6.9
10	4.9	4.9
11	6.1	6.0
12	5.2	4.9
13	5.7	5.3
14	6.9	6.5
15	6.8	7.1
16	4.9	4.8

Solución:

$$H_0; \tilde{\mu}_R - \tilde{\mu}_C = 0$$

$$H_1; \tilde{\mu}_R - \tilde{\mu}_C > 0$$

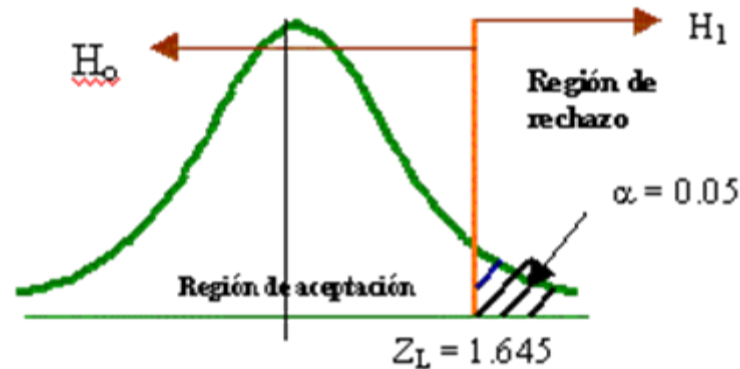
Al observar las diferencias se ve que sólo existe una $n=14$, ya que se descartan los valores de cero.

Test

Se tiene $X+ = 11$

Automóvil	Llantas radiales	Llantas con cinturón	d
1	4.2	4.1	+
2	4.7	4.9	-
3	6.6	6.2	+
4	7.0	6.9	+
5	6.7	6.8	-
6	4.5	4.4	+
7	5.7	5.7	0
8	6.0	5.8	+
9	7.4	6.9	+
10	4.9	4.9	0
11	6.1	6.0	+
12	5.2	4.9	+
13	5.7	5.3	+
14	6.9	6.5	+
15	6.8	7.1	-
16	4.9	4.8	+

$$Z = \frac{r^+ - 0.5n}{0.5\sqrt{n}} = \frac{11 - (0.5)(14)}{0.5\sqrt{14}} = 2.14$$



Decisión y conclusión:

Como 2.14 es mayor a 1.645 se rechaza H_0 y se concluye con un $\alpha = 0.05$ que las llantas radiales mejoran la economía de combustible.

NOTA:

Si $n > 10$ entonces podemos aproximar a la normal ($np > 5$)