PREDICCIÓN POBLACIONAL A PARTIR DE MUESTRAS

Laboratorio – Estadística Inferencial

Roberto Barroso Garcia
Enviado a Universidad UNIR
22 de junio de 2021
roberto.barroso.garcia@outlook.com

Contenido

Contrastes de Hipótesis

Una hipótesis estadística es una afirmación respecto a alguna característica dentro de una población.

Un contraste de hipótesis es una técnica estadística que se utiliza para comprobar la validez de una afirmación (hipótesis) en base a la información recogida en una muestra de observaciones.

Mediante ciertas técnicas se proponen unos intervalos de aceptación y rechazo, donde con cierto margen de error veremos si nuestra hipótesis cae dentro o fuera de dichos rangos, y de esta forma darla por cierta o falsa.

La hipótesis que predecimos se suele designar por H_o y se llama Hipótesis nula porque parte del supuesto que la diferencias entre el valor verdadero del parámetro y su valor hipotético es debida al azar, es decir no hay diferencia.

La hipótesis contraria se designa por H1 y se llama Hipótesis alternativa

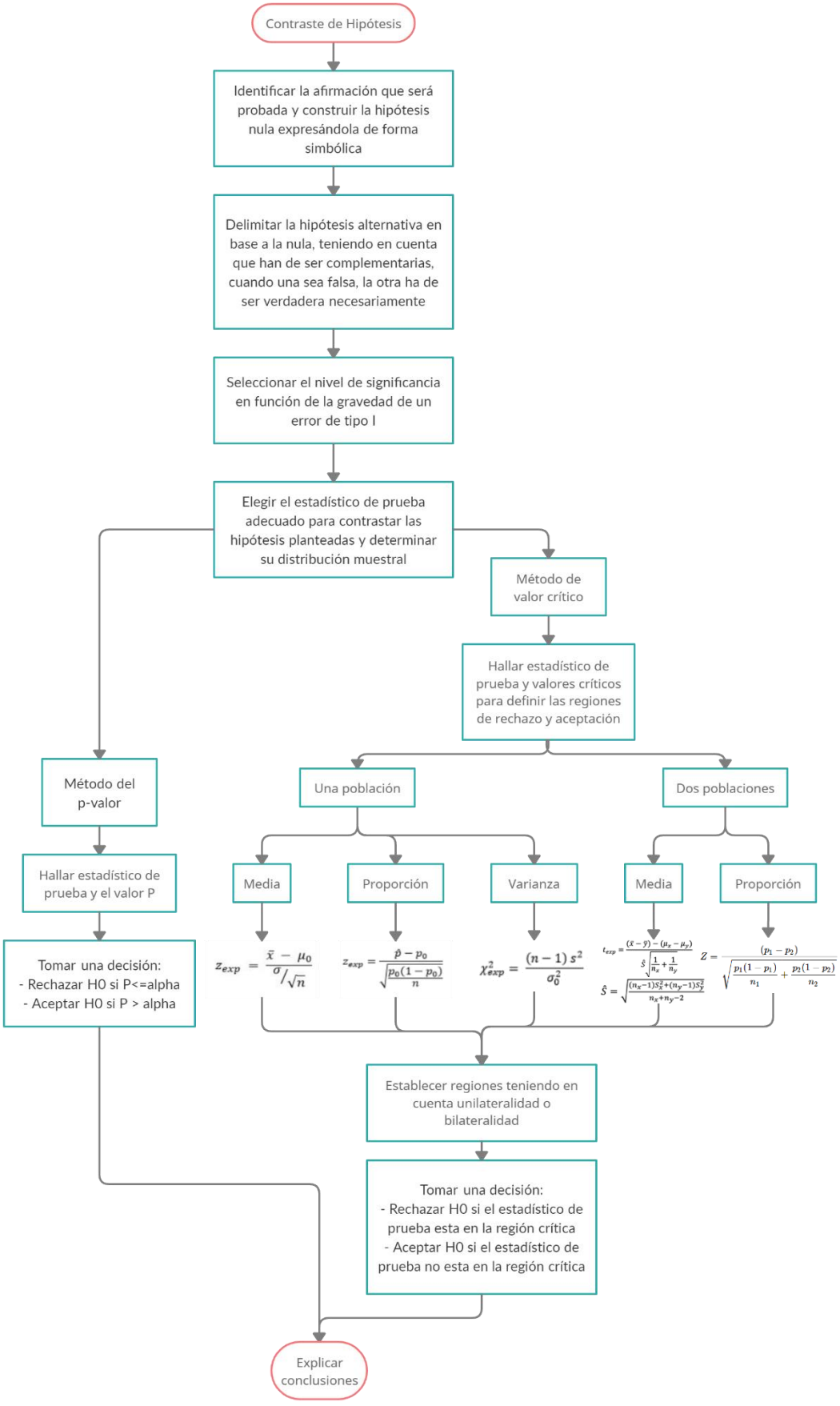
Los contrastes pueden ser unilaterales o bilaterales (también llamados de una o dos colas) según establezcamos las hipótesis, si las definimos en términos de igual y distinto estamos ante una hipótesis bilateral, si suponemos una dirección (en términos de mayor o menor) estamos ante uno unilateral.

El objetivo es hallar a partir de una muestra aleatoria, que permitan aceptar o rechazar una hipótesis predicha, sobre el valor de un parámetro desconocido de la población.

Para ello:

1. Enunciamos nuestras hipótesis, nula y alternativa, establecer si se trata de un problema bilateral o unilateral.
2. Elegir un nivel de significación y construir la zona de aceptación, intervalo fuera del cual sólo se encuentran el 100% de los casos más raros. A la zona de rechazo la llamaremos región crítica, y su área es el nivel de significación.
3. Verificar la hipótesis extrayendo una muestra cuyo tamaño se ha decidido en el paso anterior y obteniendo de ella el correspondiente estadístico (media o proporción dependiendo del caso).
4. Decidir. Si el valor calculado en la muestra cae dentro de la zona de aceptación se acepta la hipótesis y si no se rechaza.

Algoritmo General



Escenario 1

Se trata de un estudio comparativo de la tasa de abandono escolar en Madrid y Barcelona.

El tipo de información a recoger es la de individuos que afirman haber abandonado los estudios en una encuesta realizada a personas entre 16 y 24 años.

Se han recogido los datos y como resultado tenemos los siguientes totales:

	Individuos encuestados	Individuos que han abandonado
Madrid	12500	1273
Barcelona	9700	1611

Se va a realizar un contraste de hipótesis para 2 poblaciones, en los que se van a estudiar las proporciones. Y donde vamos a suponer con un nivel de confianza del 98%, que la tasa de abandono escolar en Barcelona es mayor que en Madrid.

Tenemos las siguientes hipótesis:

- H0: La proporción en la tasa de abandono escolar es igual en las 2 ciudades
- H1: La proporción en la tasa de abandono escolar en Barcelona es mayor que en Madrid.

Formulación:

- H_0: $p1 - p2 = 0$
- H_1: $p1 - p2 < 0$
- $\alpha = 0.02$

* Se trata de un contraste unilateral (izquierdo)

```r
# Importacion de librerias
library(readr)
library(dplyr)

# Importacion de los datos
Datos_abandono <- read_delim("Datos_abandono.csv",
                             ";", escape_double = FALSE, trim_ws = TRUE)

# Calculo del numero de individuos del estudio
Datos_abandono %>% filter( CCAA %in% "Madrid")->Datos_Madrid
Datos_abandono %>% filter( CCAA %in% "Barcelona")->Datos_Barcelona

n1 = nrow(Datos_Madrid)
n2 = nrow(Datos_Barcelona)

# Calculo del numero de individuos que han abandonado
Datos_Madrid %>% filter( ABANDONO %in% "Si")->Datos_Madrid_si
Datos_Barcelona %>% filter( ABANDONO %in% "Si")->Datos_Barcelona_si

x1 = nrow(Datos_Madrid_si)
x2 = nrow(Datos_Barcelona_si)

# Proporciones
P1 = x1/n1
P2 = x2/n2

# P: Proporción conjunta
P = (n1*P1+n2*P2)/(n1+n2)

# Zexp: Calificación Z
Zexp = (P1-P2)/sqrt(P*(1-P)*(1/n1+1/n2))
pvalor = pnorm(Zexp)

# Cálculo de límites
alpha = 0.02
lim_region = qnorm(alpha)
```

Data	
Datos_abandono	22200 obs. of 2 variables
Datos_Barcelona	9700 obs. of 2 variables
Datos_Barcelona_si	1611 obs. of 2 variables
Datos_Madrid	12500 obs. of 2 variables
Datos_Madrid_si	1273 obs. of 2 variables
Values	
alpha	0.02
lim_region	-2.05374891063182
n1	12500L
n2	9700L
P	0.12990990990991
P1	0.10184
P2	0.166082474226804
pvalor	1.39818513512147e-45
x1	1273L
x2	1611L
Zexp	-14.1215833408257

Tenemos que z_α = -2,053 por lo que como $z_{exp} < z_\alpha$ rechazamos por estar cayendo en la región de crítica.

Conclusión: Se rechaza la hipótesis nula con un nivel de confianza del 98%, y podemos decir que la proporción de abandono escolar en Barcelona es mayor que en Madrid.

Escenario 2

Se trata de un estudio comparativo de la edad a la que se tienen hijos en España actualmente.

El tipo de información a recoger es la edad a la que han sido padres los individuos preguntados, que han tenido hijos este último año.

Se va a realizar un contraste de hipótesis para 1 población a partir de la media de los resultados. Y donde vamos a suponer con un nivel de confianza del 99%, que la edad a la que se tienen hijos en España es de 31 Años.

- H0: $\mu = \mu 0$
- H1: $\mu \neq \mu 0$

Donde $\mu 0$ es la media = 31:

- H0: $\mu = 31$
- H1: $\mu \neq 31$
- $\alpha = 0.01$

*Se trata de un contraste bilateral

```r
# Importacion de librerias
library(readr)
library(dplyr)

# Importacion de los datos
Datos_edades <- read_delim("edades.csv",
                           ";", escape_double = FALSE, trim_ws = TRUE)

# Media oficial en España
media_oficial = 31

# Construccion de los datos
tot_muestras = nrow(Datos_edades)
x = c(Datos_edades[["Edad"]])
media = mean(x)
varianza = var(x)

# Cálculo de límites
alpha = 0.01
lim_region = abs(qnorm(alpha/2))

# Zexp: Calificación z
Zexp = (media-media_oficial)/sqrt(varianza/tot_muestras)
pvalor = pnorm(Zexp)

# Zona de aceptación
lim_abajo = media_oficial-lim_region*varianza/sqrt(tot_muestras)
lim_arriba = media_oficial+lim_region*varianza/sqrt(tot_muestras)
```

Data	
Datos_edades	1748 obs. of 2 variables
values	
alpha	0.01
lim_abajo	29.9016075815603
lim_arriba	32.0983924184397
lim_region	2.5758293035489
media	30.8489702517162
media_oficial	31
pvalor	0.0673959446687212
tot_muestras	1748L
varianza	17.8283503986566
x	num [1:1748] 27 38 27 30 36 26 33 28 33 24 ...
Zexp	-1.49546974829225

Tenemos que $z_{\alpha/2} = -2{,}575$ por lo que como $-z_{\alpha/2} < z_{exp} < z_{\alpha/2}$ podemos aceptar la hipótesis nula al encontrarse Z_{exp} en la región de aceptación.

Conclusión: Se acepta la hipótesis nula con un nivel de confianza del 99%, y podemos afirmar que la edad media de tener hijos en España es de 31 años.