Universidad de Alcalá

Estadística

Práctica 9

Contrastes de hipótesis

Universidad de Alcalá

Contenido

1. Introducción 3

2. Contraste de hipótesis para la media poblacional 5

3. Ejemplo: Contraste de hipótesis sobre la media de calificaciones de una población de estudiantes (muestra grande) 7

4. Contraste de hipótesis para proporciones poblacionales 15

5. Ejemplo: Contraste de hipótesis sobre la proporción de estudiantes que tienen móvil Android... 17

6. Contraste de hipótesis sobre la diferencia de medias 20

7. Ejemplo: Contraste de hipótesis sobre la diferencia de medias de calificaciones de estudiantes del turno de mañana y del turno de tarde (muestras grandes e independientes) 21

8. Contraste de hipótesis sobre la diferencia de proporciones 24

9. Ejemplo: Contraste de hipótesis sobre la diferencia de proporciones de estudiantes que tienen móvil Android en el turno de mañana y en el turno de tarde (muestras grandes e independientes) 25

10. Ejercicios propuestos 28

1. Introducción

Con esta práctica se utiliza R y RStudio para determinar si se puede aceptar una hipótesis, con un nivel de confianza determinado, para el valor de un parámetro de una población (media) a partir de una muestra de la misma.

En estadística hay que diferenciar entre conceptos “parámetros” y “estadísticos”. El término “parámetro” suele utilizarse en referencia a una población mientras que el término “estadístico” suele utilizarse en referencia a una muestra extraída de dicha población. En la siguiente tabla se muestra la nomenclatura utilizada habitualmente.

PARÁMETROS (POBLACIONALES)	ESTADÍSTICOS (MUESTRALES)
Media poblacional (μ)	Media muestral ($\bar{x}$)
Desviación típica o estándar poblacional (σ)	Desviación típica o estándar muestral (s)
Varianza poblacional (σ^2)	Varianza muestral (s^2)
Tamaño población (N)	Tamaño muestra (n)
Proporción poblacional (p)	Proporción muestral ($\hat{p}$)

En un contraste de hipótesis:

$$H_0: Hipótesis\ nula \quad \text{vs} \quad H_A: Hipótesis\ altenativa$$

Se manejan varios conceptos:

- **Nivel de significación del contraste (α)**: Es la probabilidad de rechazar la hipótesis H_0 cuando dicha hipótesis es cierta. Se denomina error tipo I.
- **Nivel de confianza del contraste (1-α).** Es la probabilidad de no rechazar la hipótesis H_0 cuando dicha hipótesis es cierta.
- Potencia del contraste. Es la probabilidad de rechazar la hipótesis H_0 cuando dicha hipótesis es falsa.
- Debilidad del contraste (1-potencia del contraste). Es la probabilidad de no rechazar la hipótesis H_0 cuando dicha hipótesis es falsa. Se denomina error tipo II.

Se resumen en la siguiente tabla.

	Decisión: Rechazar H_0	Decisión: No rechazar H_0	Probabilidad de acertar
Realidad: Hipótesis H_0 CIERTA	FALLO Error tipo I	**ACIERTO**	(1 – α) es la probabilidad de acertar cuando la hipótesis es cierta
Realidad: Hipótesis H_0 FALSA	**ACIERTO**	FALLO Error tipo II	(1 – potencia) es la probabilidad de acertar cuando la hipótesis es falsa

Existen dos métodos para realizar el contraste de hipótesis:

1. Comprobando la región de no rechazo
2. Calculando el p-valor

2. Contraste de hipótesis para la media poblacional

Se calcula el valor del estadístico de contraste (z_o si es muestra grande, o t_o si es muestra pequeña), después se calculan los límites de la región de aceptación, y si el valor del estadístico de contraste está:

- Dentro de la región, entonces no se puede rechazar la hipótesis H_0, por lo que se acepta y se rechaza HA.
- Fuera de la región, entonces se rechaza la hipótesis H_0.y se acepta HA.

Otra opción es calcular el p-valor y si el p-valor es:

- Mayor que α, entonces no se puede rechazar H_0.
- Menor que α, entonces se rechaza H_0.

Varianza conocida	Contraste de hipótesis	Región de aceptación y código R para calcularla	p-valor y código R para calcularlo
No $n \geq 30$	$H_0: \mu = \mu_o$ vs $H_A: \mu < \mu_o$	$(-z_\alpha, +\infty)$ `> z.alfa=qnorm(1-alfa,0,1)`	$P\{Z \leq Z_o\}$ `> p.valor=pnorm(zo,0,1)`
	$H_0: \mu = \mu_o$ vs $H_A: \mu > \mu_o$	$(-\infty, z_\alpha)$ `> z.alfa=qnorm(1-alfa,0,1)`	$P\{Z \geq Z_o\}$ `> p.valor=1-pnorm(zo,0,1)`
	$H_0: \mu = \mu_o$ vs $H_A: \mu \neq \mu_o$	$(-z_{\alpha/2}, z_{\alpha/2})$ `> z.alfa.medios=qnorm(1-alfa/2,0,1)`	$2P\{Z \geq \lvert Z_o \rvert\}$ `> p.valor=2*(1-pnorm(abs(zo),0,1))`
No $n < 30$	$H_0: \mu = \mu_o$ vs $H_A: \mu < \mu_o$	$(-t_{\alpha,\text{n-1}}, +\infty)$ `> t.alfa=qt(1-alfa,n-1)`	$P\{t \leq t_o\}$ `> p.valor=pt(to,n-1)`
	$H_0: \mu = \mu_o$ vs $H_A: \mu > \mu_o$	$(-\infty, t_{\alpha,\text{n-1}})$ `> t.alfa=qt(1-alfa,n-1)`	$P\{t \geq t_o\}$ `> p.valor=1-pt(to,n-1)`
	$H_0: \mu = \mu_o$ vs $H_A: \mu \neq \mu_o$	$(-t_{\alpha/2,\text{n-1}}, t_{\alpha/2,\text{n-1}})$ `> t.alfa.medios=qt(1-alfa/2,n-1)`	$2P\{t \geq \lvert t_o \rvert\}$ `> p.valor=2*(1-pt(abs(to),n-1))`

Cálculo del estadístico de contraste

Estadístico de contraste	Código R

$Z_o = \frac{\overline{x} - \mu_o}{\frac{s}{\sqrt{n}}}$	`x=muestra de X`
	`m=mean(x)`
	`s=sd(x)`
	`n=length(x)`
	`mo=valor a comprobar en la hipótesis`
	
	`zo=(m-mo)/(s/sqrt(n))`

$t_o = \frac{\overline{x} - \mu_o}{s/\sqrt{n}}$	`x=muestra de X`
	`m=mean(x)`
	`s=sd(x)`
	`n=length(x)`
	`mo=valor a comprobar en la hipótesis`
	
	`to=(m-mo)/(s/sqrt(n))`

En lugar de calcular las fórmulas, se pueden usar las funciones z.test y t.test de R.

Varianza conocida	Contraste de hipótesis	Código R

No
$n \geq 30$
	$H_0: \mu = \mu_o$ vs $H_A: \mu < \mu_o$	`install.packages("BSDA")`
		`library(BSDA)`
		
		`> z.test(x, alternative="less", mu=mo, sigma.x=s)`

	$H_0: \mu = \mu_o$ vs $H_A: \mu > \mu_o$	`> z.test(x, alternative="greater", mu=mo, sigma.x=s)`
	$H_0: \mu = \mu_o$ vs $H_A: \mu \neq \mu_o$	`> z.test(x, alternative="two.sided", mu=mo, sigma.x=s)`
No $n < 30$	$H_0: \mu = \mu_o$ vs $H_A: \mu < \mu_o$	`> t.test(x, alternative="less", mu=mo)`
	$H_0: \mu = \mu_o$ vs $H_A: \mu > \mu_o$	`> t.test(x, alternative="greater", mu=mo)`
	$H_0: \mu = \mu_o$ vs $H_A: \mu \neq \mu_o$	`> t.test(x, alternative="two.sided", mu=mo)`

3. Ejemplo: Contraste de hipótesis sobre la media de calificaciones de una población de estudiantes (muestra grande)

ENUNCIADO

Mediante una encuesta, se saben las calificaciones de acceso a la universidad de 74 estudiantes de la asignatura Estadística del Grado en Ingeniería en Sistemas de Información de la Universidad de Alcalá del curso 2021-22, de una población de 108 matriculados. Si se supone que las calificaciones tienen una distribución Normal, responder a las siguientes preguntas:

a) **¿Se puede afirmar con una confianza del 95% que la media de notas de acceso de toda la población es diferente a 8.6?**
   1. **Resolverlo comprobando la región de aceptación.**
   2. **Resolverlo calculando el p-valor aplicando fórmulas**
   3. **Resolverlo usando la función** `z.test()` **del paquete BSDA de R**

b) **¿Se puede afirmar con una confianza del 95% que la media de notas de acceso de toda la población de alumnos matriculados en la asignatura es inferior a 8.6?**
   1. **Resolverlo comprobando la región de aceptación.**
   2. **Resolverlo calculando el p-valor aplicando fórmulas**
   3. **Resolverlo usando la función** `z.test()` **del paquete BSDA de R**

c) **¿Se puede afirmar con una confianza del 95% que la media de notas de acceso de toda la población de alumnos matriculados en la asignatura es superior a 8.6?**
   1. **Resolverlo comprobando la región de aceptación**
   2. **Resolverlo calculando el p-valor aplicando fórmulas**
   3. **Resolverlo usando la función** `z.test()` **del paquete BSDA de R**

SOLUCIÓN

Primero hay que leer los datos de las notas, disponibles en el archivo encuesta.csv:

```
> encuesta = read.csv2("encuesta.csv")
> (nota=encuesta$NOTA)
 [1]  8.500  7.100  8.635  8.624  8.200  8.700  7.210  7.630  8.400  8.300
[11]  8.200  9.100  9.789 10.107  8.015  7.310  7.500  8.710  8.336  9.210
[21]  7.800 10.300  7.990  6.900  7.800 10.000  8.590  7.000  8.050 10.799
[31]  7.994  8.550  7.340  6.750  9.560  7.417  6.936  7.210  7.680 10.277
[41]  7.860 10.260  7.270  5.800  7.300  7.140  8.600  7.500  8.000  7.540
[51]  7.292  7.830  6.750  9.806  6.800  6.445  6.650  7.804 10.270  7.600
[61]  7.870  7.000  7.085  7.480  8.070  5.820  6.500  9.900  7.500  6.500
[71]  9.456  8.000  7.800  7.654
```

No se conoce la varianza poblacional y es una muestra grande (74 >30). Al ser una muestra grande no hay que comprobar la normalidad.

Primero se crean las variables alfa, n, s, m y mo.

```
> (alfa=1-0.95)
```

```
[1] 0.05
> (n=length(nota))
[1] 74
> (s=sd(nota))
[1] 1.131097
> (m=mean(nota))
[1] 8.022581
> (mo=8.6)
[1] 8.6
```

a) ¿Se puede afirmar con una confianza del 95% que la media de notas de acceso de toda la población es diferente a 8.6?

Se formula el siguiente contraste de hipótesis:

$$H_0: \mu = 8.6 \text{ vs } H_A: \mu \neq 8.6$$

a.1) Resolverlo comprobando la región de aceptación.

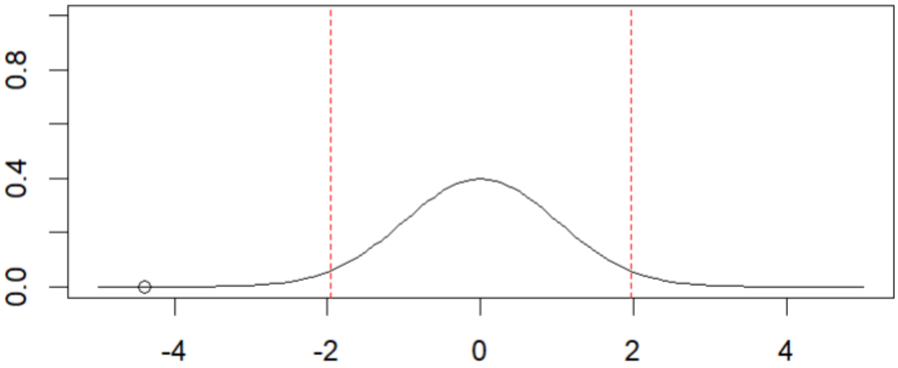
La región de aceptación es $\left(-z_{\alpha/2}, z_{\alpha/2}\right)$.

```
> (zo=(m-mo)/(s/sqrt(n)))
[1] -4.391442
> (z.alfa.medios=qnorm(1-alfa/2,0,1))
[1] 1.959964
> (-z.alfa.medios<=zo)&(zo<=z.alfa.medios)
[1] FALSE
```

Como zo (-4.39) no está en la región de aceptación $(-1.96,1.96)$, entonces se rechaza la hipótesis nula (H0) y se acepta la alternativa (HA), por lo que la respuesta es **SÍ se puede afirmar con una confianza del 95% que la media poblacional es diferente a 8.6**.

Podemos hacer la comprobación visualmente.

```
> curve(dnorm(x,0,1),-5,5, add = TRUE)
> abline(v=-z.alfa.medios, col="red", lty=2)
> abline(v=z.alfa.medios, col="red", lty=2)
```



En el diagrama, el área bajo la función entre las dos líneas es 1-alfa (0.95), el área desde la primera línea hacia la izquierda es alfa/2 (0.025), y el área desde la segunda línea hacia la derecha es alfa/2 (0.025). El área total es 1.

La región de aceptación está limitada por las dos líneas rojas, y zo es el punto dibujado. Se comprueba visualmente que zo está fuera de la región.

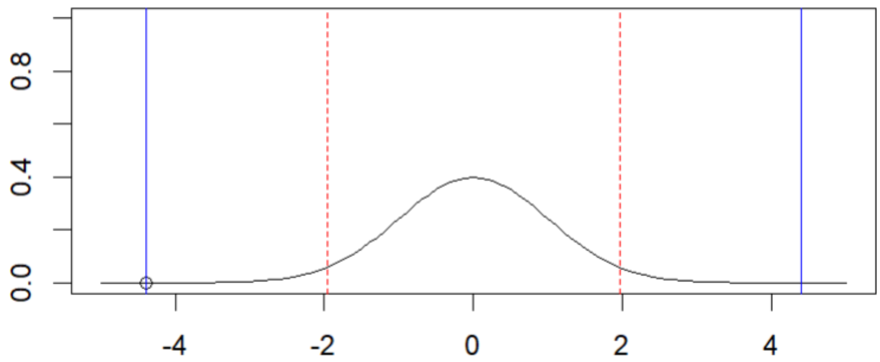
a.2) Resolverlo calculando el p-valor aplicando fórmulas

```
> (p.valor=2*(1-pnorm(abs(zo),0,1)))
[1] 0.00001126015
```

Como el p-valor (0.00001126) es menor que alfa (0.05), como ocurrió en el apartado anterior, se rechaza la hipótesis nula y se acepta la alternativa de que la media poblacional es diferente a 8.6.

En el mismo diagrama del apartado anterior podemos dibujar una línea vertical azul continua en el punto zo y otra en -zo.

```
> abline(v=zo, col="blue")
> abline(v=-zo, col="blue")
```



el área desde la primera línea roja hacia la izquierda es alfa/2 (0.025), y el área desde la segunda línea hacia la derecha es alfa/2 (0.025). El área desde la primera línea azul continua hacia la izquierda es p-valor/2 y el área desde la segunda línea azul continua hacia la derecha es p-valor/2, y la suma de ambas es p-valor. Se observa visualmente que p-valor < alfa.

a.3) Resolverlo usando la función `z.test()` del paquete BSDA de R

```
> z.test(nota, alternative="two.sided", mu=mo, sigma.x=s, conf.level=0.95)
      One-sample z-Test
data:  nota
z = -4.3914, p-value = 0.00001126
alternative hypothesis: true mean is not equal to 8.6
95 percent confidence interval:
 7.764871 8.280292
sample estimates:
mean of x
 8.022581
```

El p-valor es el mismo que en apartado anterior. Como el p-valor es menor que alfa (0.05), se rechaza la hipótesis nula y se acepta la alternativa de que la media poblacional es diferente a 8.6.

El resultado de la función `z.test()` incluye más información:

- Indica que el valor del estadístico de contraste zo (llamado z en el resultado de la función) es -4.3914, que coincide con los apartados anteriores.
- Indica que la hipótesis alternativa es que la media fuera diferente a 8.6, como así se indicó al llamar al a la función con el parámetro `alternative="two.sided"`
- Indica que el nivel de confianza es del 95%, como así se indicó al llamar a la función con el parámetro `conf.level=0.95`.
- Indica que el intervalo de confianza para la media poblacional con un nivel del 95% es (7.764871, 8.280292). Puede comprobarse que 8.6 está fuera del intervalo.
- Indica que la media muestral es 8.022581.

b) ¿Se puede afirmar con una confianza del 95% que la media de notas de acceso de toda la población de alumnos matriculados en la asignatura es inferior a 8.6?

Se formula el siguiente contraste de hipótesis:

$$H_0: \mu = 8.6 \text{ vs } H_A: \mu < 8.6$$

b.1) Resolverlo comprobando la región de aceptación

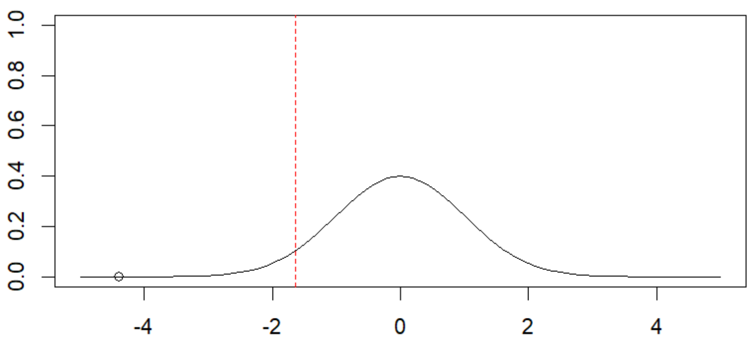
La región de aceptación es $(-z_\alpha, +\infty)$.

```
> (zo=(m-mo)/(s/sqrt(n)))
[1] -4.391442
> (z.alfa=qnorm(1-alfa,0,1))
[1] 1.644854
> (zo>-z.alfa) #Como es FALSE se acepta la alternativa. Respuesta SI
[1] FALSE
```

Como zo (-4.39) está fuera de la región de aceptación $(-1.64, +\infty)$, entonces se rechaza la hipótesis nula (H0), y se acepta la hipótesis alternativa (HA), por lo que la respuesta es **SÍ se puede afirmar con una confianza del 95% que la media poblacional es menor a 8.6**.

Podemos hacer la comprobación visualmente, dibujando la curva de la función de densidad de probabilidad normal Z: N(0,1) para la variable estadístico de contraste, señalando un punto para el valor calculado para el estadístico (zo), y una línea vertical roja con trazo discontinuo, en el extremo inferior de la región de aceptación (-z.alfa), que no tiene límite superior, pues sería -infinito. Puede comprobarse visualmente que el punto zo está fuera de la región de aceptación.

```
> plot(zo,0, xlim = c(-5,5), ylim=c(0,1))
> curve(dnorm(x,0,1),-5,5, add = TRUE)
> abline(v=-z.alfa, col="red", lty=2)
```



En el diagrama, el área bajo la función desde la línea roja hacia la izquierda es alfa (0.05), mientras que el área hacia la derecha es 1-alfa (0.95). La región de aceptación es desde la línea roja hasta $+\infty$, y zo es el punto dibujado. Se comprueba visualmente que zo está fuera de la región.

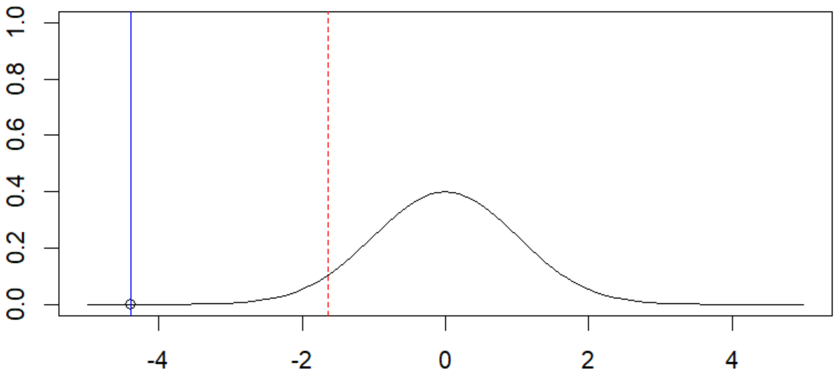
b.2) Resolverlo calculando el p-valor aplicando fórmulas

```
> (p.valor=pnorm(zo,0,1))
[1] 0.000005630077
```

Como el p-valor (0.00000563) es menor que alfa (0.05), se rechaza la hipótesis H0, como en el apartado anterior.

En el mismo diagrama del apartado anterior podemos dibujar una línea vertical azul continua en el punto zo.

```
> abline(v=zo, col="blue")
```



En el diagrama, el área bajo la función desde la línea roja discontinua hacia la izquierda es alfa (0.05), mientras que el área desde la línea azul continua hacia la izquierda es el p-valor (0.00000563). Se observa visualmente que p-valor < alfa.

b.3) Resolverlo usando la función `z.test()` del paquete BSDA de R

```
> z.test(nota, alternative="less", mu=mo, sigma.x=s, conf.level=0.95)
      One-sample z-Test
data:  nota
z = -4.3914, p-value = 0.00000563
alternative hypothesis: true mean is less than 8.6
95 percent confidence interval:
       NA 8.238858
sample estimates:
mean of x
 8.022581
```

El p-valor es el mismo que en el apartado anterior.

c) ¿Se puede afirmar con una confianza del 95% que la media de notas de acceso de toda la población es superior a 8.6?

Se formula el siguiente contraste de hipótesis:

$$H_0: \mu = 8.6 \text{ vs } H_A: \mu > 8.6$$

c.1) Resolverlo comprobando la región de aceptación

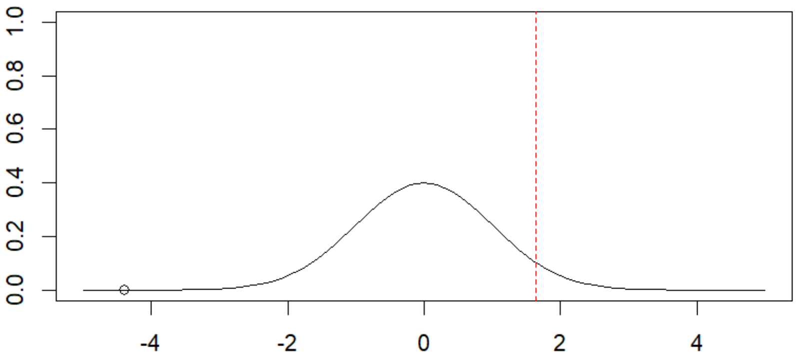
La región de aceptación es $(-\infty, z_\alpha)$.

```
> (zo=(m-mo)/(s/sqrt(n)))
[1] -4.391442
> (z.alfa=qnorm(1-alfa,0,1))
[1] 1.644854
> (zo<z.alfa)
[1] TRUE
```

Como zo (-4.39) está en la región de aceptación $(-\infty, 1.64)$, entonces no se rechaza la hipótesis nula, y se rechaza la hipótesis alternativa, por lo que la respuesta es que **NO se puede afirmar con una confianza del 95% que la media poblacional sea superior a 8.6**.

Podemos hacer la comprobación visualmente.

```
> plot(zo,0, xlim = c(-5,5), ylim=c(0,1))
> curve(dnorm(x,0,1),-5,5, add = TRUE)
> abline(v=z.alfa, col="red", lty=2)
```



En el diagrama, el área bajo la función desde la línea hacia la derecha es alfa (0.05), mientras que el área hacia la izquierda es 1-alfa (0.95). La región de aceptación es desde $-\infty$ hasta la línea roja, y zo es el punto dibujado. Se comprueba visualmente que zo está dentro de la región

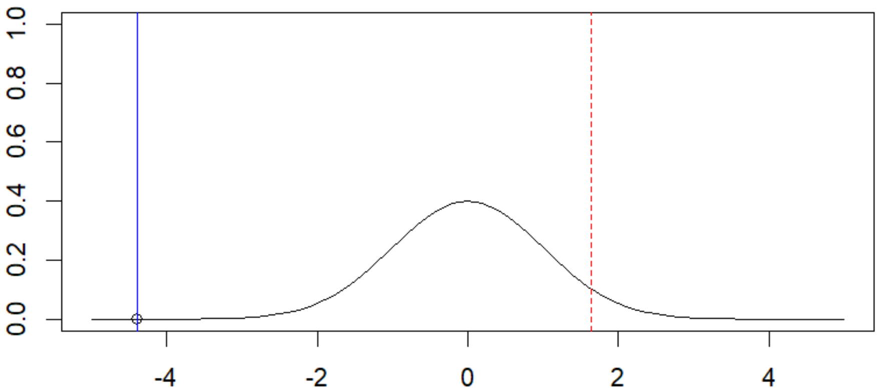
c.2) Resolverlo calculando el p-valor aplicando fórmulas

```
> (p.valor=1-pnorm(zo,0,1))
[1] 0.9999944
```

Como el p-valor (0.9999944) es mayor que alfa (0.05), no se rechaza la hipótesis nula, y se rechaza la hipótesis alternativa de que la media poblacional es mayor que 8.6.

En el mismo diagrama del apartado anterior podemos dibujar una línea vertical azul continua en el punto zo.

```
> abline(v=zo, col="blue")
```



En el diagrama, el área bajo la función desde la línea roja discontinua hacia la derecha es alfa (0.05), mientras que el área desde la línea azul continua hacia la derecha es p-valor (0.999). Se observa visualmente que p-valor > alfa.

c.3) Resolverlo usando la función `z.test()` del paquete BSDA de R

```
> z.test(nota, alternative="greater", mu=mo, sigma.x=s, conf.level=0.95)
      One-sample z-Test
data:  nota
z = -4.3914, p-value = 1
alternative hypothesis: true mean is greater than 8.6
95 percent confidence interval:
 7.806304       NA
sample estimates:
mean of x
 8.022581
```

El p-valor es el mismo que en apartado anterior (0.9999944), pero está redondeado a 1.

4. Contraste de hipótesis para proporciones poblacionales

Se calcula el valor del estadístico de contraste (z_o), después se calculan los límites de la región de aceptación, y si el valor del estadístico de contraste está:

- Dentro de la región, entonces no se puede rechazar la hipótesis H_0, por lo que se acepta y se rechaza HA.
- Fuera de la región, entonces se rechaza la hipótesis H_0.y se acepta H_A.

Otra opción es calcular el p-valor y si el p-valor es:

- Mayor que α, entonces no se puede rechazar H_0.
- Menor que α, entonces se rechaza H_0.

Contraste de hipótesis	Región de aceptación y código R para calcularla	p-valor y código R para calcularlo
$H_0: p = p_o$ vs $H_A: p < p_o$	$(-z_\alpha, +\infty)$ `> z.alfa=qnorm(1-alfa,0,1)`	$P\{Z \leq Z_o\}$ `> p.valor=pnorm(zo,0,1)`
$H_0: p = p_o$ vs $H_A: p > p_o$	$(-\infty, z_\alpha)$ `> z.alfa=qnorm(1-alfa,0,1)`	$P\{Z \geq Z_o\}$ `> p.valor=1-pnorm(zo,0,1)`
$H_0: p = p_o$ vs $H_A: p \neq p_o$	$(-z_{\alpha/2}, z_{\alpha/2})$ `> z.alfa.medios=qnorm(1-alfa/2,0,1)`	$2P\{Z \geq \|Z_o\|\}$ `> p.valor=2*(1-pnorm(abs(zo),0,1))`

Cálculo del estadístico de contraste.

Estadístico de contraste	Código R

$Z_o = \frac{\hat{p} - p_o}{\sqrt{\frac{p_o(1-p_o)}{n}}}$	`x=muestra de X`
	`n=length(x)`
	`p=proporción muestral`
	`po=valor a comprobar en la hipótesis`
	
	`> zo=(p-po)/sqrt(po*(1-po)/n)`

En lugar de calcular la fórmula, se puede usar la función prop.test de R.

Contraste de hipótesis	Código R
$H_0: p = p_o$ vs $H_A: p < p_o$	`> prop.test(ng, n, p=po, alternative="less", conf.level=0.95, correct=FALSE)`
$H_0: p = p_o$ vs $H_A: p > p_o$	`> prop.test(ng, n, p=po, alternative="greater", conf.level=0.95, correct=FALSE)`
$H_0: p = p_o$ vs $H_A: p \neq p_o$	`> prop.test(ng, n, p=po, alternative="two.sided", conf.level=0.95, correct=FALSE)`

Donde `n` es el tamaño de la muestra, y `ng` es el tamaño del grupo de datos de la muestra para el que se calcula la proporción, es decir la proporción muestral sería `ng/n`.

Universidad de Alcalá

5. Ejemplo: Contraste de hipótesis sobre la proporción de estudiantes que tienen móvil Android

ENUNCIADO

Mediante una encuesta, se sabe que 50 de una muestra de 74 estudiantes de la asignatura Estadística del Grado en Ingeniería en Sistemas de Información de la Universidad de Alcalá del curso 2021-22 tienen teléfonos móviles con sistema operativo Android. La población es de 108 estudiantes matriculados. Responder a la siguiente pregunta:

¿Se puede afirmar con una confianza del 95% que más de la mitad de los alumnos matriculados en la asignatura tienen teléfono Android?

a) **Resolverlo comprobando la región de aceptación**
b) **Resolverlo calculando el p-valor aplicando fórmulas**
c) **Resolverlo usando la función `prop.test()`**

SOLUCIÓN

Se formula el siguiente contraste de hipótesis, ya que la mitad es el 50%, es decir una proporción de 0.5:

$$H_0: p = 0.5 \text{ vs } H_A: p > 0.5$$

a) Resolverlo comprobando la región de aceptación

La región de aceptación es $(-\infty, z_\alpha)$.

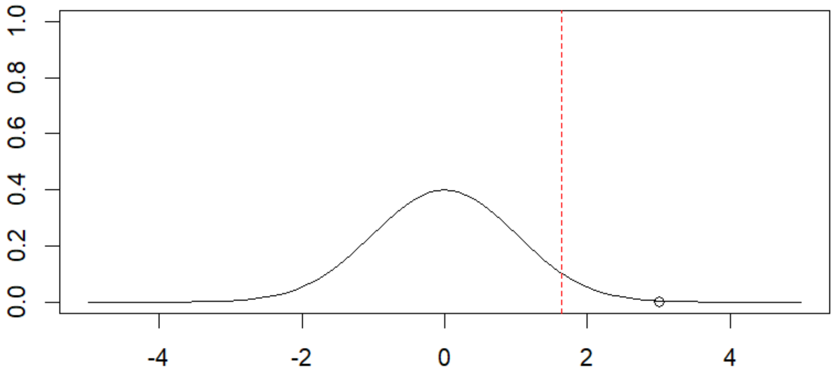
```
> so=encuesta$SO
> (n=length(so))
[1] 74
> so.android=so[so=="Android"]
> (ng=length(so.android))
[1] 50
> (p=ng/n)
[1] 0.6756757
> po=0.5
> alfa=0.05

> (zo=(p-po)/sqrt(po*(1-po)/n))
[1] 3.022439

> (z.alfa=qnorm(1-alfa,0,1))
[1] 1.644854
> (zo<z.alfa)
[1] FALSE
```

Como zo (3.02) está fuera de la región de aceptación $(-\infty, 1.64)$, entonces se rechaza la hipótesis nula (H0), y se acepta la hipótesis alternativa (HA), por lo que la respuesta es **SÍ se puede afirmar**

con una confianza del 95% que la proporción poblacional de alumnos con Android es superior al 50%.

Podemos hacer la comprobación visualmente.

```
> plot(zo,0, xlim = c(-5,5), ylim=c(0,1))
> curve(dnorm(x,0,1),-5,5, add = TRUE)
> abline(v=z.alfa, col="red", lty=2)
```



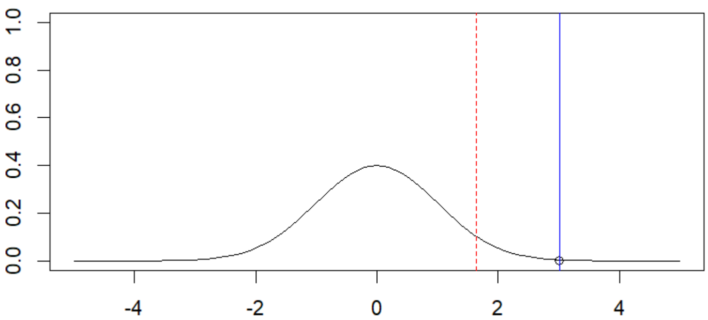
En el diagrama, el área bajo la función desde la línea hacia la izquierda es 1-alfa (0.95), mientras que el área hacia la derecha es alfa (0.055). La región de aceptación es desde $-\infty$ hasta la línea roja, y zo es el punto dibujado. Se comprueba visualmente que zo está fuera de la región.

b) Resolverlo calculando el p-valor aplicando fórmulas

```
> (p.valor=1-pnorm(zo,0,1))
[1] 0.001253735
```

Como el p-valor (0.001) es menor que alfa (0.05), se rechaza la hipótesis H0, como en el apartado anterior.

En el mismo diagrama del apartado anterior podemos dibujar una línea vertical azul continua en el punto zo.



En el diagrama, el área bajo la función desde la línea roja discontinua hacia la derecha es alfa (0.05), mientras que el área desde la línea azul continua hacia la derecha es el p-valor (0.001). Se observa visualmente que p-valor < alfa.

c) Resolverlo usando la función `prop.test()`

```
> prop.test(ng, n, p=po, alternative="greater", conf.level=0.95, correct=FALSE)
      1-sample proportions test without continuity correction
data:  ng out of n, null probability po
X-squared = 9.1351, df = 1, p-value = 0.001254
alternative hypothesis: true p is greater than 0.5
95 percent confidence interval:
 0.5813442 1.0000000
sample estimates:
        p
0.6756757
```

El p-valor es el mismo que en el apartado anterior.

6. Contraste de hipótesis sobre la diferencia de medias

Por simplicidad se utilizará sólo el método del p-valor usando las funciones z.test y t.test de R.

Si x e y son dos vectores con las muestras de las poblaciones X e Y independientes, entonces se utilizarán las funciones indicadas en la siguiente tabla.

Varianzas conocidas	Contraste de hipótesis	Función R
No $n, m \geq 30$	$H_0: \mu_X = \mu_Y$ vs $H_A: \mu_X < \mu_Y$	`z.test(x, y, alternative="less", sigma.x=sd(x), sigma.y=sd(y), conf.level=valor)`
	$H_0: \mu_X = \mu_Y$ vs $H_A: \mu_X > \mu_Y$	`z.test(x, y, alternative="greater", sigma.x=sd(x), sigma.y=sd(y), conf.level=valor)`
	$H_0: \mu_X = \mu_Y$ vs $H_A: \mu_X \neq \mu_Y$	`z.test(x, y, alternative="two.sided", sigma.x=sd(x), sigma.y=sd(y), conf.level=valor)`
No $n, m < 30$ $\sigma_X \neq \sigma_Y$	$H_0: \mu_X = \mu_Y$ vs $H_A: \mu_X < \mu_Y$	`t.test(x, y, alternative="less", conf.level=valor, var.equal=FALSE)`
	$H_0: \mu_X = \mu_Y$ vs $H_A: \mu_X > \mu_Y$	`t.test(x, y, alternative="greater", conf.level=valor, var.equal=FALSE)`
	$H_0: \mu_X = \mu_Y$ vs $H_A: \mu_X \neq \mu_Y$	`t.test(x, y, alternative="two.sided", conf.level=valor, var.equal=FALSE)`
No $n, m < 30$ $\sigma_X = \sigma_Y$	$H_0: \mu_X = \mu_Y$ vs $H_A: \mu_X < \mu_Y$	`t.test(x, y, alternative="less", conf.level=valor, var.equal=TRUE)`
	$H_0: \mu_X = \mu_Y$ vs $H_A: \mu_X > \mu_Y$	`t.test(x, y, alternative="greater", conf.level=valor, var.equal=TRUE)`
	$H_0: \mu_X = \mu_Y$ vs $H_A: \mu_X \neq \mu_Y$	`t.test(x, y, alternative="two.sided", conf.level=valor, var.equal=TRUE)`

7. Ejemplo: Contraste de hipótesis sobre la diferencia de medias de calificaciones de estudiantes del turno de mañana y del turno de tarde (muestras grandes e independientes)

ENUNCIADO

Mediante una encuesta, se saben las calificaciones de acceso a la universidad de 42 estudiantes del turno de mañana y de 32 estudiantes del turno de tarde de la asignatura Estadística del Grado en Ingeniería en Sistemas de Información de la Universidad de Alcalá del curso 2021-22, de una población de 52 matriculados en el turno de mañana y 56 en el turno de tarde. Si se supone que las calificaciones tienen una distribución Normal, responder a la siguiente pregunta:

a) **¿Se puede afirmar con una confianza del 95% que la media de notas de acceso de los alumnos de mañana es diferente a la media de notas de acceso de los alumnos de tarde?**
b) **¿Se puede afirmar con una confianza del 95% que la media de notas de acceso de los alumnos de mañana es mayor que la media de notas de acceso de los alumnos de tarde?**
c) **¿Se puede afirmar con una confianza del 95% que la media de notas de acceso de los alumnos de mañana es menor que la media de notas de acceso de los alumnos de tarde?**

SOLUCIÓN

Primero hay que crear los vectores con los datos de las dos muestras.

```
> encuesta=read.csv2("encuesta.csv")

> encuesta.mañana=encuesta[(encuesta$GRUPO=="A1")|(encuesta$GRUPO=="A2"),]
> encuesta.tarde=encuesta[(encuesta$GRUPO=="B1")|(encuesta$GRUPO=="B2"),]
> notaM=encuesta.mañana$NOTA
> notaT=encuesta.tarde$NOTA
```

a) ¿Se puede afirmar con una confianza del 95% que la media de notas de acceso de los alumnos de mañana es diferente a la media de notas de acceso de los alumnos de tarde?

Se formula el contraste:

$$H_0: \mu_{Mañana} = \mu_{Tarde} \text{ vs } H_A: \mu_{Mañana} \neq \mu_{Tarde}$$

```
> z.test(notaM, notaT, alternative="two.sided", sigma.x=sd(notaM),
sigma.y=sd(notaT), conf.level=0.95)

      Two-sample z-Test
data:  notaM and notaT
z = 2.9876, p-value = 0.002812
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.2592794 1.2483159
sample estimates:
mean of x mean of y
 8.348548  7.594750
```

Como el p-valor (0.0028) es menor que alfa (0.05), se rechaza la Hipótesis H0 y se acepta HA, por lo que **SÍ se puede afirmar con una confianza del 95% que la media de notas de acceso de los alumnos de mañana es diferente a la media de notas de acceso de los alumnos de tarde**.

b) ¿Se puede afirmar con una confianza del 95% que la media de notas de acceso de los alumnos de mañana es mayor que la media de notas de acceso de los alumnos de tarde?

Se formula el contraste:

$$H_0\colon \mu_{Mañana} = \mu_{Tarde} \text{ vs } H_A\colon \mu_{Mañana} > \mu_{Tarde}$$

```
> z.test(notaM, notaT, alternative="greater", sigma.x=sd(notaM),
sigma.y=sd(notaT), conf.level=0.95)

      Two-sample z-Test
data:  notaM and notaT
z = 2.9876, p-value = 0.001406
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 0.3387848        NA
sample estimates:
mean of x mean of y
 8.348548  7.594750
```

Como el p-valor (0.0014) es menor que alfa (0.05), se rechaza la Hipótesis H0 y se acepta HA, por lo que **SÍ se puede afirmar con una confianza del 95% que la media de notas de acceso de los alumnos de mañana es mayor que la media de notas de acceso de los alumnos de tarde**.

c) ¿Se puede afirmar con una confianza del 95% que la media de notas de acceso de los alumnos de mañana es menor que la media de notas de acceso de los alumnos de tarde?

Se formula el contraste:

$$H_0: \mu_{Mañana} = \mu_{Tarde} \text{ vs } H_A: \mu_{Mañana} < \mu_{Tarde}$$

```
> z.test(notaM, notaT, alternative="less", sigma.x=sd(notaM), sigma.y=sd(notaT), conf.level=0.95)

      Two-sample z-Test
data:  notaM and notaT
z = 2.9876, p-value = 0.9986
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
      NA 1.16881
sample estimates:
mean of x mean of y
 8.348548  7.594750
```

Como el p-valor (0.9986) es mayor que alfa (0.05), no se puede rechazar la Hipótesis H0 y, por tanto, se rechaza HA, por lo que **NO se puede afirmar con una confianza del 95% que la media de notas de acceso de los alumnos de mañana es menor que la media de notas de acceso de los alumnos de tarde**.

Universidad de Alcalá

8. Contraste de hipótesis sobre la diferencia de proporciones

Por simplicidad se utilizará sólo el método del p-valor usando la función prop.test de R.

Si x e y son dos vectores con las muestras de las poblaciones X e Y independientes, entonces se utilizarán las funciones indicadas en la siguiente tabla.

Contraste de hipótesis	Función R
$H_0: p_1 = p_2$ vs $H_A: p_1 < p_2$	`prop.test(c(ng1,ng2), c(n1,n2), alternative="less", conf.level=valor, correct=FALSE)`
$H_0: p_1 = p_2$ vs $H_A: p_1 > p_2$	`prop.test(c(ng1,ng2), c(n1,n2), alternative="greater", conf.level=valor, correct=FALSE)`
$H_0: p_1 = p_2$ vs $H_A: p_1 \neq p_2$	`prop.test(c(ng1,ng2), c(n1,n2), alternative="two.sided", conf.level=valor, correct=FALSE)`

`n1 y n2` son los tamaños de las muestras, y `ng1 y ng2` son los tamaños de los grupos de datos de las muestras para los que se calculan las proporciones, es decir, las proporciones muestrales serían `p1=ng1/n1 y p2=ng2/n2.`

Universidad de Alcalá

9. Ejemplo: Contraste de hipótesis sobre la diferencia de proporciones de estudiantes que tienen móvil Android en el turno de mañana y en el turno de tarde (muestras grandes e independientes)

ENUNCIADO

Mediante una encuesta se sabe que, en una muestra de 42 estudiantes del turno de mañana de la asignatura Estadística, 27 tienen teléfonos móviles con sistema operativo Android, y que, en una muestra de 32 estudiantes del turno de tarde, 23 tienen también teléfono Android. Responder a las siguientes preguntas:

a) **¿Se puede afirmar con una confianza del 95% que la proporción de alumnos con Android en el turno de mañana es diferente a la proporción de alumnos con Android en el turno de tarde?**
b) **¿Se puede afirmar con una confianza del 95% que la proporción de alumnos con Android en el turno de mañana es mayor que la proporción de alumnos con Android en el turno de tarde?**
c) **¿Se puede afirmar con una confianza del 95% que la proporción de alumnos con Android en el turno de mañana es menor que la proporción de alumnos con Android en el turno de tarde?**

SOLUCIÓN

Primero creamos variables con el número de alumnos total y los que tiene Android en cada turno:

```
> soM=encuesta.mañana$SO
> soT=encuesta.tarde$SO

> (nM=length(soM))
[1] 42
> (nT=length(soT))
[1] 32

> (nMA=length(soM[soM=="Android"]))
[1] 27
> (nTA=length(soT[soT=="Android"]))
[1] 23
```

a) ¿Se puede afirmar con una confianza del 95% que la proporción de alumnos con Android en el turno de mañana es diferente a la proporción de alumnos con Android en el turno de tarde?

Se formula el contraste:

$$H_0: p_{Mañana} = p_{Tarde} \text{ vs } H_A: p_{Mañana} \neq p_{Tarde}$$

```
> prop.test(c(nMA,nTA), c(nM,nT), alternative="two.sided", conf.level=0.95,
correct=FALSE)

      2-sample test for equality of proportions without continuity correction
data:  c(nMA, nTA) out of c(nM, nT)
X-squared = 0.47737, df = 1, p-value = 0.4896
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.2886515  0.1368658
sample estimates:
   prop 1    prop 2
0.6428571 0.7187500
```

Como el p-valor (0.4896) es mayor que alfa (0.05), no se puede rechazar la Hipótesis H0 y, por tanto se rechaza HA, por lo que **NO se puede afirmar con una confianza del 95% que la proporción de alumnos con Android en el turno de mañana es diferente a la proporción de alumnos con Android en el turno de tarde**.

b) ¿Se puede afirmar con una confianza del 95% que la proporción de alumnos con Android en el turno de mañana es mayor que la proporción de alumnos con Android en el turno de tarde?

Se formula el contraste:

$$H_0: p_{Mañana} = p_{Tarde} \text{ vs } H_A: p_{Mañana} > p_{Tarde}$$

```
> prop.test(c(nMA,nTA), c(nM,nT), alternative="greater", conf.level=0.95,
correct=FALSE)

      2-sample test for equality of proportions without continuity correction
data:  c(nMA, nTA) out of c(nM, nT)
X-squared = 0.47737, df = 1, p-value = 0.7552
alternative hypothesis: greater
95 percent confidence interval:
 -0.2544456  1.0000000
sample estimates:
   prop 1    prop 2
0.6428571 0.7187500
```

Como el p-valor (0.7552) es mayor que alfa (0.05), no se puede rechazar la Hipótesis H0 y, por tanto se rechaza HA, por lo que **NO se puede afirmar con una confianza del 95% que la proporción de alumnos con Android en el turno de mañana es mayor que la proporción de alumnos con Android en el turno de tarde**.

Universidad de Alcalá

c) ¿Se puede afirmar con una confianza del 95% que la proporción de alumnos con Android en el turno de mañana es menor que la proporción de alumnos con Android en el turno de tarde?

Se formula el contraste:

$$H_0: p_{Mañana} = p_{Tarde} \text{ vs } H_A: p_{Mañana} < p_{Tarde}$$

```
> prop.test(c(nMA,nTA), c(nM,nT), alternative="less", conf.level=0.95,
correct=FALSE)

      2-sample test for equality of proportions without continuity correction
data:  c(nMA, nTA) out of c(nM, nT)
X-squared = 0.47737, df = 1, p-value = 0.2448
alternative hypothesis: less
95 percent confidence interval:
 -1.0000000  0.1026599
sample estimates:
   prop 1    prop 2
0.6428571 0.7187500
```

Como el p-valor (0.2448) es mayor que alfa (0.05), no se puede rechazar la Hipótesis H0 y, por tanto se rechaza HA, por lo que **NO se puede afirmar con una confianza del 95% que la proporción de alumnos con Android en el turno de mañana es menor que la proporción de alumnos con Android en el turno de tarde**.

10. Ejercicios propuestos

1) Mediante una encuesta, se sabe el tiempo en minutos del viaje a la Escuela Politécnica de 74 estudiantes de la asignatura Estadística del Grado en Ingeniería en Sistemas de Información de la Universidad de Alcalá del curso 2021-22, de una población de 108 matriculados. Responder a las siguientes preguntas:
    a. ¿Se puede afirmar con una confianza del 95% que la media del tiempo de viaje de toda la población es inferior o igual a 50 minutos?
        1. Resolverlo comprobando la región de no rechazo.
        2. Resolverlo calculando el p-valor aplicando fórmulas
        3. Resolverlo calculando el p-valor usando la función z.test() del paquete BSDA de R.