Pontificia Universidad JAVERIANA Bogotá

DISTRIBUCION NORMAL E INFERENCIA

ING. HERIBERTO FELIZZOLA JIMENEZ

RUTA DE CIENCIA DE DATOS

Educación Continua

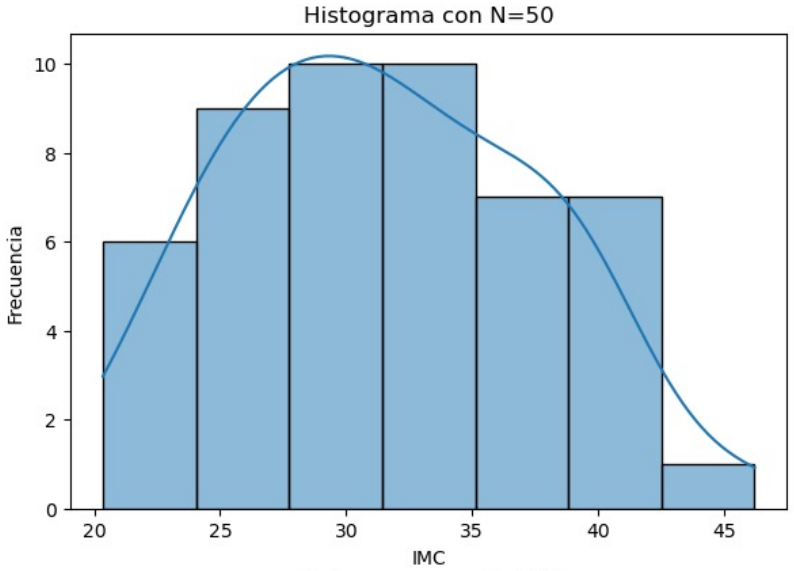
Generamos experiencias educativas

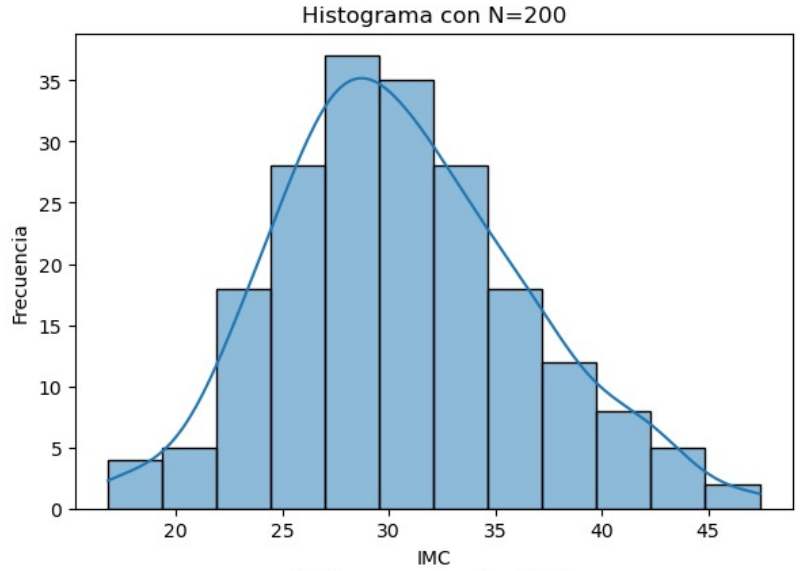
CONTENIDO

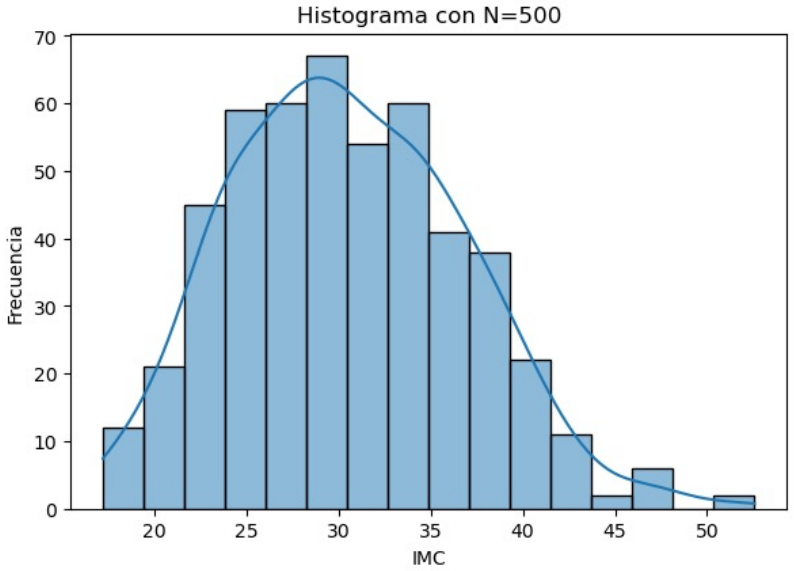
1. Aproximación a la Distribución Normal
2. Utilidad de la Distribución Normal
3. Características de la Distribución Normal
4. Distribución Normal Estándar
5. Cálculo de Probabilidades en la Distribución Normal
6. Intervalos de confianza
7. Prueba de Hipótesis
8. Error Tipo I y Tipo II
9. Procedimiento de Prueba de Hipótesis

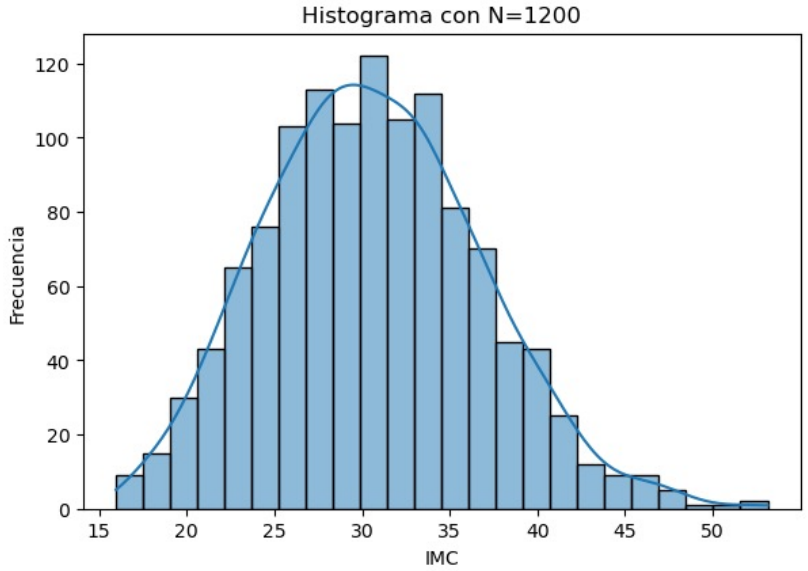
APROXIMACIÓN A LA DISTR. NORMAL

- En las gráficas se presenta la distribución de índice de masa corporal IMC de una población utilizando diferentes tamaño de muestra.
- Al aumentar el tamaño de la muestra se puede observar que la distribución toma una forma particular de campana.
- Hay gran concentración de individuos con IMC entre 25-35, con un media aproximada de 30.
- Además, muy pocos casos en los extremos.
- Este es un comportamiento que se observa en diferentes fenómenos, sistemas o procesos.

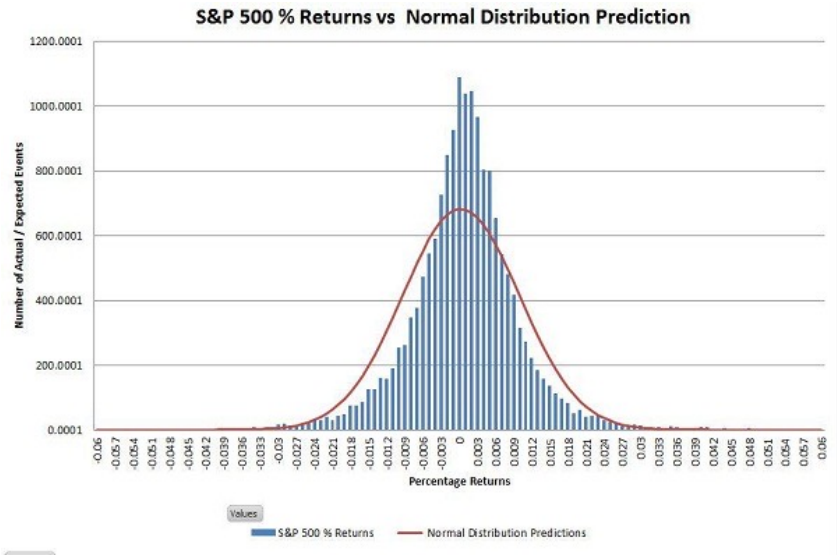


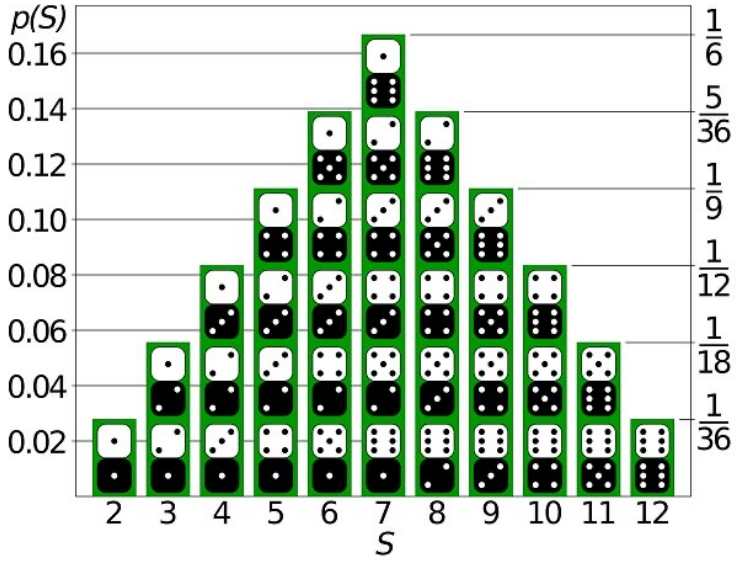


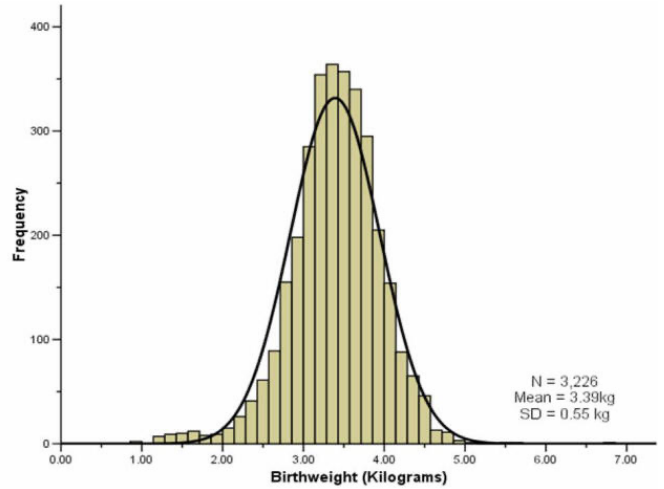


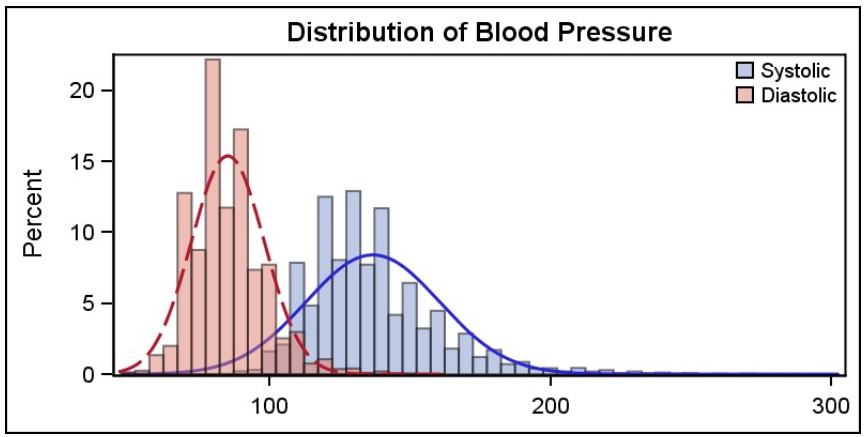


PROCESOS QUE SE COMPORTAN NORMAL





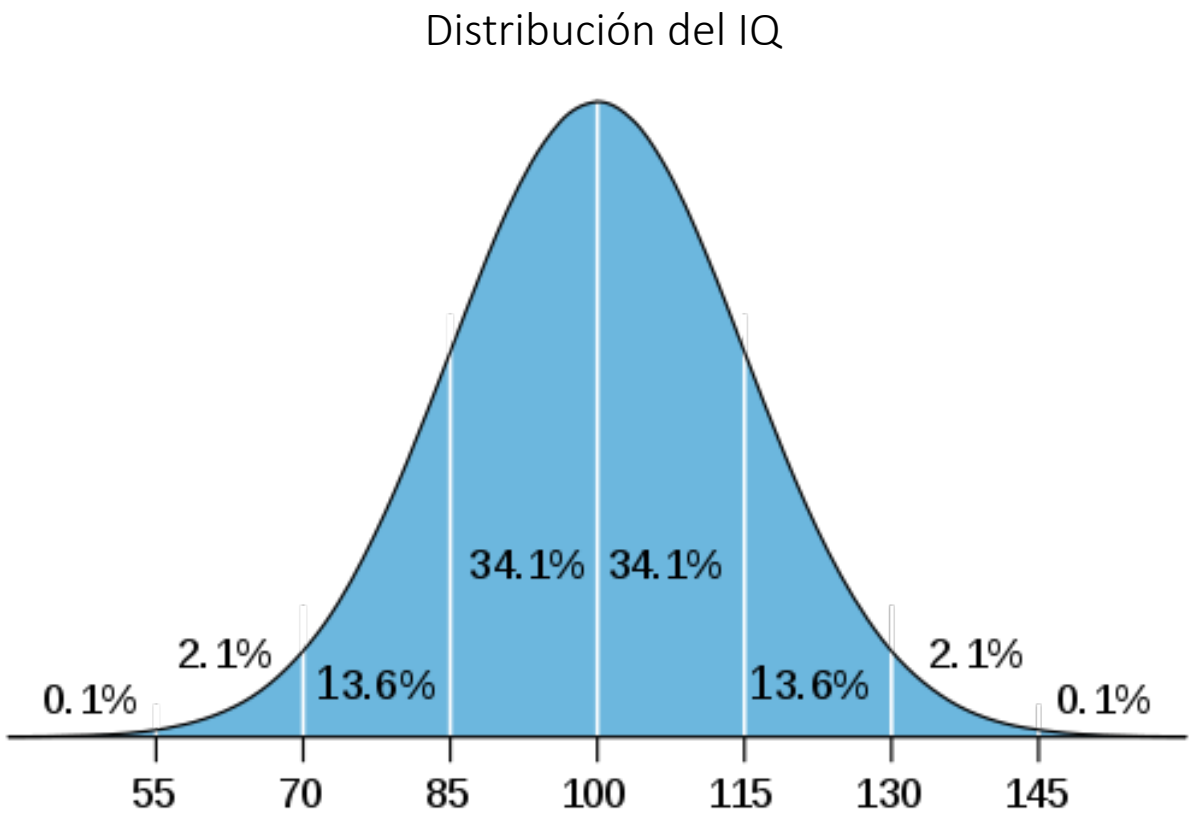




¿Conoce otros fenómenos que se comportan de la misma manera?

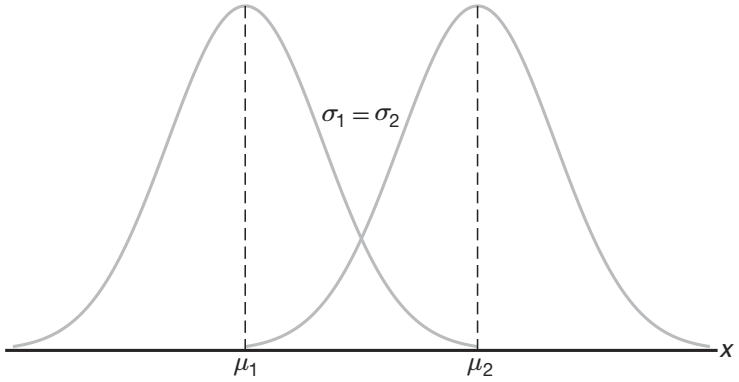
¿QUE UTILIDAD TIENE LA DISTR. NORMAL?

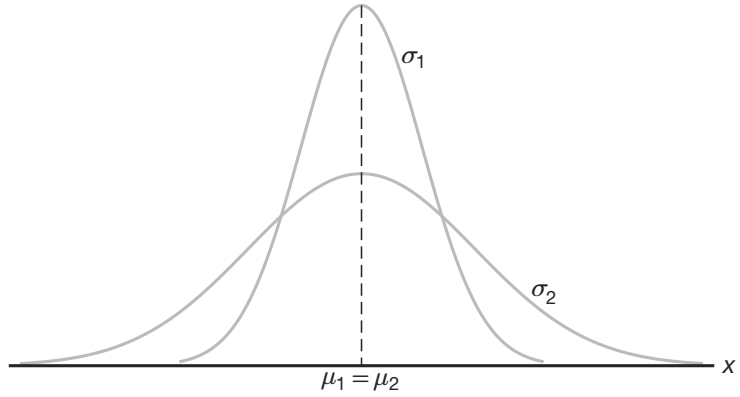
- Podemos aprovechar el enorme potencial que tiene esta distribución para hacer inferencias sobre:
  - Los resultados de un proceso o sistema
  - Los cambios en los fenómenos, sistemas o procesos
  - Resultados inesperado o atípicos
- A partir de esto se pueden tomar decisiones o generar conclusiones válidas para entender y estudiar los sistemas.

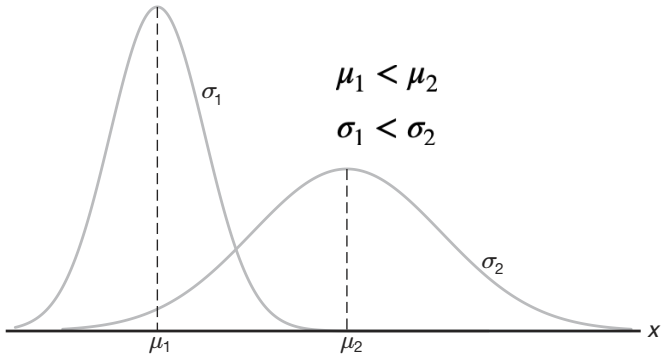


CARACTERISTICAS DE LA DISTR. NORMAL

- Es la distribución de probabilidad continua más importante en todo el campo de la estadística.
- Su gráfica, denominada curva normal o campana de gauss.
- Describe de manera aproximada muchos fenómenos que ocurren en la naturaleza, la industria y la investigación.
- La FORMA de la distribución normal tiene dos parámetros:
  - Media: μ -> parámetro de localización
  - Desviación estándar: σ -> parámetro de forma
- Estos parámetros se pueden estimar con la media y desviación estándar muestral, cuando el tamaño de la muestra es lo suficientemente grande n > 30.
- Una mayor desviación muestra una curva más plana (platicúrtica), lo que indica *menor concentración* de valores en el centro.
- Una menor desviación muestra una curva mas empinada (leptocúrtica), lo que indica una *mayor concentración* de valores hacia el centro.

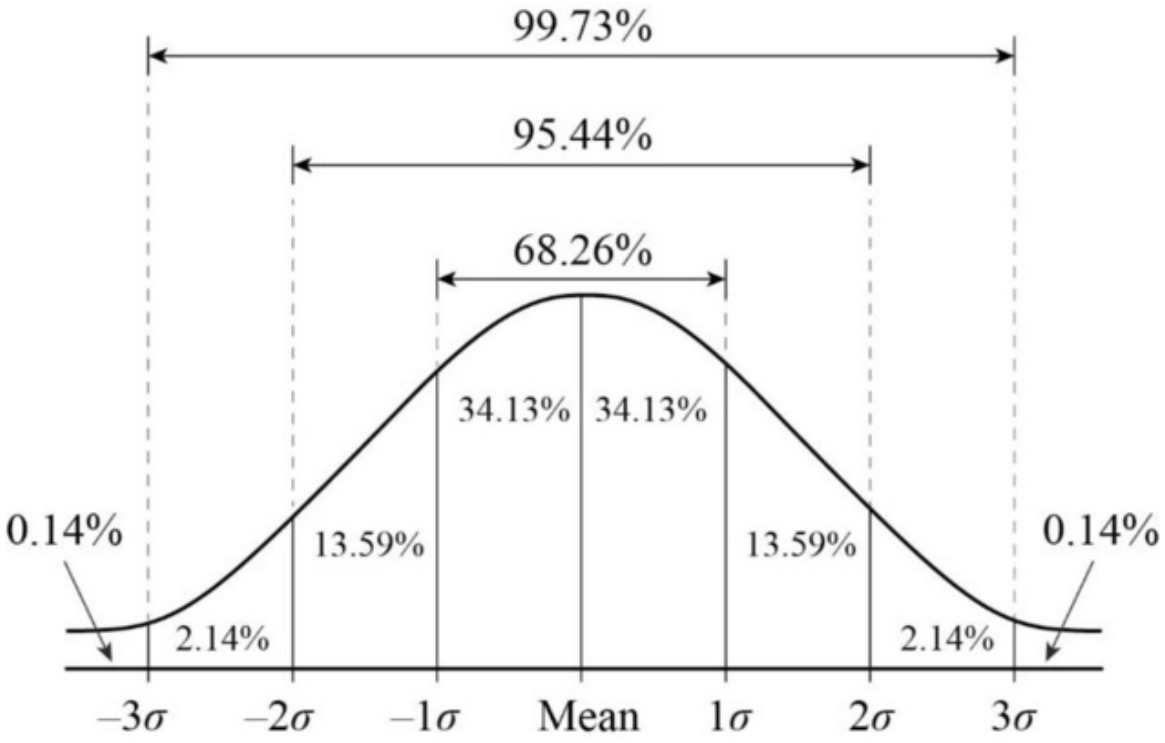






CARACTERISTICAS DE LA DISTR. NORMAL

- La moda, la mediana y la media coinciden.
- La curva es simétrica alrededor de la media.
- La curva tiene sus puntos de inflexión a 1, 2 y 3 desviaciones estándar.
- La curva normal se aproxima al eje horizontal de manera asintótica, conforme nos alejamos de la media en cualquier dirección.
- El área total bajo la curva y sobre el eje horizontal es igual a uno (cumpliendo con un axioma de las probabilidades).

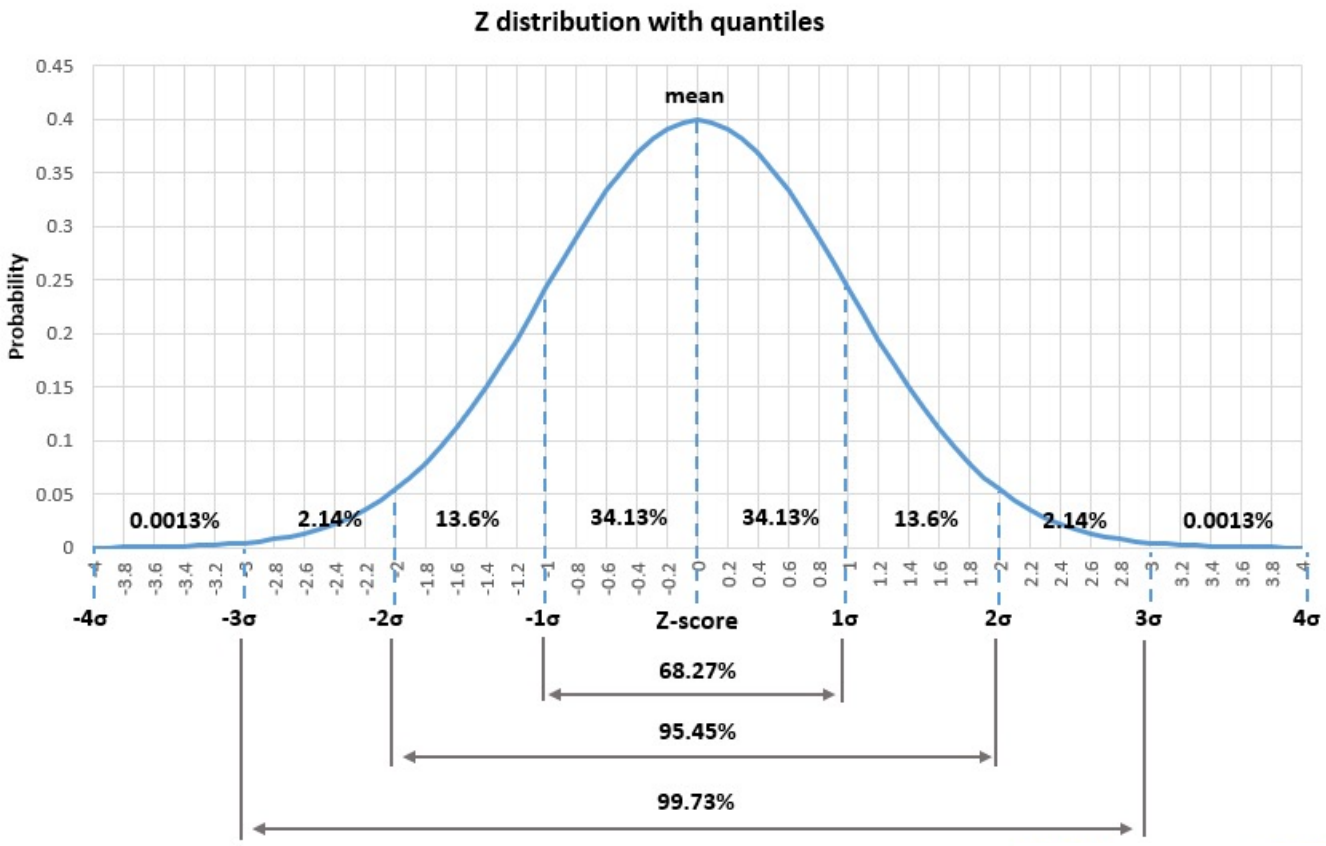


La función de densidad de la normal

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \quad -\infty < x < \infty$$

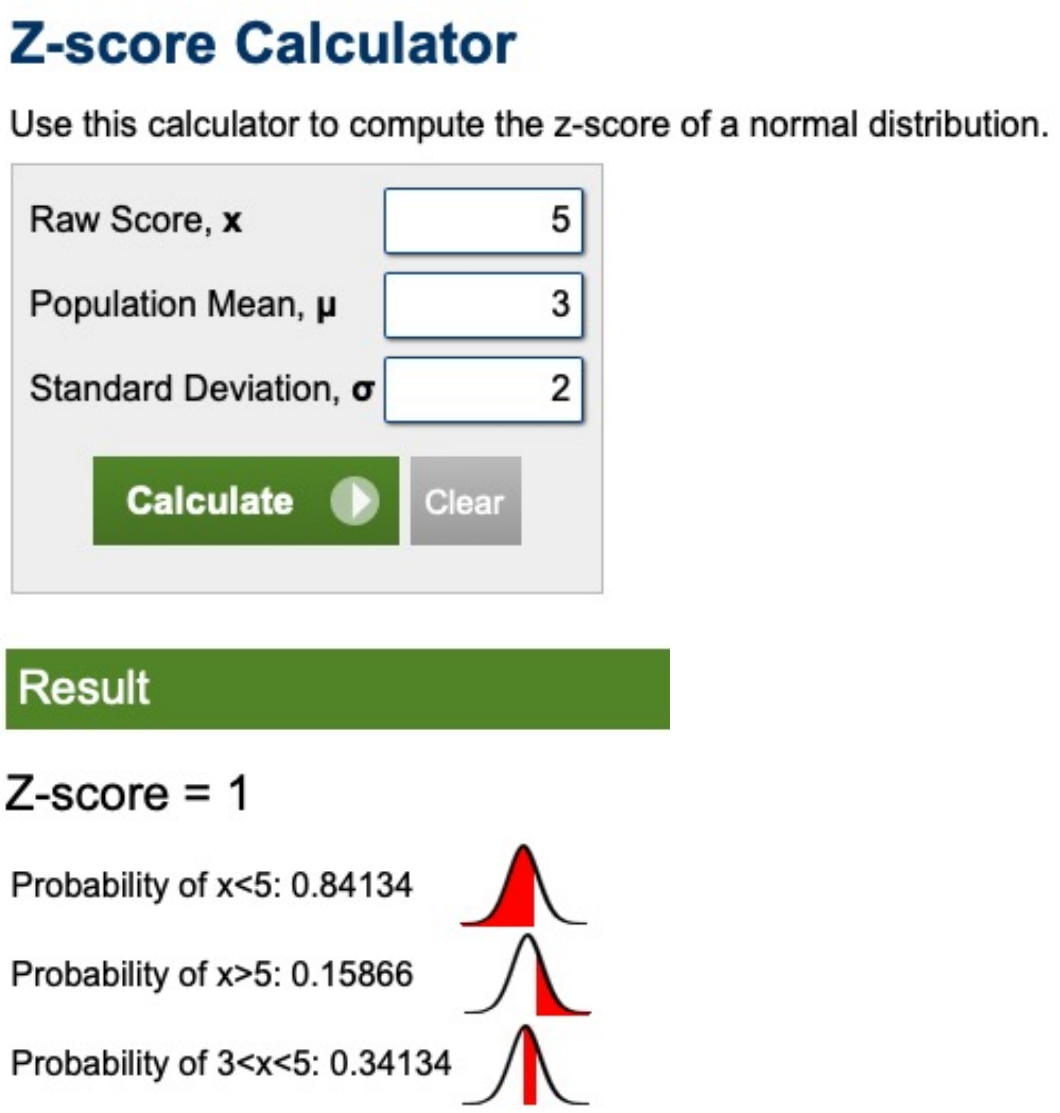
DISTRIBUCION NORMAL ESTANDAR

- Dado que la distribución normal cambia de acuerdo a sus parámetro, media y desviación.
- Entonces tendremos posibilidades infinitas de formas.
- Esto dificulta el modelado, el cálculo de probabilidades y la inferencia.
- Para solucionar esto, se utiliza la distribución normal estándar - Z.
- Donde, $Z = \frac{X-\mu}{\sigma}$
- Esta nueva distribución tiene:
  - $\mu = 0$.
  - $\sigma = 1$



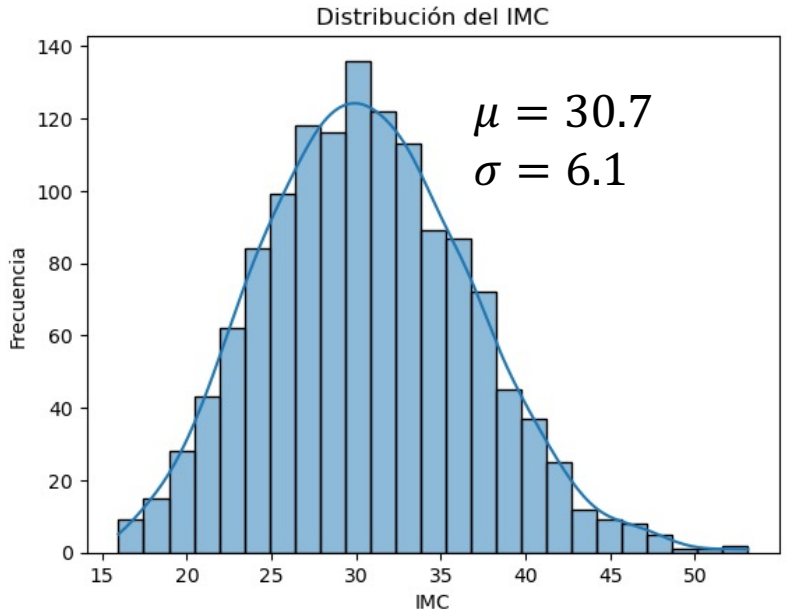
¿COMO CALCULAR LAS PROBABILIDADES CON LA NORMAL?

- Existen múltiples aplicaciones, paquetes estadísticos y tablas que permiten el cálculo de las probabilidades en una distribución normal.
- Por facilidad, en esta primera parte de la clase vamos a utilizar una aplicación on-line, en la segunda parte de la clase utilizaremos Python.
- Ingresemos a: https://www.calculator.net/z-score-calculator.html



CALCULEMOS PROBABILIDADES…

- Suponga que se tiene una población de individuos cuyo índice de masa corporal se distribuye normal con media 30.7 y desviación estándar 6.1
- Suponga que queremos responder la siguiente pregunta:
- *¿Cuál es la probabilidad de que un individuo seleccionado al azar tenga un IMC menor a 22?*



Raw Score, **x** 22

Population Mean, **μ** 30.7

Standard Deviation, **σ** 6.1

Calculate Clear

Result

Z-score = -1.42623

Probability of x<22: 0.076901

Probability of x>22: 0.9231

Probability of 22<x<30.7: 0.4231

Steps:

$$\text{Z score} = \frac{x - \mu}{\sigma}$$

$$= \frac{22 - 30.7}{6.1}$$

$$= -1.42623$$

P-value from Z-Table:

P(x<22) = 0.076901

P(x>22) = 1 - P(x<22) = 0.9231

P(22<x<30.7) = 0.5 - P(x<22) = 0.4231

CALCULEMOS PROBABILIDADES…

Pontificia Universidad JAVERIANA Bogotá

Con el procedimiento utilizado anteriormente, ahora calcule la probabilidad de que cualquier individuo seleccionado al azar se encuentre en alguna de las categorías definidas en la tabla.

BMI	Clasificación	Z1	Z2	P
< 18.5	Bajo			
18.5 – 24.9	Normal			
25.0 – 29.9	Sobrepeso			
30.0 – 34.9	Obesidad I			
35.0 – 39.9	Obesidad II			
> 40	Obesidad III			

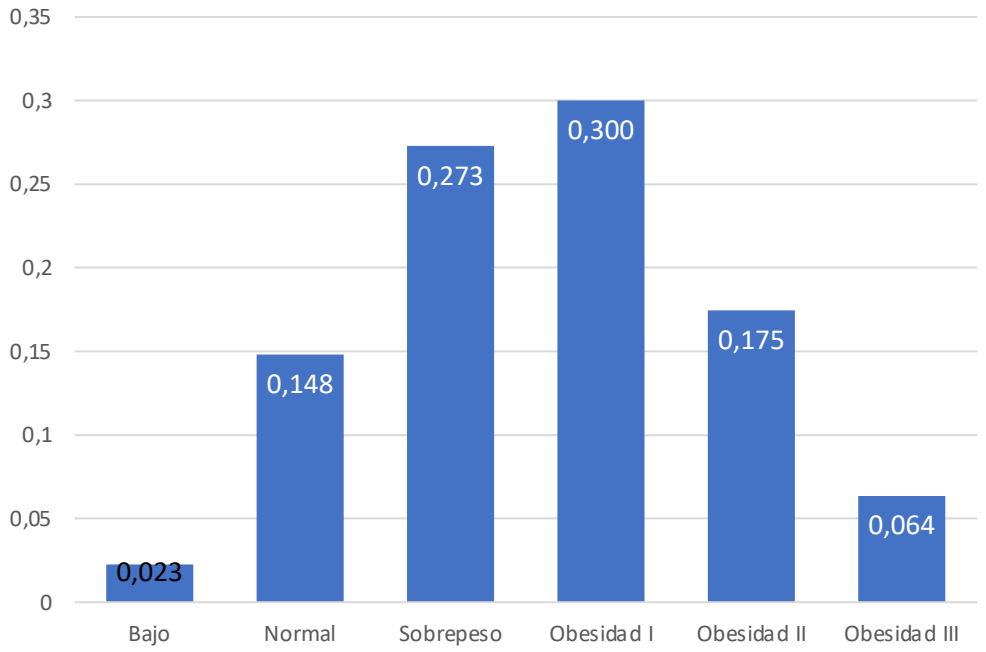
¿Qué puede concluir acerca de las características de la población estudiada de acuerdo al IMC?
¿Qué implicaciones tiene este tipo de comportamiento en la salud pública, los costos del sistema, entre otros?

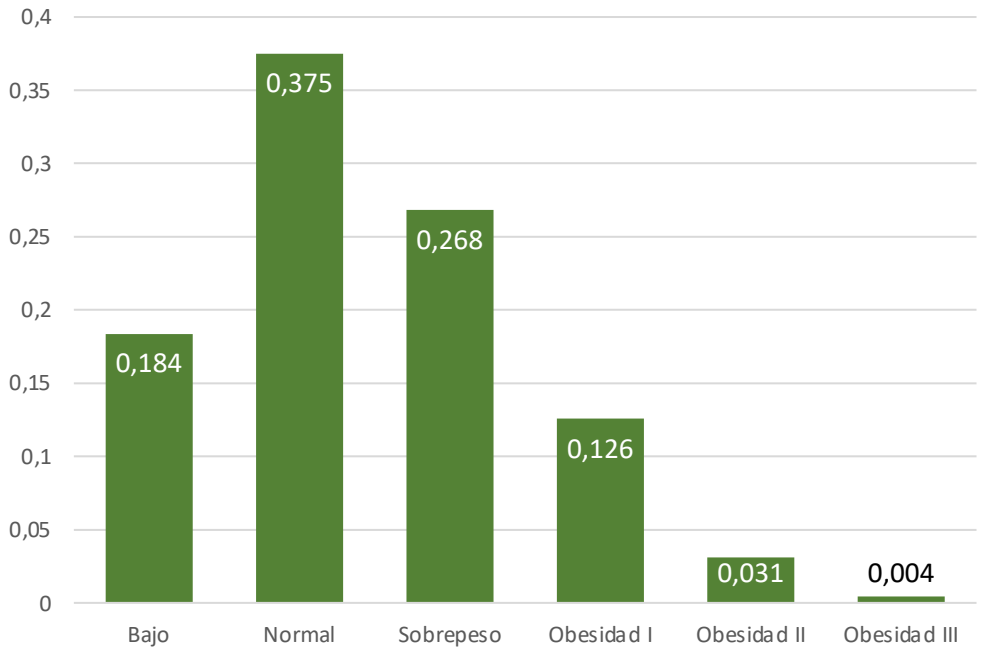
Clasificación del BMI según la OMS

BMI	Nutritional status
Below 18.5	Underweight
18.5–24.9	Normal weight
25.0–29.9	Pre-obesity
30.0–34.9	Obesity class I
35.0–39.9	Obesity class II
Above 40	Obesity class III

Fuente: World Health Organization https://www.euro.who.int/en/health-topics/disease-prevention/nutrition/a-healthy-lifestyle/body-mass-index-bmi

¿QUE EFECTOS PUEDE TENER UN CAMBIO EN LA MEDIA?





Después de un cambio en el IMC promedio ¿Qué puede observar en las probabilidades de cada categorías?

ESTIMACIÓN PUNTUAL

- En el análisis y control de sistemas o procesos nos interesa conocer sus parámetro estadísticos, ejemplo:
  - El costo promedio de tratamiento de una enfermedad
  - El tiempo promedio de permanencia en hospitalización o urgencias
  - Promedio de atenciones por día
  - La proporción de individuos que posee cierta enfermedad
  - La proporción de errores en diagnósticos
- El estimador puntual de un parámetro es un estadístico que genera un valor numérico simple, y que se utiliza para proporcionar una estimación del valor del parámetro desconocido.
- Los estimadores puntuales (estadísticos) más recomendados para estimar estos parámetros son (aunque estos dependen de la distribución), respectivamente:
  - La media muestral $\hat{\mu} = \bar{X}$
  - La varianza muestral $\hat{\sigma}^2 = S^2$
  - La desviación muestral $\hat{\sigma} = S$
  - La proporción de cierta categoría en la muestra: $\hat{p} = \frac{X}{n}$, donde X es el número individuos en una categoría

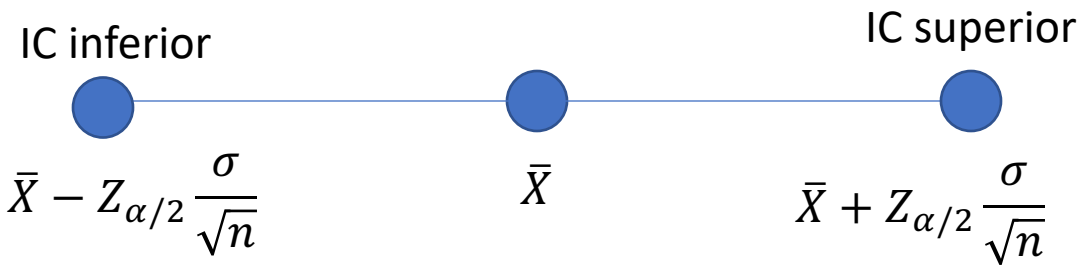
ESTIMACIÓN POR INTERVALO

- La estimación puntual genera un estadístico que depende de la muestra. Por tanto, varía con cada muestra que se tome.
- Esto nos lleva a cuestionarnos sobre qué tanta certeza hay sobre el verdadero valor del parámetro.
- Una forma de tener esta información es conocer el error estándar del estadístico.
- Por tanto, saber qué tan precisa es la estimación consiste en calcular un intervalo de confianza IC que indique un rango "donde puede estar el parámetro" con cierto nivel de seguridad o confianza.
- Casi siempre calculamos los IC con niveles de confianza que van de 90%, 95% o 99%, dependiendo de la naturaleza del proceso, del impacto en el cliente y los recursos con los que se cuenta para ejecutar el estudio.
- Un intervalo de confianza del 95% podría ser interpretado como: 95 de cada 100 IC calculados a partir de muestras distintas contendrán el verdadero valor del parámetro poblacional.

¿COMO SE CONSTRUYE UN INTERVALO DE CONFIANZA?

- La mayoría de las veces, la fórmula para encontrar el intervalo de confianza de los parámetros tiene la siguiente estructura:
  - IC = Estadístico de la Muestra ± Margen de Error
  - Margen de Error = K * Medida de Variabilidad
  - Estadístico = Media, Varianza, Proportion, etc.
  - Factor de Confianza, K = Constante basada en una distribución de probabilidad estadística

- A continuación, se ilustra el cálculo del intervalo de confianza para la media μ



- Donde,
  - $\bar{X}$ es la media muestral
  - $Z_{\alpha/2}$ es una constante que se extrae de la distribución normal estándar. Para un nivel de confianza del 95% se utiliza $\mathrm{Z}_{\alpha/2} = 1.96$
  - $\alpha/2 = (1 - 0.95)/2$
  - σ es la desviación estándar
  - n el tamaño de la muestra

INTERVALOS PARA MEDIA, VARIANZA Y PROPORCIONES

Intervalo de confianza para la Media: μ

$$\bar{x} - \frac{z_{\alpha/2}\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + \frac{z_{\alpha/2}\sigma}{\sqrt{n}}$$

Intervalo de confianza para la Varianza: σ^2

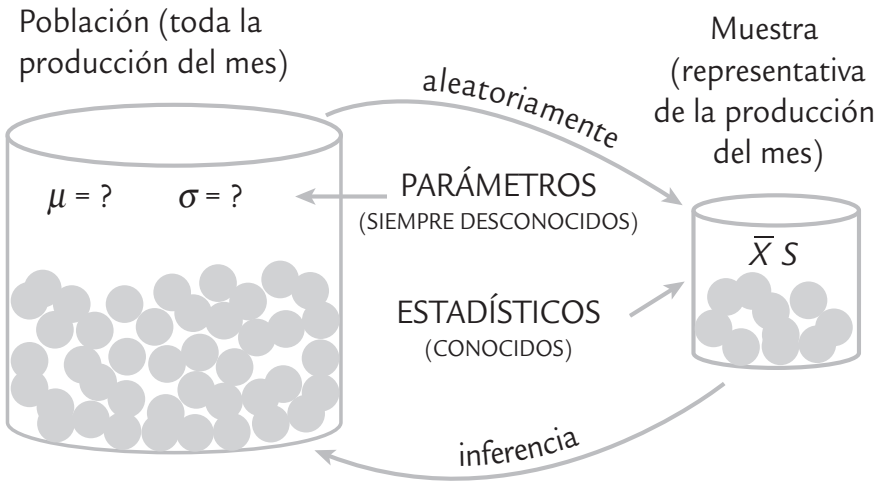
$$\frac{(n-1)s^2}{\chi^2_{\alpha/2,n-1}} \leq \sigma^2 \leq \frac{(n-1)s^2}{\chi^2_{1-\alpha/2,n-1}}$$

Intervalo de confianza para la proporción: p

$$\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq p \leq \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

PRUEBA DE HIPÓTESIS

- Una hipótesis estadística es una afirmación con respecto a una característica desconocida de una población de interés.
- Ejemplo:
  - El salario promedio en Colombia es de 1.2 millones
  - El porcentaje de registros defectuoso en una muestra es del 4%
  - La efectividad de la vacuna contra el covid19 es del 90%
  - La demanda de un articulo por mes es 5000 unidades
  - El tiempo promedio de espera es de 15 minutos
- La prueba de hipótesis nos permite decidir si la afirmación se encuentra apoyada en evidencia experimental o datos muestrales.
- Por lo general, se indaga sobre parámetros estadísticos o la forma de la distribución de una población.
- Las decisiones se toman basado en probabilidad. Por tanto, no hay 100% de certeza en inferencias realizadas.



PRUEBA DE HIPÓTESIS EN PARAMÉTROS ESTADÍSTICOS

- Suponga que se quiere hacer seguimiento a la oportunidad en la atención médica especializada.
- Se ha fijado un valor meta de 10 días máximo.
- Interesa detectar aumentos en el tiempo promedio por encima de los 10 días.
- Para controlar el tiempo promedio se toman muestras de 9 casos/día.
- Se sabe que la desviación estándar en el tiempo de oportunidad es de $\sigma = 1.5$ días.

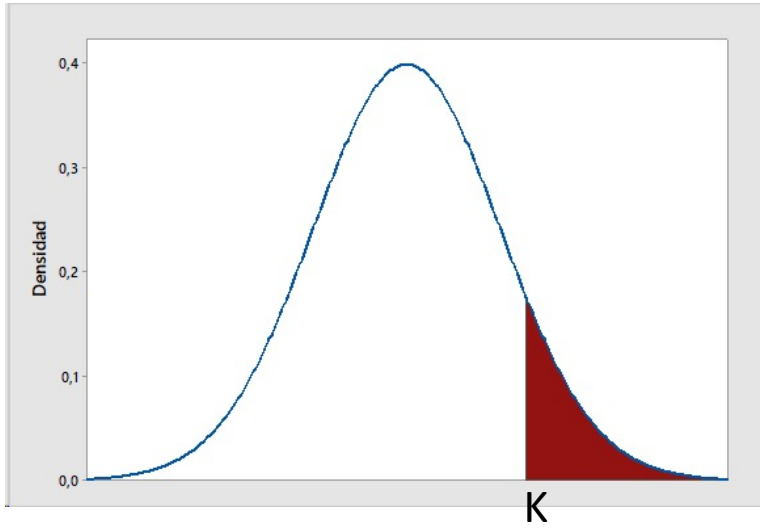
HIPÓTESIS ESTADÍSTICAS

- En toda prueba de hipótesis se plantean dos hipótesis:
  - Hipótesis nula – H_0: es la afirmación sobre la característica de la población (parámetro o forma).
  - Hipótesis alterna – H_1: representa una forma de negación de la hipótesis y se toma verdadera cuando la hipótesis nula es rechazada.
- En el ejemplo:
  - $\mathrm{H_0}: \mu = 10\ días$
  - $\mathrm{H_1}: \mu > 10\ días$
- Las hipótesis deben ser mutuamente excluyentes y colectivamente exhaustivas.
- Por tanto:
  - Si se quiere indagar: $\mathrm{H_1}: \mu \neq \theta \rightarrow \mathrm{H_0}: \mu = \theta$
  - Si se quiere indagar: $\mathrm{H_1}: \mu > \theta \rightarrow \mathrm{H_0}: \mu \leq \theta$
  - Si se quiere indagar: $\mathrm{H_1}: \mu < \theta \rightarrow \mathrm{H_0}: \mu \geq \theta$
- Donde, θ es el parámetro estadístico sobre el que se hace la inferencia.

REGIÓN CRÍTICA

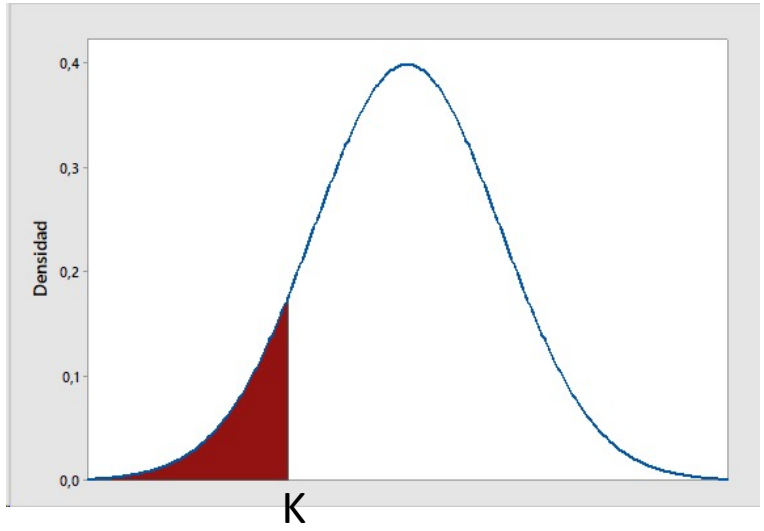
- Para rechazar la hipótesis nula se define un valor de referencia que define una *región crítica*.
- Si el estadístico cae en la región crítica se rechaza la hipótesis nula.
- La ubicación de la región crítica se define de acuerdo a la hipótesis alterna.

La región crítica esta sombreada en rojo y esta delimitada por un valor crítico K



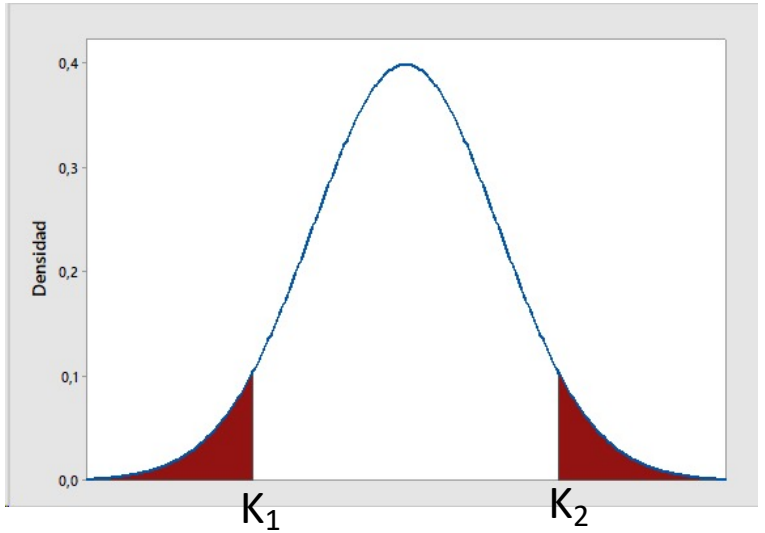
$H_1 : \mu > \mu_0$

Prueba de cola derecha



$H_1 : \mu < \mu_0$

Prueba de cola izquierda

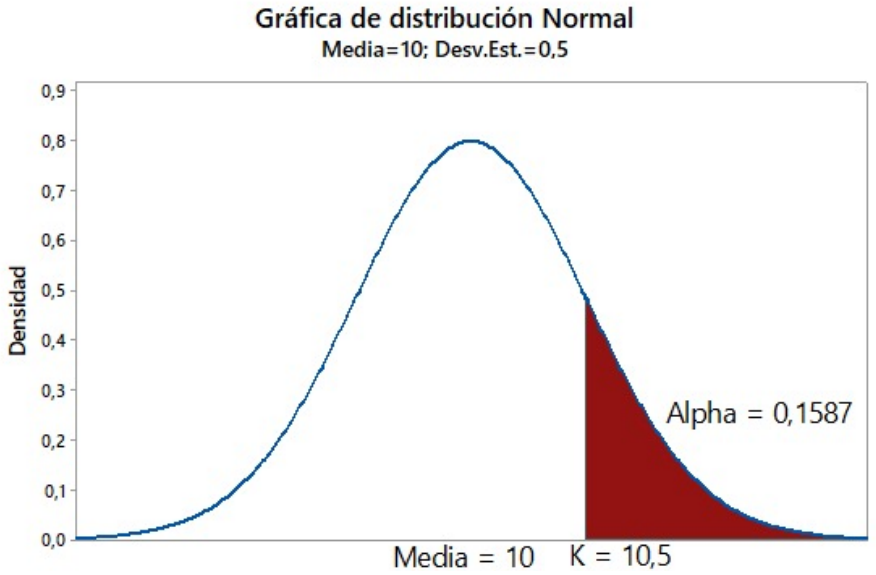


$H_1 : \mu \neq \mu_0$

Prueba Bilateral

ERROR TIPO I

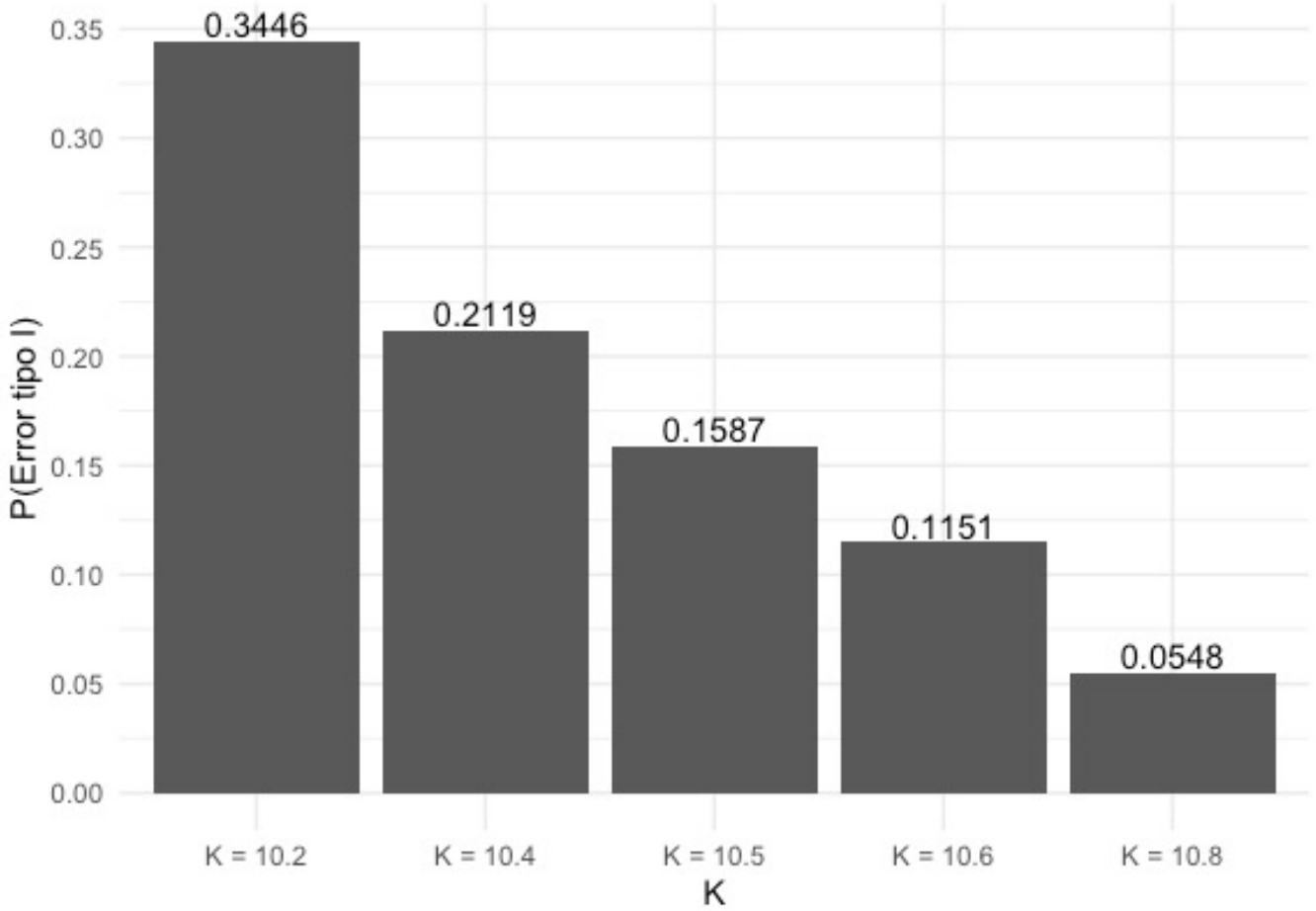
- Para la oportunidad de atención suponga que se ha establecido como valor crítico K = 10.5 días.
- Esto quiere decir que si el tiempo de oportunidad promedio $\overline{X}$ de la muestra es mayor a 10.5 días se infiere que μ es mayor a 10 minutos (se rechaza H_0).
- Tomar esta decisión implica que hay un riesgo de equivocarse.
- Esto es, afirmar que $\mu > 10$ cuando realmente $\mu = 10$.
- Error tipo I o riesgo tipo I: *Rechazar H_0 dado que es verdadera.*
- La probabilidad asociada: $\alpha = \mathrm{P}(\text{Error tipo I})$.
- El α también es denominado nivel de significancia.
- De acuerdo al teorema del límite central - TLC, si la oportunidad de atención tiene una desviación estándar $\sigma = 1.5\ días$, entonces la media tiene una desviación $\sigma_{\bar{X}} = {}^{\sigma}/_{\sqrt{n}} = 1.5/\sqrt{9} = 0.5\ días$



$$\alpha = P(\bar{X} > 10.5 \mid \mu = 10)$$
$$= P\left(Z > \frac{10.5 - 10}{1.5/\sqrt{9}}\right)$$
$$= P(Z > 1) = 0.1587$$

Si K = 10.5 existe una probabilidad de 0.1587 de inferir que la media es mayor a 10 días, cuando realmente esto es falso.

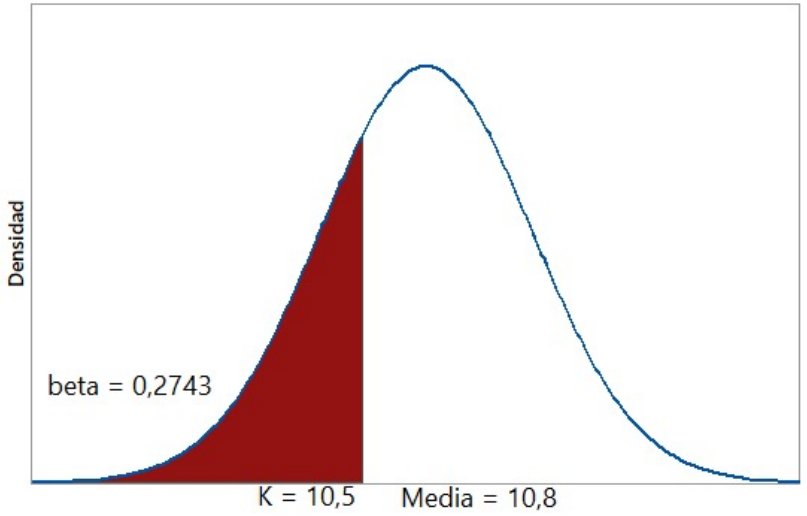
K VS EL RIESGO TIPO I

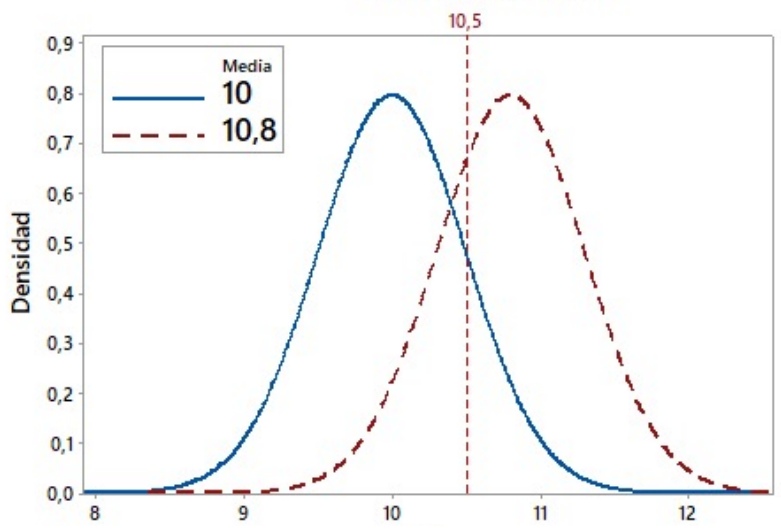


En la gráfica se puede observar que a medida que el valor de K se aleja de $\mu = 10$ disminuye el riesgo tipo I

ERROR TIPO II

- Si la oportunidad promedio de una muestra es menor a K = 10.5 entonces se infiere que el tiempo promedio $\mu = 10.$
- Existe un riesgo o posibilidad de equivocarse en esta decisión, ya que, es probable que valores promedios por debajo 10.5 pertenezcan a poblaciones con $\mu > 10.$
- Error tipo o riesgo tipo II: No rechazar H_0 dado que es falso.
- La probabilidad asociada: $\beta = \mathrm{P}(\text{Error tipo II})$.
- El error tipo II implica plantear un valor alterno para $\mu = 10$ minutos, por ejemplo $\mu_1 = 10.8$ días.
- $\beta = \mathrm{P}(\overline{\mathrm{X}} < 10.5 \mid \mu = 10.8)$
- $\beta = \mathrm{P}\left(\mathrm{Z} < \frac{10.5-10.8}{1.5/\sqrt{9}}\right) = \mathrm{P}(\mathrm{Z} < -0.6) = 0.2743$
- Esto quiere decir que hay una probabilidad 0.2743 de inferir que $\mu = 10$ días, cuando realmente es $\mu_1 = 10.8$ minutos.





ERROR TIPO I VS ERROR TIPO II

		Hipótesis verdadera	
		H_0	H_1
Hipótesis rechazada	H_0	Error tipo I α Nivel de Significancia (Falso Positivo)	Decisión acertada 1- α Nivel de Confianza
	H_1	Decisión acertada 1- β Potencia de la prueba	Error Tipo II β (Falso Negativo)

CONCLUSIÓN DÉBIL VS CONCLUSIÓN ROBUSTA

- Recuerde que para un valor de K = 10.5:
  - P(Error tipo I) = α = 0.1587
  - P(Error tipo II) = β = 0.2743
- Uno de los parámetros de la prueba de hipótesis es la P(Error tipo I) o α con valores de 0.01, 0.05 o 0.1, por tanto cuando se rechaza H_0, existe un probabilidad predefinida y relativamente baja de equivocarse en esta decisión.
- Por esta razón, rechazar H_0 es considerada un conclusión robusta.
- En cambio, cuando no se rechaza H_0 el riesgo de equivocarse en esta decisión depende del valor especifico de una media alternativa hipotética.
- Entre mas cerca μ_1 de μ_0 mayor el riesgo de no detectar el cambio.
- Por esta razón, no rechazar H_0 es considerada conclusión débil.

PROCEDIMIENTO

- Se recomienda el uso de la siguiente secuencia de pasos en la aplicación de la metodología de comprobación de hipótesis.
    1. **Parámetro de interés:** A partir del contexto del problema, identifique el parámetro de interés.
    2. **Hipótesis nula, H_0:** Enunciar la hipótesis nula, H_0.
    3. **Hipótesis alternativa, H_1:** Especifique una hipótesis alternativa adecuada, H_1.
    4. **Cálculos:** Calcule estadísticos muestrales.
    5. **Estadístico de prueba:** Determine un estadístico de prueba apropiado.
    6. **Nivel de confianza y nivel de significancia:** fije un nivel de confianza y el α o nivel de significancia
    7. **Tome una decisión sobre H_0:** Compare el valor p y con el nivel de significancia fijado previamente.
    8. **Concluya:** Decida si se rechaza o no H0 y comuníquelo en el contexto del problema.

EJEMPLO: PRUEBA DE HIPÓTESIS PARA LA MEDIA

- **Problema:** Se ha fijado un objetivo de 10 días para oportunidad. Si en una una muestra la oportunidad promedio es mayor a 10 días se revisará el proceso para analizar las causas y realizar ajustes.
- **Parámetros estadísticos:** Se sabe que la desviación estándar de la oportunidad de atención es $\sigma = 1.5$ días. Utilice un nivel de confianza del 95%, por tanto un $\alpha = 0.05$ para probar las hipótesis.
- **Muestreo:** Para determinar la oportunidad promedio se toman muestras de 9 casos. Los datos de un día muestran los siguientes resultados: 11, 9, 11, 10, 10, 12, 13, 12, 10
- **Media muestral:**

$$\overline{X} = \frac{11 + 9 + 11 + 10 + 10 + 12 + 13 + 12 + 10}{9} = 10.89$$

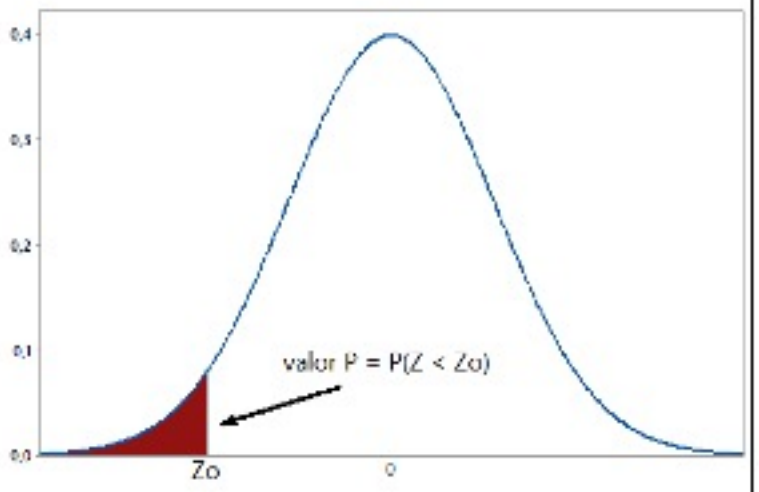
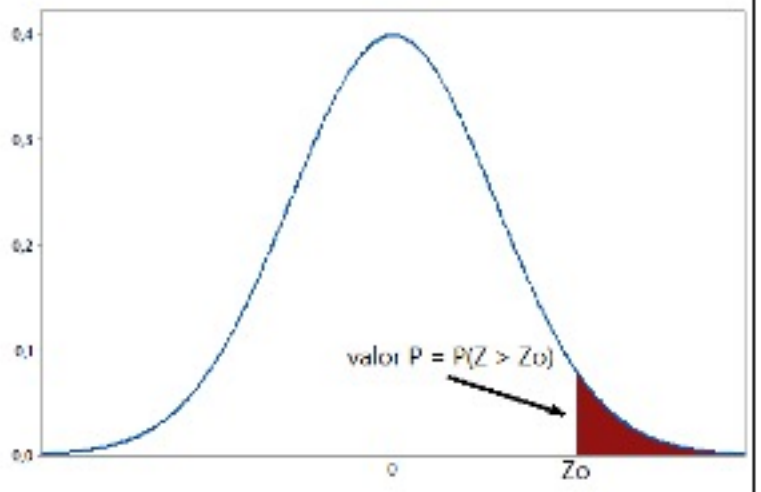
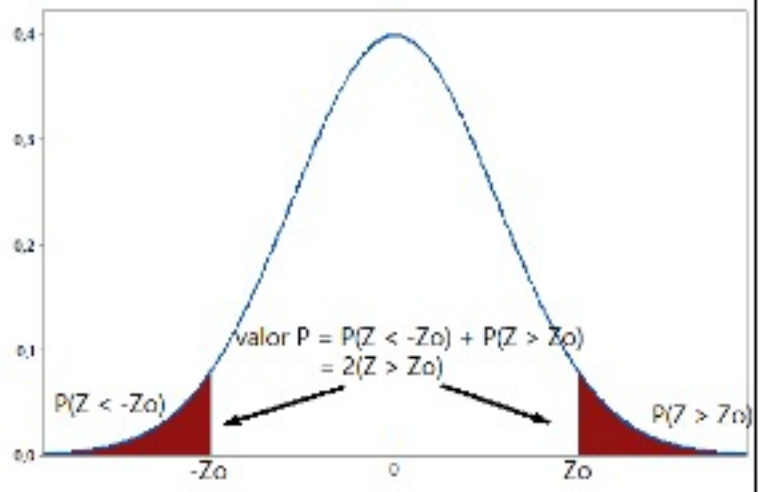
- **El intervalo de confianza bilateral del 95% para la media μ:**

$$IC_{95\%}\mu = \overline{X} \pm Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = 10.89 \pm 1.96\frac{1.5}{\sqrt{9}} = [9.91,\ 11.87]$$

GUIA PARA CALCULAR Y CONCLUIR CON EL VALOR P

Se rechaza la hipótesis nula H_0 si **valor P $< \alpha$**

No se rechaza la hipótesis nula H_0 si **valor P $> \alpha$**

Prueba de cola izquierda	Prueba de cola derecha	Prueba bilateral
$H_1: \mu < \mu_0$	$H_1: \mu > \mu_0$	$H_1: \mu \neq \mu_0$
Valor P $= P(Z < Z_0)$	Valor P $= P(Z > Z_0)$	Valor P $= 2 \times P(Z > Z_0)$
		

EJEMPLO: PRUEBA DE HIPÓTESIS PARA LA MEDIA

Hipótesis estadísticas:

$$H_0 : \mu = 10$$

$$H_1 : \mu > 10$$

Estadístico de prueba:

$$Z_0 = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} = \frac{10.89 - 10}{1.5/\sqrt{9}} = 1.78$$

Calculo del valor P:

Raw Score, **x**	1.78
Population Mean, **μ**	0
Standard Deviation, **σ**	1

Calculate | Clear

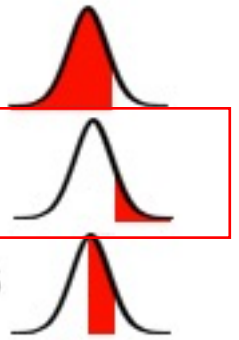
Resultado:

Z-score = 1.78

Probability of x<1.78: 0.96246

Probability of x>1.78: 0.037538

Probability of 0<x<1.78: 0.46246



Conclusión:

- Con un valor P = 0.037
- $\alpha = 0.05$
- Dado que el valor P < α
- Se rechaza H_0
- Se infiere que la oportunidad media de atención es mayor a 10 días

¿PREGUNTAS?