BIOESTADÍSTICA BÁSICA

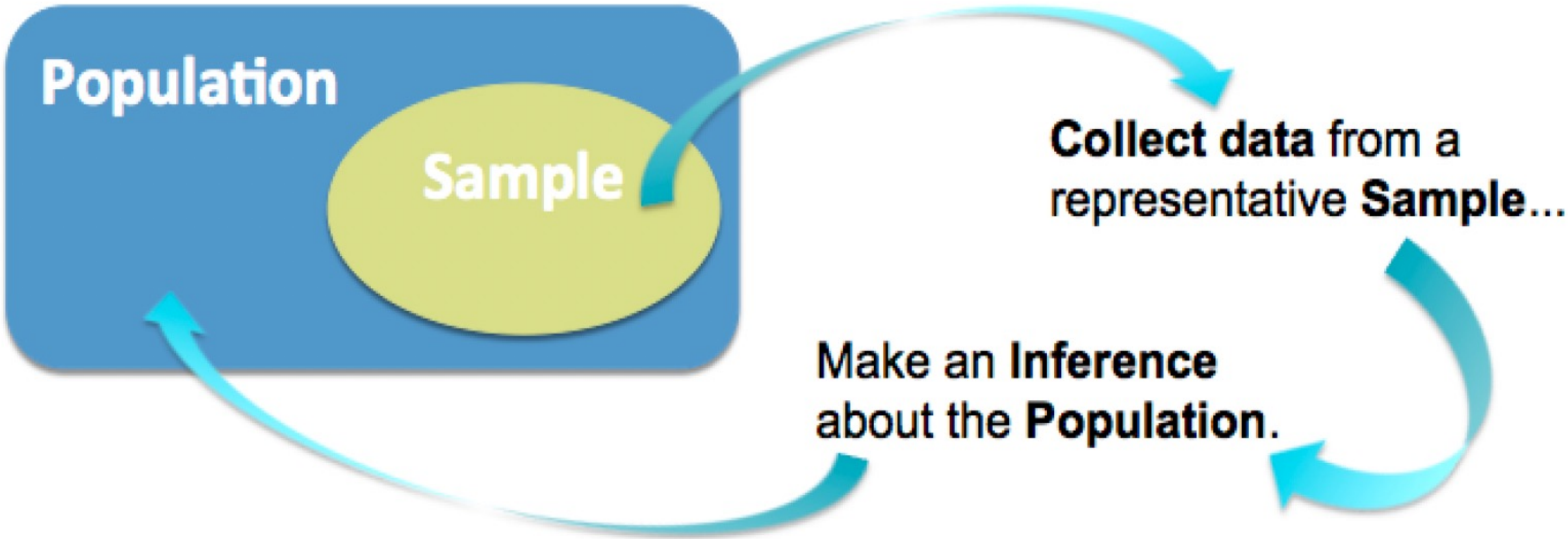
INTERVALOS DE CONFIANZA
TAMAÑO DE LA MUESTRA
CONTRASTES DE HIPÓTESIS
PRUEBAS Z Y KOLMOGOROV

CLASE 4

ESTADÍSTICA INFERENCIAL

Es un proceso donde una muestra es analizada y, con base en su información, se infiere, se deduce o se concluye sobre lo que está sucediendo en una población.

El propósito de la estimación es reflejar el valor del parámetro poblacional. Una buena estimación proporcionará técnicas correctas para encontrar los verdaderos parámetros poblacionales.


Population
Sample
Collect data from a representative Sample...
Make an Inference about the Population.

ALTERNATIVAS PARA LA SIGNIFICANCIA DE LA INFERENCIA ESTADÍSTICA

a. Estimación de intervalos

b. Pruebas de hipótesis

Intervalo de confianza: dos valores númericos que definen un rango dentro del cual se pretende que se encuentre el parámetro de interés con un cierto nivel de confianza.

Para ello se debe conocer la distribución de la variable, la cual puede adoptar diferentes formas. La forma más común en variables continuas que se distribuyen en forma normal, es la distribución Z.

ESTIMACIÓN POR INTERVALOS Y CÁLCULO MUESTRAL

PRIMERO DEBEMOS ENTENDER ALGUNOS CONCEPTOS IMPORTANTES

1. NIVEL DE CONFIANZA

2. SD & ERROR ESTÁNDAR

El nivel de confianza deseado se indica por:

1 - α % = NIVEL DE CONFIANZA

α: nivel de significación - es el máximo error que queremos cometer. Es decir, es la probabilidad de que el valor estimado NO esté en el intervalo calculado.

Mientras > sea el nivel de confianza < será el nivel de α

Al 95%, el valor de α es de 0,05

1 - α = 0,95 $\rightarrow$ 1-0.95 = α $\rightarrow$ 0,05 = α

Los extremos del intervalo: Limite inferior y limite superior (LIC-LSC)

Para tener resultados fiables se necesita un nivel de confianza cercano a 1, es decir:
0,90 / 0,95 / 0,99

Existen dos métodos para poder estimar la media de una población através de intervalos de confianza

MUESTRAS GRANDES:

-muestras compuestas de 30 o más datos (teórica)

*También puede ser utilizado para muestras < 30 datos, siempre que se tenga conocimiento de:

- la distribución poblacional de los datos es normal

- valor de la varianza poblacional o de la desviación estándar poblacional.

MUESTRAS PEQUEÑAS:

-muestras compuestas de menos de 30 datos (teórica)

*Cuando se desconoce el valor de la varianza poblacional o de la desviación estándar poblacional, PERO la distribución de los datos de la población sea normal.

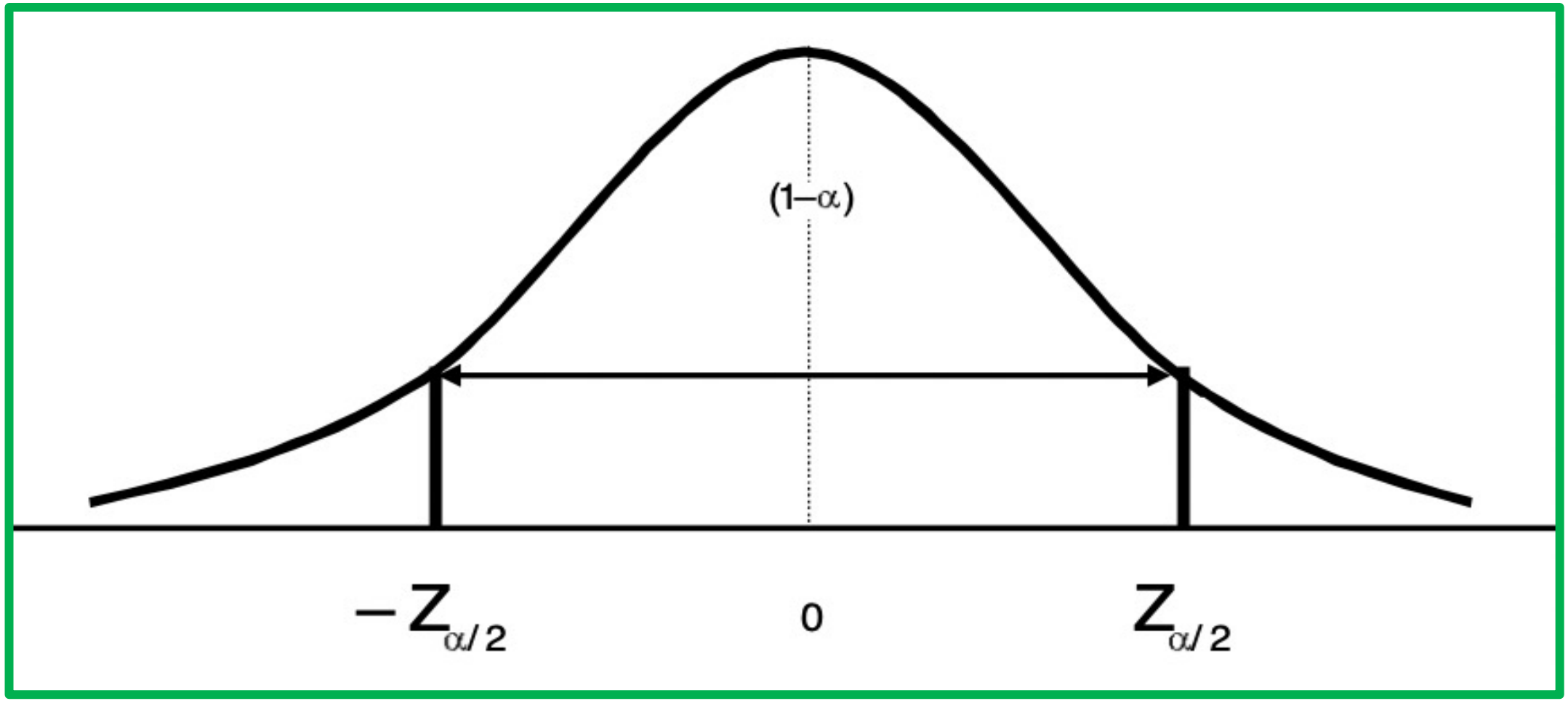
Existen dos métodos para poder estimar la media de una población através de intervalos de confianza

MUESTRAS GRANDES:

Z

MUESTRAS PEQUEÑAS:

T Test


(1–α)
$-Z_{\alpha/2}$
0
$Z_{\alpha/2}$

4 CONCEPTOS CLAVES

DESVIACIÓN ESTÁNDAR

Medida de la dispersión de los datos, cuanto mayor sea la dispersión mayor es la sd.

La desviación estándar es un índice para usar cuando se pretende describir la variabilidad de una variable continua en una muestra.

ERROR ESTÁNDAR

El error estándar de la media se debe usar cuando se pretende cuantificar el error cometido al estimar la media poblacional mediante la media muestral. Oscilaciones de la media muestral (media obtenida en base a los datos medidos en la muestra utilizada) alrededor de la media poblacional (verdadero valor de la media).

No es, un índice de variabilidad, sino una medida del error que se comete al tomar la media calculada en una muestra como estimación de la media de la población. A partir del error estándar se construye el intervalo de confianza de la medida correspondiente.

ENTONCES EN TÉRMINOS FORMALES EL ERROR ESTÁNDAR ES:

POBLACIÓN

$$SE = \frac{\sigma}{\sqrt{N}}$$

MUESTRA

$$SE = \frac{s}{\sqrt{N}}$$

SE: Standard Error

- σ: la desviación estándar de la población
- s: desviación estándar de la muestra
- N: número de observaciones de la muestra

ERROR ESTÁNDAR

MARGEN DE ERROR

Medida del error que se comete al tomar la media calculada en una muestra como estimación de la media de la población

El margen de error es una estadística que expresa la cantidad de error de muestreo aleatorio en los resultados de una encuesta

Cuanto mayor sea el margen de error, menos confianza se debe tener en que el resultado de una encuesta reflejaría el resultado de una encuesta de toda la población .

TAMAÑO MUESTRAL

A través del 'Margin of Error', podemos encontrar el tamaño de la muestra:

$$Error = z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

La desviación típica de las tallas de los hombres de 18 o más años de un país, vale 4

Con esa información, se solicita lo siguiente:

Con un nivel de confianza del 99% se desea estimar el tamaño de la muestra tomando en cuenta que el margen de error no sea mayor a medio centímetro.

$$Error = z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

Despejando la expresión anterior tenemos:

$$n = \frac{(z_{\alpha/2} \times \sigma)^2}{Error^2}$$

Por tanto, necesitamos determinar el valor de Z:

$Z_{\frac{\alpha}{2}}$ → 1 - α = 0,99 → 1-0.99 = α → 0,01 = α → $\alpha/2$ = 0,005

tamaño muestra.xlsx

Sustituyendo en la expresión anterior de n, tenemos

$$n = \frac{(2{,}58 \times 4)^2}{0{,}5^2} = 424{,}63$$

Necesitamos, por tanto, un tamaño de muestra de 425 hombres.

PRUEBAS DE HIPÓTESIS

PRUEBAS DE HIPÓTESIS

HIPÓTESIS DE INVESTIGACIÓN

Suposición que motiva al estudio, según la experiencia.

-Descriptiva: describo las caracteristicas de la poblacion (variables)

-Correlación/Asociaciones

-Diferenciales: Magnitud de diferencia entre grupos

-Causalidad: Relación

HIPÓTESIS ESTADÍSTICA

Proposición sobre los parámetros de una o más poblaciones.

HIPÓTESIS NULA Y ALTERNATIVA

H_0: Depende del test que se analice.

Sustenta igualdad de medias/medianas, independencia, no correlación, normalidad, etc

H_1: Es una afirmación contraria a la H_0 y es la hipótesis que va en el sentido lógico de lo que busca el investigador.

CON LA SIGNIFICANCIA SE PUEDE ACEPTAR O RECHAZAR LA H0

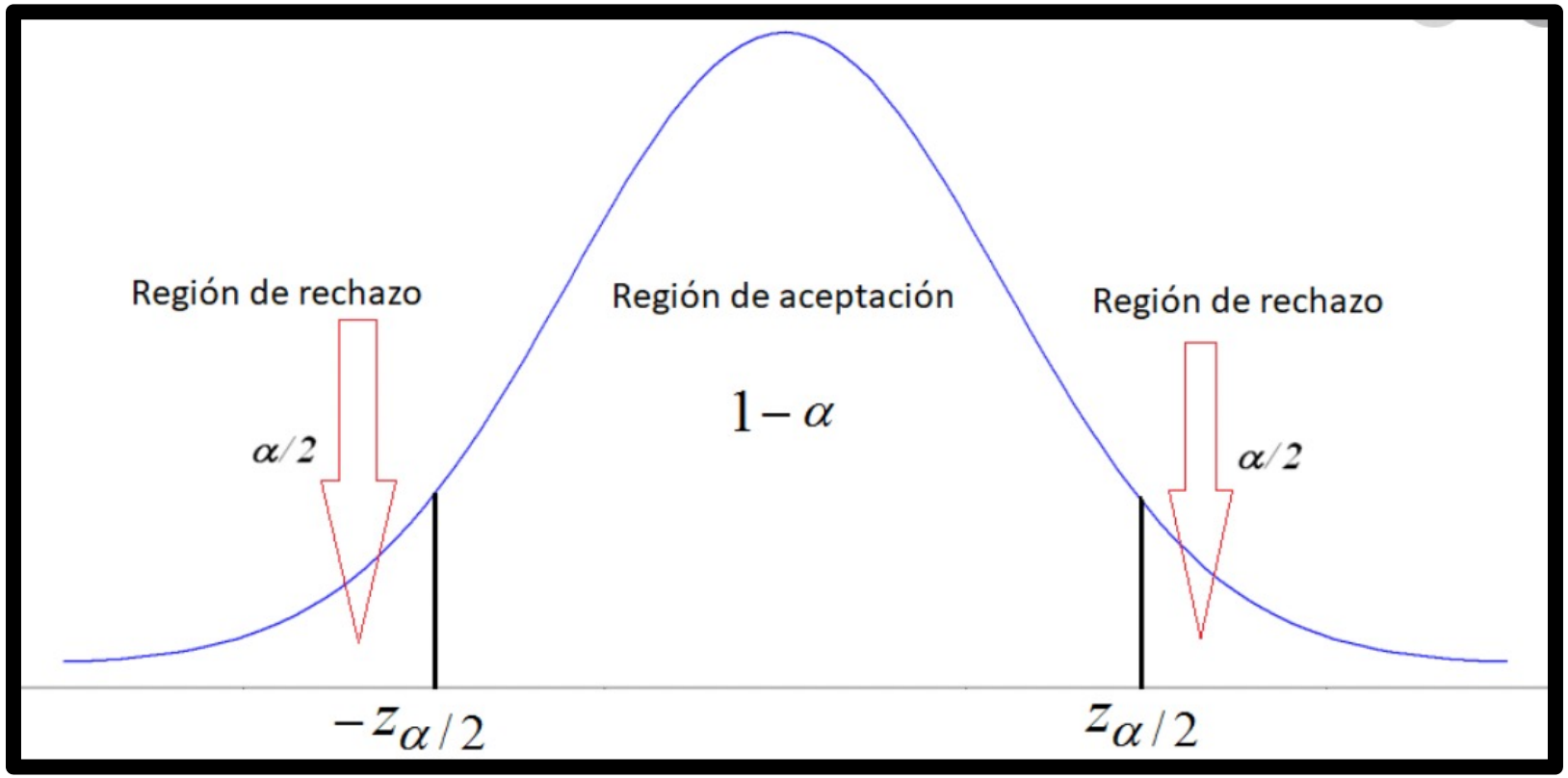
La seleccción de la prueba de hipótesis estadística depende de la escala de medición de las variables estudiadas:

1. cuantitativa (continua/discreta)
2. cualitativa (nominal/ordinal)

y en función de lo que se pretende comparar: medias/varianza/proporciones/riesgo/asociación

REGIÓN DE RECHAZO

En el eje de las Z (distribución normal), se hace una división para establecer la zona o región de rechazo, es decir, aquellos valores que tienen menor probabilidad de ocurrir si la hipótesis nula es verdadera.



α = 0,05 lo que implica el rechazo de la H_0, con probablidad de cometer un error debido al muestreo del 5%.

REGLA DE DECISIÓN

El valor obtenido del estadístico se compara con los valores de distribución de la prueba de hipótesis, y así se determina si el valor se encuentra o no, en la zona de rechazo.

La H_0 se rechaza a favor de la H_1 cuando está en la zona de rechazo, si ese valor no se encuentra ahí entonces no se puede rechazar la H_0.

Entonces:

Rechazar la H_0 o No rechazar la H_0

ERRORES EN PRUEBAS DE HIPÓTESIS

		Ho	
		Verdadera	Falsa (H1 verdadera)
Ho	Aceptar	Decisión correcta $1 - \alpha$	Decisión incorrecta Error de tipo II β
	Rechazar (Aceptar H1)	Decisión incorrecta Error de tipo I α	Decisión correcta $1 - \beta$

P < α rechazo H_0

$\alpha = 0.1/0.05/0.01$

SIGNIFICANCIA CLÍNICA
VS
SIGNIFICANCIA ESTADÍSTICA

La significancia estadística representa la estimación de cometer errores (α y β) ante la decisión de rechazar o aceptar la H_0.

Pero, el hallazgo y peso de los resultados depende de la traducción clínica/biológica de los mismos. Por ejemplo, se puede tener significancia estadística; y, no significancia biológica o visceversa.

PASOS PARA LOS CONTRASTE DE HIPÓTESIS

Proponer una hipótesis que se considera como verdadera, llamada **hipótesis nula**. La inversa de la hipótesis nula se llama **hipótesis alternativa**.

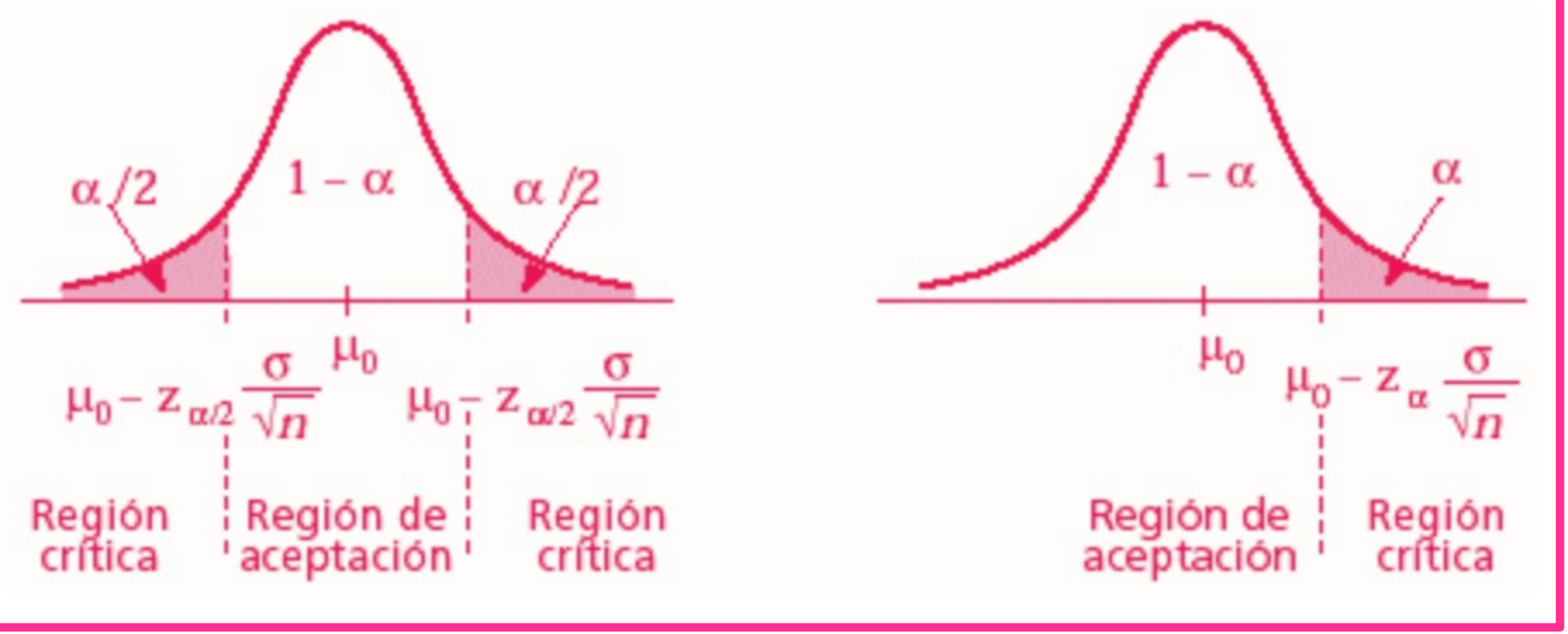
Definir las leyes de probabilidad de la población y de la muestra (en general, se considera una **distribución normal**).

Determinar la **zona de aceptación** de la hipótesis nula, mediante intervalos de confianza.

Fijar posibles zonas de rechazo, donde no se admite la hipótesis nula, que se conocen genéricamente como **región crítica**.

Cuando la región crítica está situada a los dos lados de la zona de aceptación de la hipótesis nula, el contraste se denomina **bilateral** o de dos colas; si está sólo a un lado de la región crítica, se llama **unilateral** o de una cola.

Regiones críticas y de aceptación en contraste de hipótesis bilateral y unilateral, en el caso de la media



DOS CONCEPTOS CLAVES PARA ACEPTAR O RECHAZAR HIPÓTESIS

1. **ESTADISTICO OBSERVADO: LO QUE OBTENGO MEDIANTE LAS FÓRMULAS**

2. **ESTADISTICO CRÍTICO: EL VALOR CON EL QUE COMPARO (EXCEL)**

PRUEBAS ESTADISTICAS INFERENCIALES

PRUEBAS PARAMÉTRICAS

Deben cumplirse algunas condiciones de validez, de modo que el resultado de la prueba paramétrica sea fiable. Por ejemplo: ajusta a una distribución normal y si las varianzas son homogéneas.

- ✓ Prueba Z (diferencia medias)
- ✓ T de student (diferencia medias) (indepedientes/pareadas)
- ✓ Prueba F
- ✓ Correlación Pearson

PRUEBAS NO PARAMÉTRICAS

No deben ajustarse a ninguna distribución. Pueden por tanto aplicarse incluso aunque no se cumplan las condiciones de validez paramétricas.

- ✓ Chi2
- ✓ Fisher
- ✓ Correlación Spearman
- ✓ Mann-Whitney
- ✓ Kruskal/Wallis

✓Kolmogorov

PRUEBA Z

Diferencia de medias

CONTRASTE DE HIPÓTESIS PARA LA MEDIA POBLACIONAL (CON SD CONOCIDA)

Dada una muestra de tamaño n y conocida la desviación típica de la población σ, se desea contrastar la hipótesis nula:

$$H_0 : \mu = \mu_0$$

frente a la alternativa:

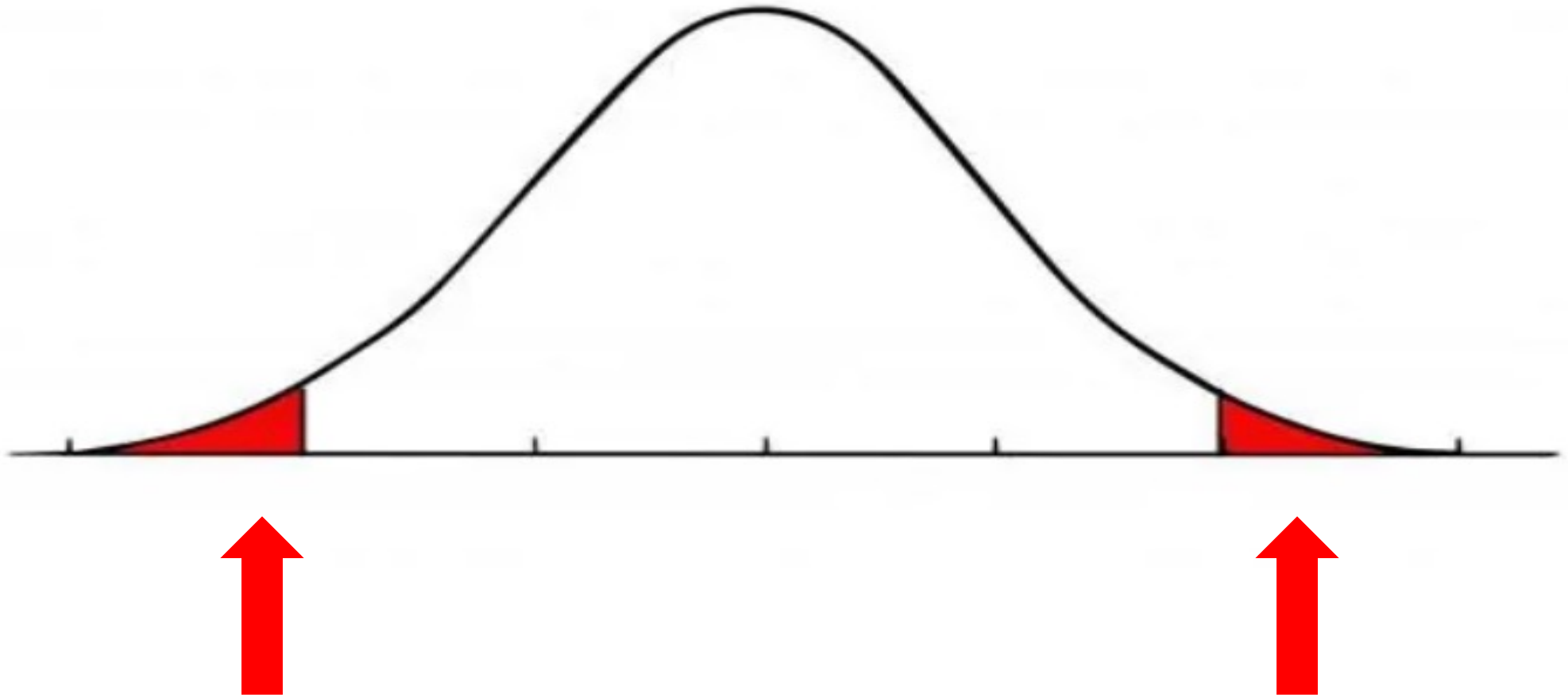
$$H_1 : \mu \neq \mu_0$$

con un nivel de significación α.

Dado el valor muestral de la media X, se determina el estadıstico de contraste:

$$z_c = \frac{\overline{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

CONTRASTE BILATERAL O DOS COLAS



REGIÓN DE RECHAZO DE LA H0 SERÁ:

$$-z \leq -z_{\alpha/2}$$

Valor obtenido mediante fórmula

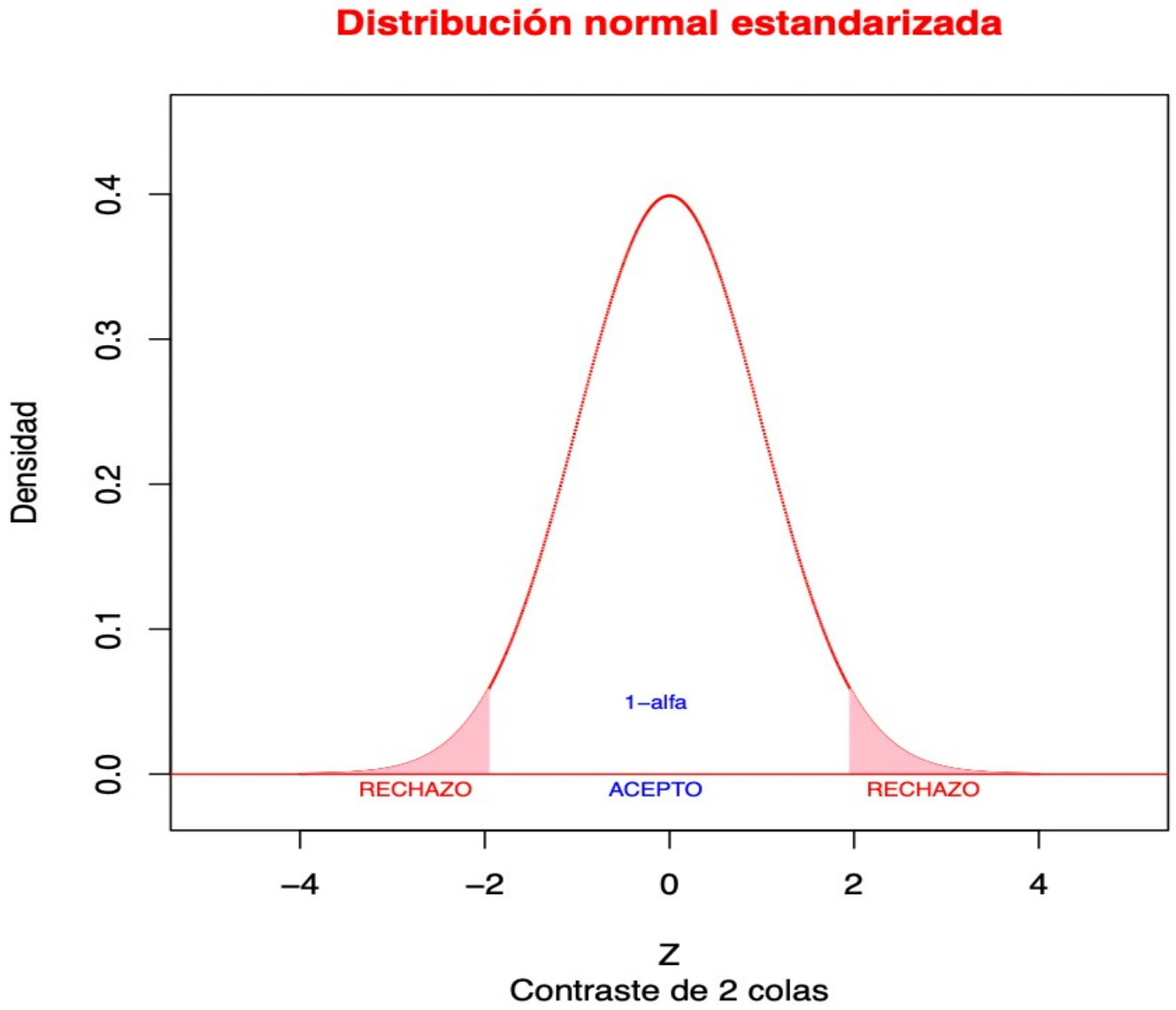
Valor obtenido mediante Excel

$$+z > +z_{\alpha/2}$$

Valor obtenido mediante fórmula

Valor obtenido mediante Excel

ZONAS DE ACEPTACIÓN Y RECHAZO DE LA HIPÓTESIS NULA



EJEMPLO: CONTRASTE DOS COLAS

Se desea contrastar con un nivel de significación del 5%, la hipótesis de que la talla media de los hombres de 18 o más años de un país, es **igual** a 180. Suponiendo que la desviación típica de las tallas en la población vale 4, contraste dicha hipótesis frente a la alternativa de que es distinta.

Los datos constituyen una muestra de n=15 hombres seleccionados al azar, cuyas alturas son:

167 167 168 168 168 169 171 172 173 175 175 175 177 182 195

$H_0: \mu = 180$

$H_1: \mu \neq 180$

$$z_c = \frac{\overline{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

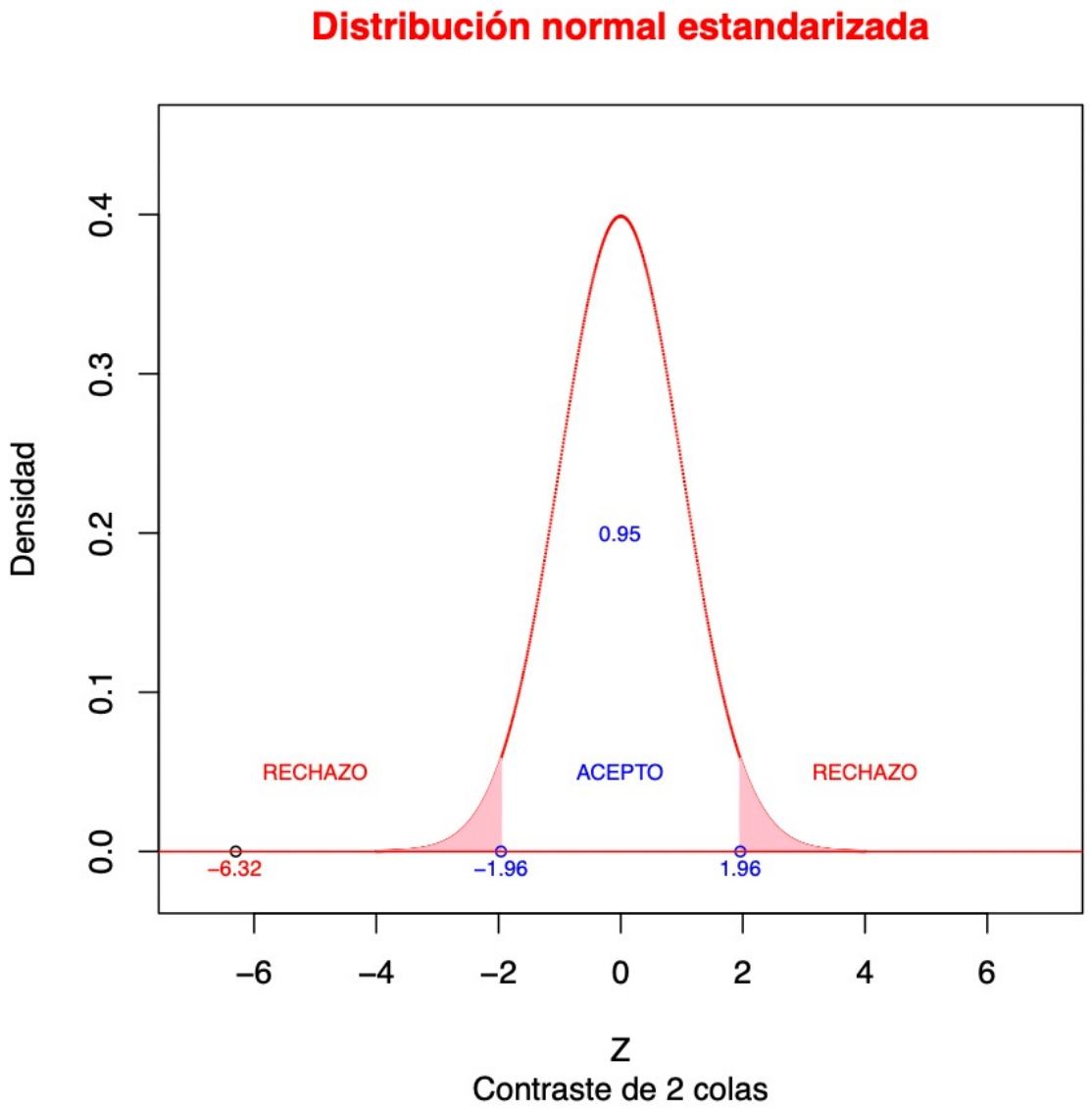
Media muestral: 173,47

$z_{\frac{\alpha}{2}}$ → 1 - α = 0,95 → 1-0.95 = α → 0,05 = α → α/2 = 0,025

$z_{\frac{\alpha}{2}}$ → 1,96 → -1,96

$$z_c = \frac{\overline{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

$$z_c = \frac{173{,}47 - 180}{\frac{4}{\sqrt{15}}} = -6{,}32$$



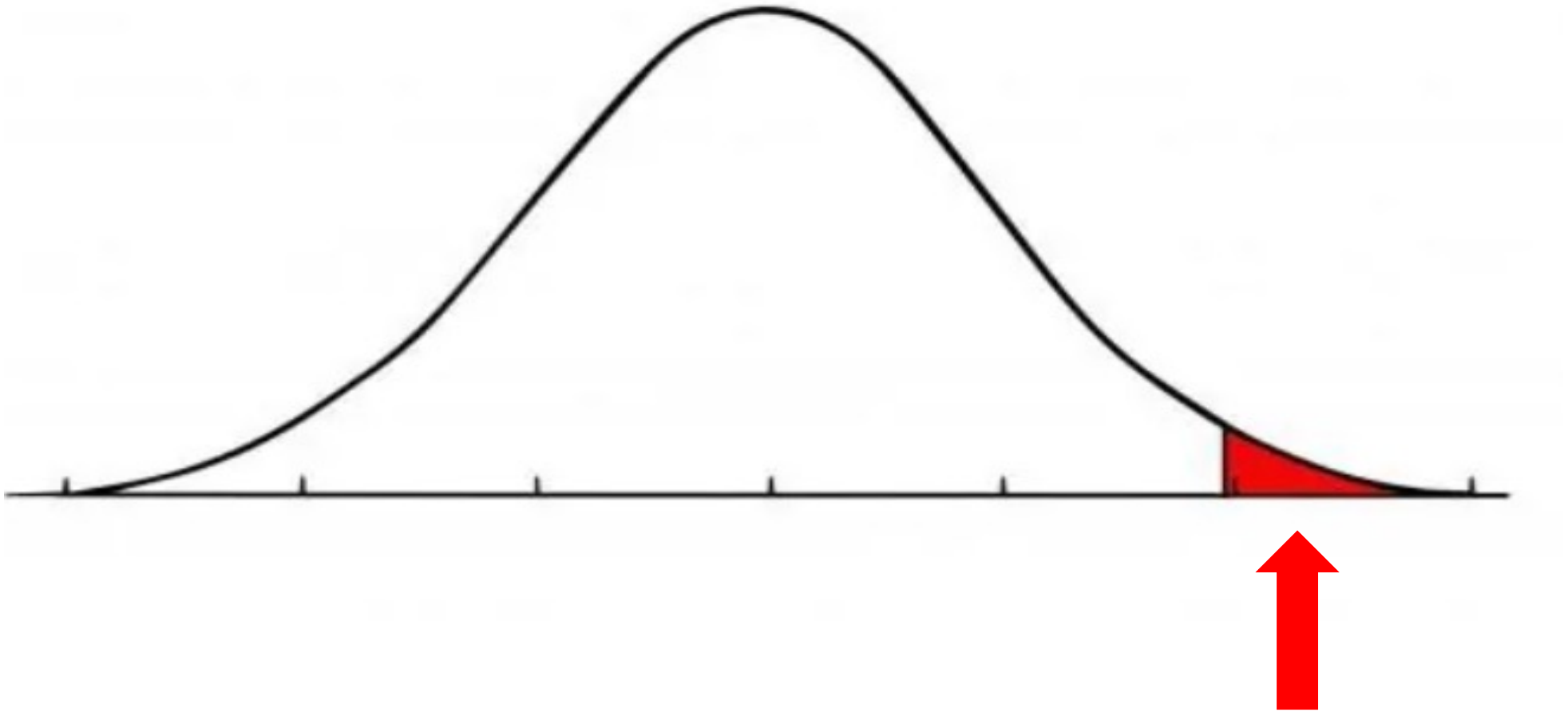
$$-z \leq -z_{\alpha/2}$$

$$-6{,}32 < -1{,}96$$

El valor de Zobs cae en la zona de rechazo.

Por lo que se rechaza la hipótesis nula que establece una talla media igual a 180 cm.

CONTRASTE UNILATERAL O UNA COLA



REGIÓN DE RECHAZO DE LA H0 SERÁ:

Derecha: $+z > +z_{\alpha}$

Valor obtenido mediante fórmula

Valor obtenido mediante Excel

Izquierda: $-z \leq -z_{\alpha}$

Valor obtenido mediante fórmula

Valor obtenido mediante Excel

EJEMPLO: CONTRASTE UNA COLA

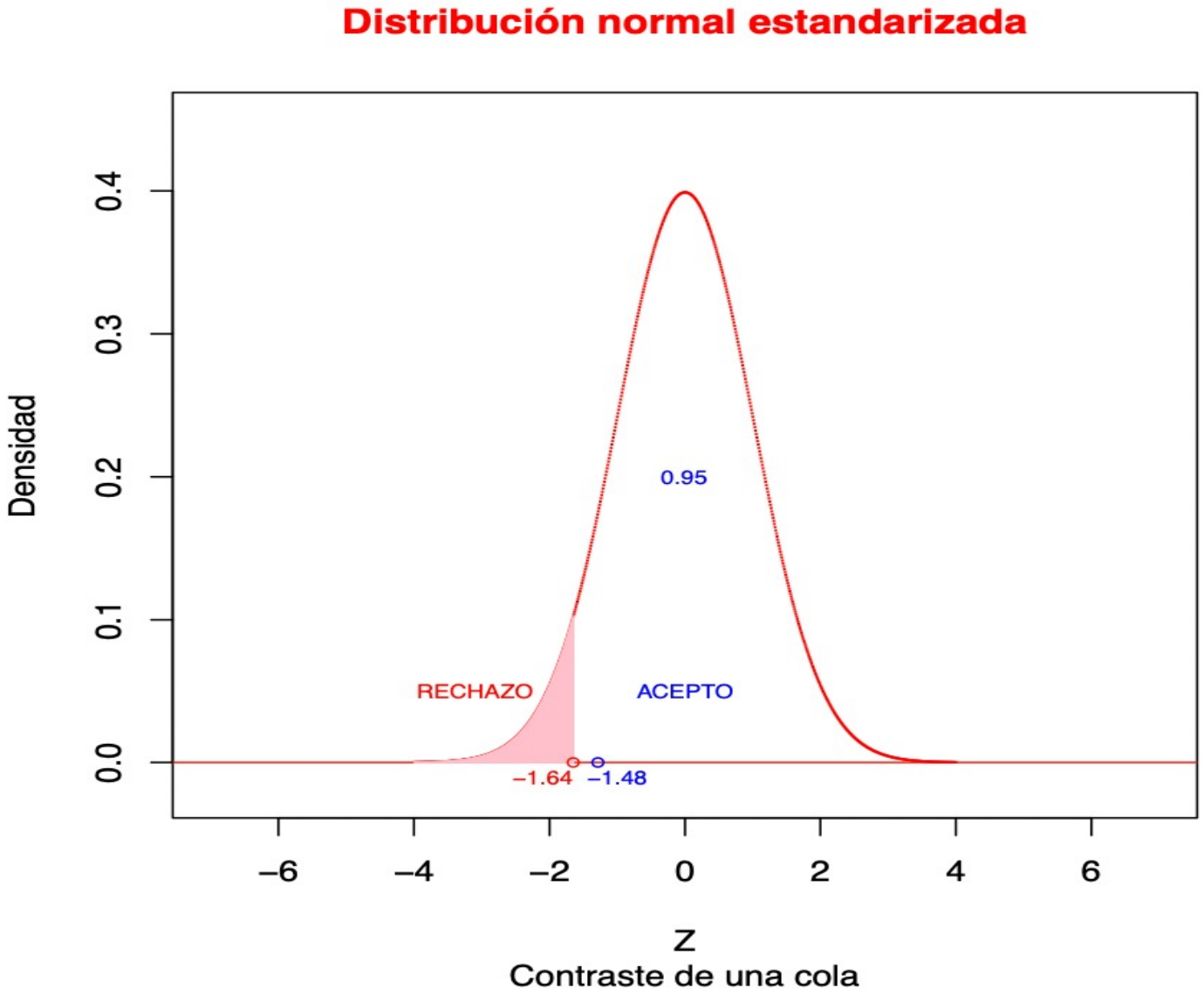
Se desea contrastar con un nivel de significación del 5%, la hipótesis de que la talla media de los hombres de 18 o más años de un país, es **igual o mayor** a 175. Suponiendo que la desviación típica de las tallas en la población vale 4, contraste dicha hipótesis frente a la alternativa de que es distinta. Con una muestra de n=15 hombres seleccionados al azar, cuyas alturas son las del apartado anterior:

$$H_0 : \mu \geq 175$$

$$H_1 : \mu < 175$$

$$z_c = \frac{173{,}47 - 175}{\frac{4}{\sqrt{15}}} = -1{,}48$$

$-Z_{crit0,05} = -1{,}64$



$$-z > -z_{\alpha}$$

$$-1{,}48 > -1{,}64$$

El valor del estadístico de contraste está en la zona de aceptación.

Por lo que no se puede rechazar la hipótesis nula que establece una talla media igual o mayor a 175 cm.

KOLMOGOROV

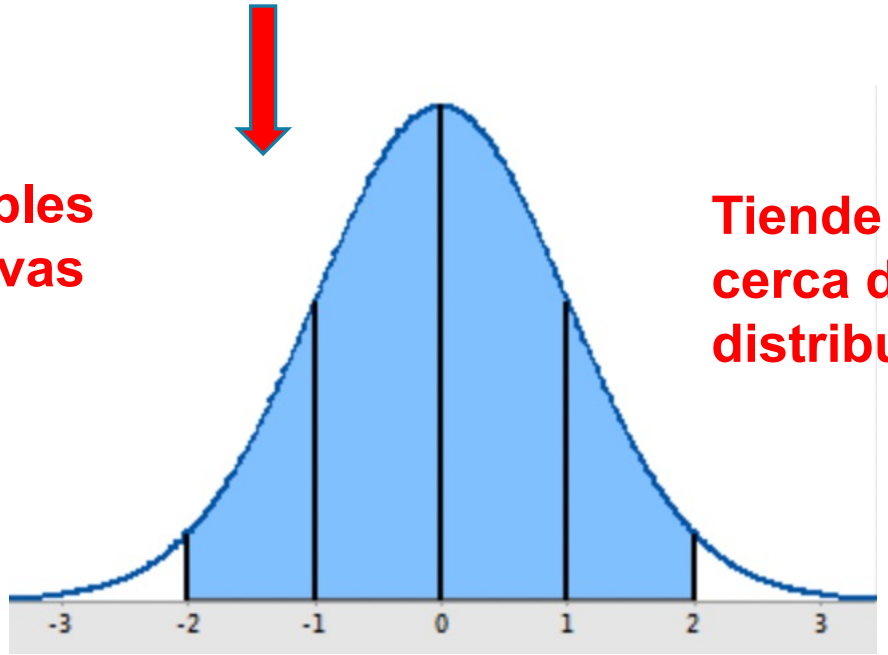
CONTRASTES DE NORMALIDAD

Objetivo de estudio	Variables	Prueba recomendada
Contraste de normalidad	Cuantitativas > 50 BASE DE DATOS	Kolmogorov-Smirnov
	Cuantitativas < 50 BASE DE DATOS	Shapiro-Wilk

KOLMOGOROV-SMIRNOV DE AJUSTE A UNA LEY DE PROBABILIDAD

La prueba de Kolmogorov-Smirnov se utiliza para decidir si una muestra proviene de una población con una distribución específica (ley uniforme – discreta – normal)

Solo se aplica a variables continuas - cuantitativas



Tiende a ser más sensible cerca del centro de la distribución que en las colas.

Kolmogorov-Smirnov compara sus datos con una distribución conocida y le permite saber si tienen la misma distribución.

La prueba no es paramétrica.

Se utiliza comúnmente como una prueba de normalidad para ver si sus datos se distribuyen normalmente.

También se utiliza para comprobar la suposición de normalidad en Análisis de Varianza.

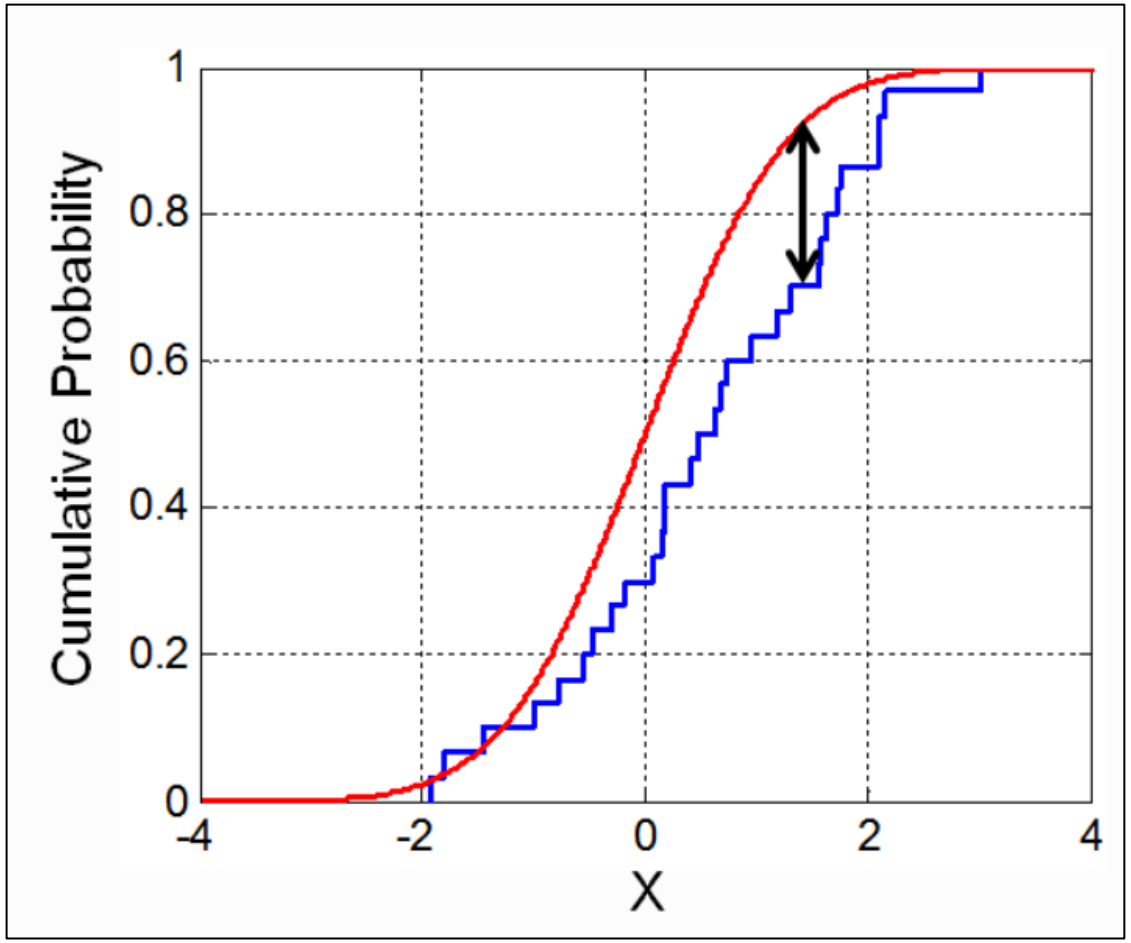
Más específicamente, la prueba compara una distribución hipotética de probabilidad conocida (por ejemplo, la distribución normal) con la distribución generada por sus datos la función de distribución empírica

DISTANCIA DE KOLMOGOROV–SMIRNOV (K-S)

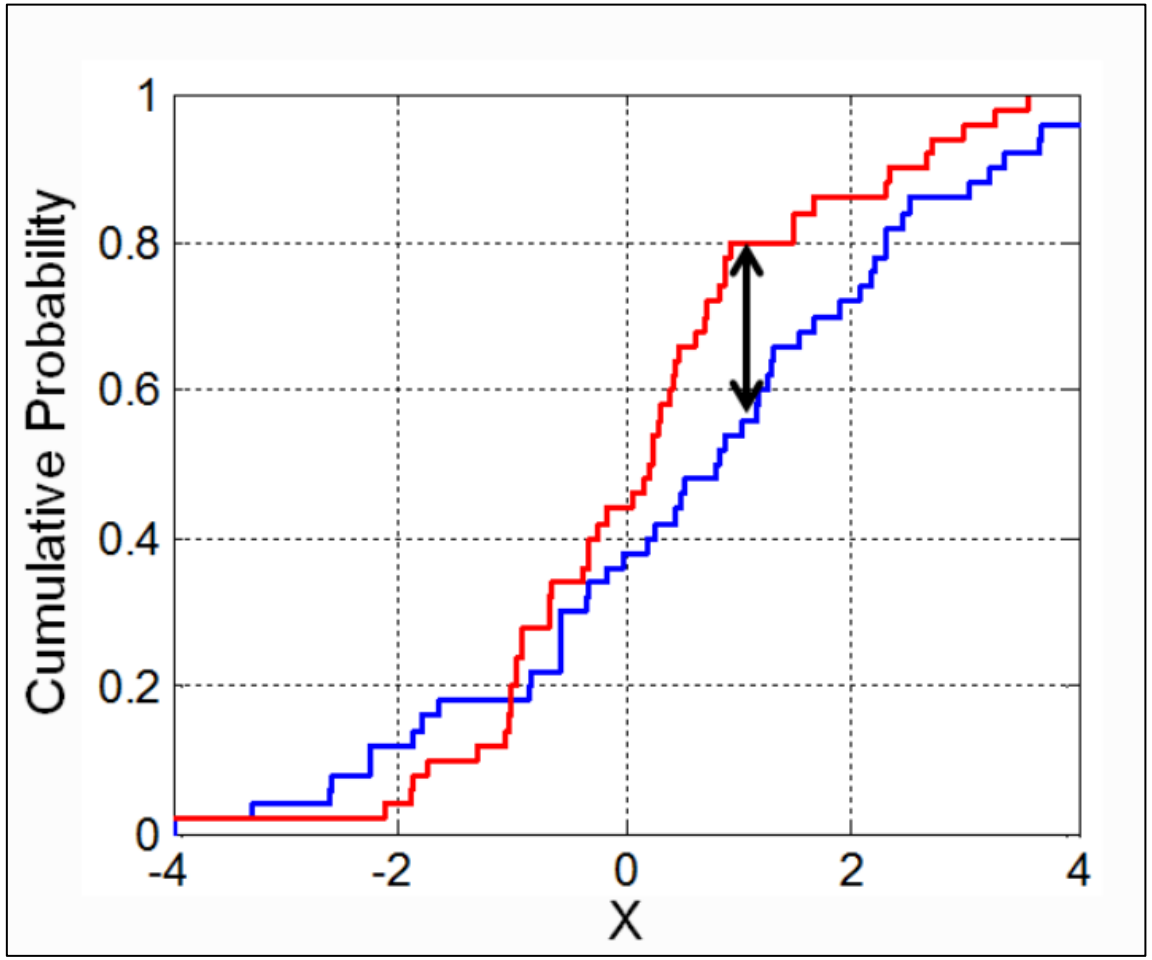
SE DEFINE COMO LA DISTANCIA VERTICAL MÁXIMA ENTRE:

una función de distribución empírica y una función de distribución acumulada teórica de referencia (normalidad).

La ventaja principal del estadístico es que es sensible a diferencias tanto en la localización como en la forma de la función de distribución acumulada.



La línea roja muestra la función de distribución acumulada teórica, la azul la función de distribución acumulada empírica, y la flecha negra es el estadístico K-S



Las líneas roja y azul muestran la función de distribución acumulada empírica de dos muestras

EN RESUMEN: KOLMOGOROV-SMIRNOV

Es un procedimiento utilizado para comprobar la hipótesis nula de la muestra que procede de una población que esta distribuida según una ley de probabilidad específica – Normal.

El estadístico se denota por D_{obs} se definida por:

$$Dobs = \max | Fo(x) - Sn(x) |$$

La distribución del estadístico de Kolmogorov-Smirnov es independiente de la distribución poblacional especificada en la hipótesis nula y los valores críticos de este estadístico están tabulados.

Donde Fo(x) es la probabilidad acumulada esperada y Sn la probabilidad observada correspondiente a la **población normal** especificada en la hipótesis nula.

NIVEL DE SIGNIFICACIÓN

0,05

ZONA DE RECHAZO

Si $D \leq D_\alpha \Rightarrow$ Aceptar H_0

Si $D > D_\alpha \Rightarrow$ Rechazar H_0