Estadística inferencial

Clase sincrónica

{desafío} latam_

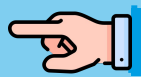
Analiza correlación, causalidad e hipótesis utilizando herramientas de Python.

- **Unidad 1: Estadística descriptiva y probabilidades**
  (Parte I)

  (Parte II)
- **Unidad 2: Variable aleatoria**
  (Parte I)

  (Parte II)
- **Unidad 3:Estadística inferencial**
- **Unidad 4: Regresión**
  (Parte I)

  (Parte II)



Te encuentras aquí

{desafío} latam_

¿Qué aprenderás en esta sesión?

A aplicar los conceptos de estadística inferencial y probabilidad en Python para describir datasets y validar hipótesis

{desafío} latam_

¿Qué entendemos por inferencia?

¿Qué conceptos o procedimientos aplicamos al realizar inferencias??

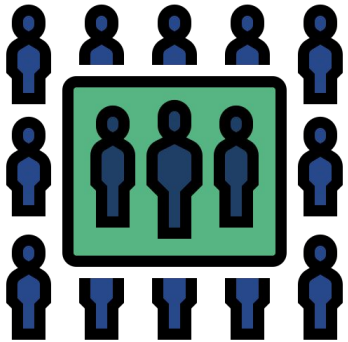
/*Estadística inferencial*/

Estadística inferencial

Descripción v/s inferencia

Hasta el momento hemos utilizado herramientas estadísticas para describir y mostrar los datos, y a partir de ello obtener información y/o conclusiones.

La **estadística inferencial** no busca describir los datos sino que, a partir de una **muestra,** busca describir cómo puede ser la **población**.



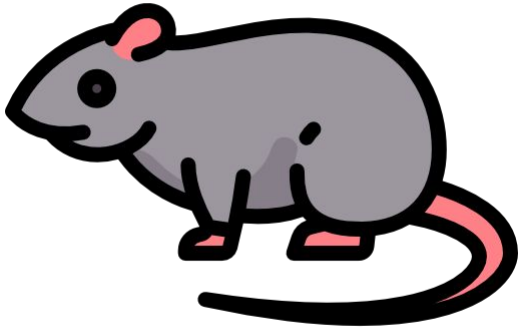
Estadística inferencial

A la población por la muestra

Cuando buscamos hacer inferencia estadística, distinguimos los datos de la población y los de una muestra determinada.

Por ejemplo, queremos estudiar la población de roedores de una región. No podemos observarlos a todos a la vez -o sería demasiado difícil- por lo que debemos obtener muestras que nos permitan inferir características de la población completa.

Para precisar, vamos a denominar de distinta manera los indicadores correspondientes a la población y a la muestra.



Estadística inferencial

Estimadores y parámetros

LLamamos **parámetros** a los indicadores correspondientes a la población, y **estimadores** a los asociados a la muestra.

Nombre	Parámetro estadístico	Estimador muestral
Media	μ	$\overline{x}$
Varianza	σ^2	S^2
Desviación	σ	S
Proporción	p	$\widehat{p}$

Si calculamos el promedio de estaturas de 100 personas adultas de Chile, ¿cómo se relaciona este promedio con el promedio nacional?

Si el promedio de estatura de los adultos en Chile es 1,62 m, y en nuestra muestra obtuvimos 1,73 m… ¿es un hecho raro? ¿muy raro?

{desafío} latam_

/*Estadísticos de prueba*/

Estadísticos de prueba

Definición

Llamamos **estadístico de prueba** a una variable aleatoria de distribución conocida que permite relacionar un indicador estadístico con su estimador respectivo.

Esto nos permitirá hacer inferencia estadística a partir de una muestra de datos, considerando las distribuciones de las variables aleatorias.

Estadísticos de prueba

Casos

Si una variable aleatoria X es tal que $X \sim N(\mu, \sigma)$ y $\{x_1, x_2, ..., x_n\}$ es una muestra aleatoria de X, entonces:

- Si conocemos el valor de σ, tenemos que:

$$\frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$$

- Si no conocemos el valor, podemos utilizar el estimador S y tenemos:

$$\frac{\bar{x} - \mu}{\frac{S}{\sqrt{n}}} \sim t_{n-1}$$

Estadísticos de prueba

Casos

Si X es una variable aleatoria y $\{x_1, x_2, ..., x_n\}$ es una muestra aleatoria de X (con n > 30):

- Si X tiene distribución desconocida y conocemos S, entonces:

$$\frac{\overline{x} - \mu}{\frac{S}{\sqrt{n}}} \approx N(0,1)$$

- SI X tiene distribución Bernouilli con probabilidad p, entonces

$$\frac{\widehat{p} - p}{\sqrt{p\,(1-p)}} \sim N\,(0,1)$$

Estadísticos de prueba

Utilidad

El uso de estadísticos de prueba nos permitirá relacionar los estimadores con sus parámetros desde una perspectiva probabilística, es decir, estimar la probabilidad de que nuestro estimador se acerque o aleje del parámetro en una magnitud dada.

Distribución T-Student

Definición

La distribución t, también conocida como distribución t de Student, es una distribución de probabilidad que se utiliza ampliamente en las pruebas de hipótesis cuando la muestra es pequeña (menor que 30) o cuando la desviación estándar de la población es desconocida.

$$t = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}}$$

t representa el estadístico de prueba t, que se utiliza para evaluar la diferencia entre la media muestral y la media poblacional.

/* Pruebas de Hipótesis */

Prueba de Hipótesis

¿Qué es?

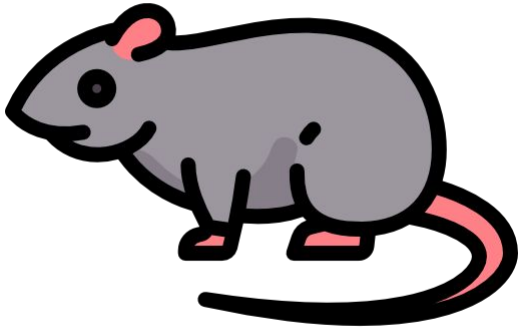
Es un procedimiento estadístico que se utiliza para evaluar si una afirmación sobre una población o un conjunto de datos es compatible con la evidencia observada o si es más razonable rechazarla en favor de una afirmación alternativa.

Pruebas de hipótesis

Aplicación

Volvamos al caso de nuestros roedores. Investigaciones de hace algún tiempo afirman que el peso promedio de los roedores es de 500 gramos, con una desviación estándar de 5 gramos. Como distribución de peso, se asume que distribuye en forma normal.

Obtenemos una muestra de 100 ratones, y se obtiene una media muestral de 496,8 gramos.



Prueba de Hipótesis

Paso 1 : Planteamiento de las hipótesis

Hipótesis nula (H0): Es la afirmación inicial o la hipótesis que se somete a prueba. Por lo general, se asume que no hay efecto o relación entre variables, o que los datos dados son correctos.

> El peso promedio de los ratones es de 500 gramos con desviación de 5 gramos.

Hipótesis alternativa (H1 o Ha): Es la afirmación que se considera como una alternativa a la hipótesis nula. Representa la idea de que hay una relación o efecto específico entre variables, o una duda sobre ellas.

> El peso promedio es diferente de 500 gramos

> El peso promedio es menor que/mayor que 500 gramos

Prueba de Hipótesis

Tipos de situaciones y errores

		Situación real	
		H1	H0
Prueba	Acepta H1	Conclusión correcta Verdadero positivo	Falso positivo Error I
	Acepta H0	Falso Negativo Error II	Concusión correcta Verdadero Negativo

Prueba de Hipótesis

Tipos de situaciones y errores

Naturalmente, no queremos cometer errores de ningún tipo pero estamos trabajando con muestras, por lo que puede ser inevitable.

En general, el error que se busca evitar más es el de tipo 1, es decir, rechazar la hipótesis nula (aceptar la alternativa) cuando esta es verdadera.

Ejemplo: al detectar una enfermedad, H0 corresponde a que el paciente esté sano (H1 es que esté enfermo). Es mejor fallar sobre diagnosticando (se puede hacer contramuestra) que subdiagosticando (el paciente no toma cuidados).

Prueba de Hipótesis

Paso 2 : Elección del nivel de significancia

El **nivel de significancia** (α) representa la probabilidad máxima que se está dispuesto a aceptar para cometer un error de tipo I (rechazar la H0 cuando es verdadera). Es común utilizar valores como 5% (0,05) o 1% (0,01) como niveles de significancia.

Si extraemos una muestra y calculamos un estimador estadístico, este puede alejarse de la media. La magnitud de esta diferencia puede modelarse probabilísticamente (mediante la distribución normal o t-student).

Prueba de Hipótesis

Paso 3 : Selección y cálculo de un estadístico de prueba

La elección del estadístico de prueba depende del tipo de datos y del objetivo de la prueba. Por ejemplo, en nuestro caso, si asumimos que la hipótesis nula es verdadera la media poblacional del peso de los roedores es 500, y la desviación estándar es 5. Por lo tanto, considerando nuestros datos iniciales tenemos que: :

$$\frac{\overline{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$$

$$\frac{\overline{496,8} - 500}{\frac{5}{\sqrt{100}}} = \frac{-3,2}{0,5} = -6,4$$

Prueba de Hipótesis

Paso 4 : Determinación de la región crítica

La región crítica es el área de la distribución de probabilidad donde se encuentran los valores que llevarían a rechazar la hipótesis nula. Estos valores se determinan según el nivel de significancia y la dirección de la prueba (unilateral o bilateral).

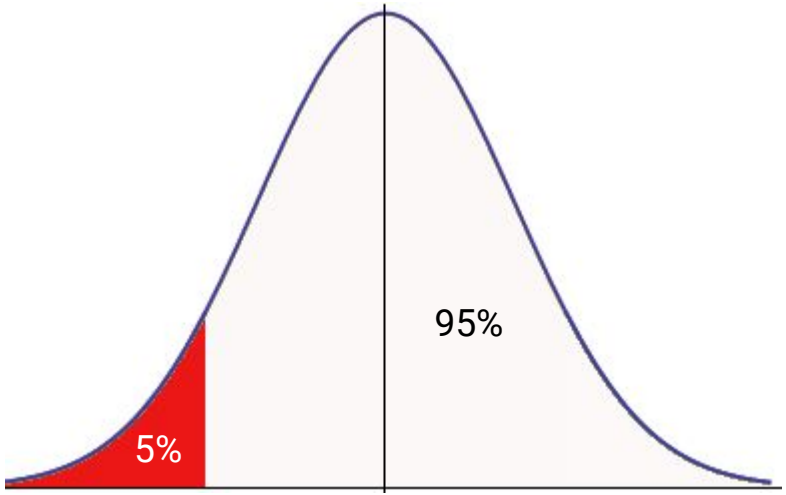
Recordemos: nos dicen que la muestra que hemos obtenido tiene media muestral igual a 496,8. ¿Cuáles podrían ser nuestras hipótesis?

- H0: la media poblacional es 500 gramos
- H1: tenemos posibilidades diferentes:
  - H1-a: la media poblacional es menor (mayor) a 500 gramos
  - H1-b: la media poblacional es diferente a 500 gramos

Prueba de Hipótesis

Paso 4 : Determinación de la región crítica

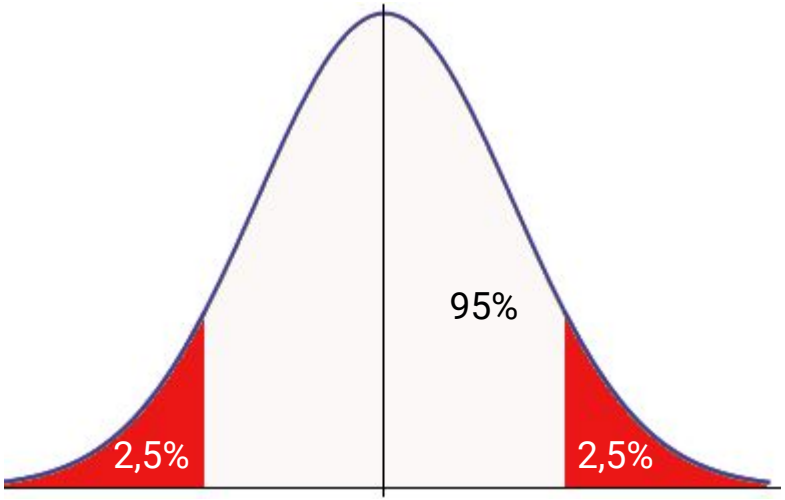
- H1-a: la media poblacional es menor (mayor) a 500 gramos.



Prueba de Hipótesis

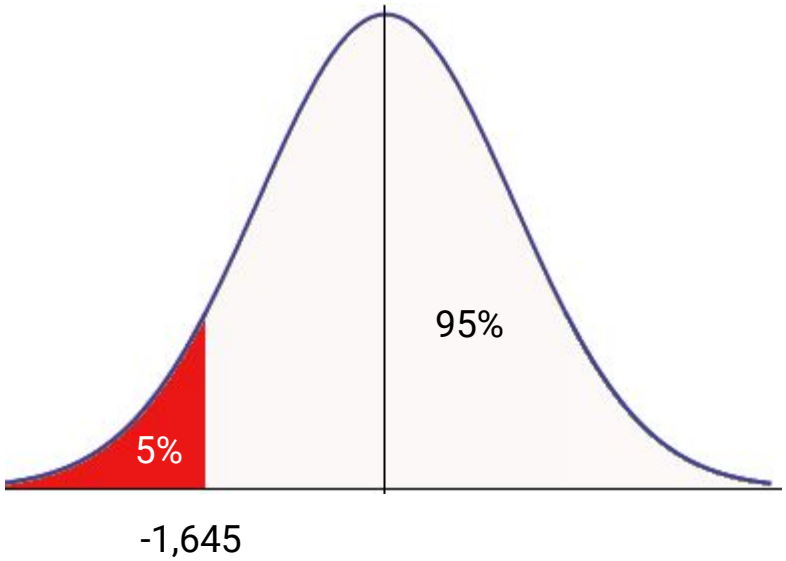
Paso 4 : Determinación de la región crítica

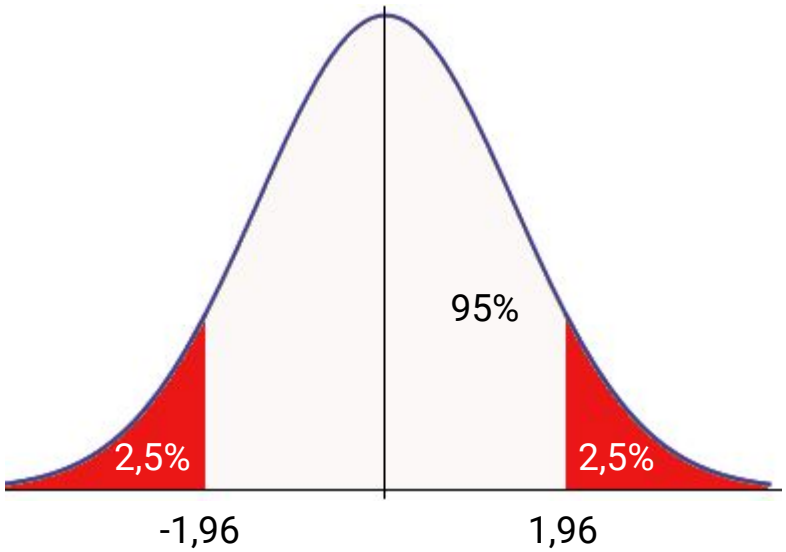
- H1-b: la media poblacional es diferente a 500 gramos.



Prueba de Hipótesis

Paso 4 : Determinación de la región crítica - P-values





Prueba de Hipótesis

Paso 5 : Toma de decisión

Se compara el valor obtenido del estadístico de prueba con los valores de la región crítica. En nuestro caso:

$$\frac{\overline{496,8} - 500}{\frac{5}{\sqrt{100}}} = \frac{-3,2}{0,5} = -6,4$$

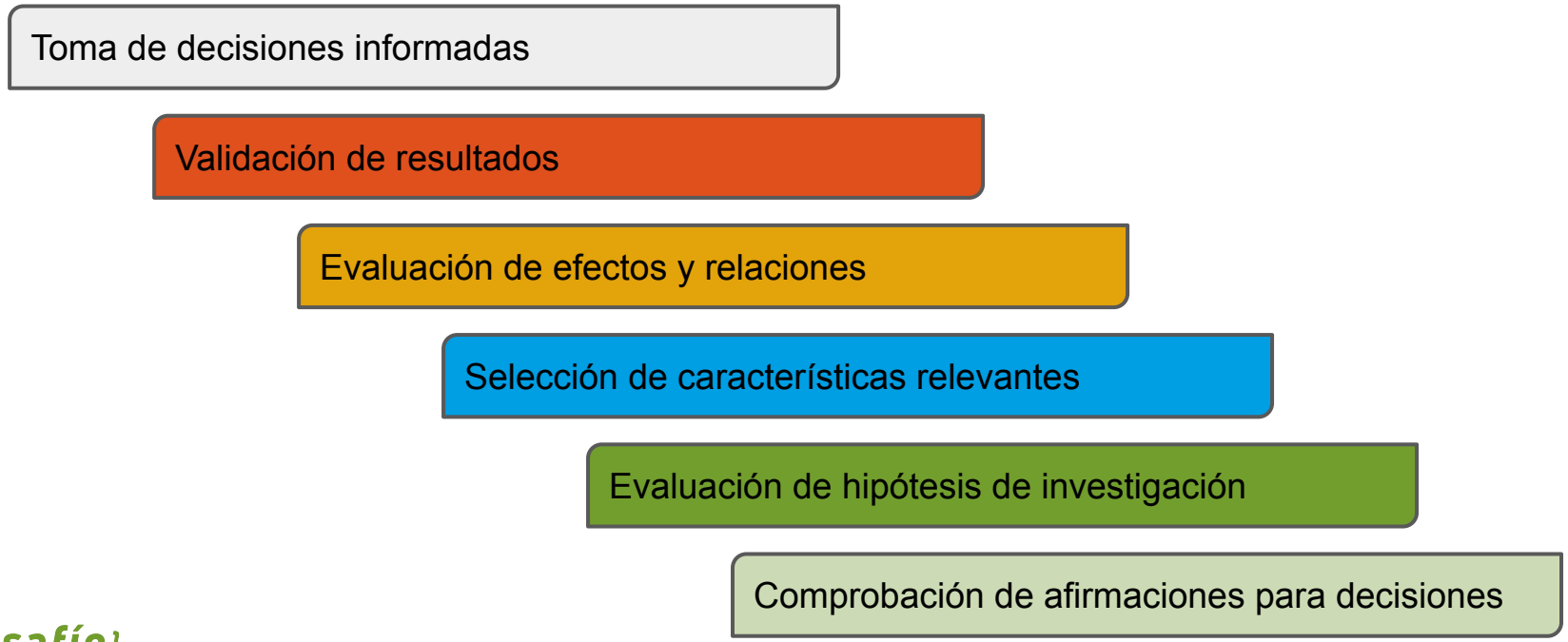
El valor que hemos obtenido para la media, en nuestro caso, se encuentra a más de 1,645 desviaciones estándar de la media, en sentido negativo.

Esto ocurre con probabilidad menor al 5% si la media es la declarada.

Por lo tanto, se RECHAZA la hipótesis nula y se acepta H1, que plantea que la media es menor.

Prueba de Hipótesis

Importancia



P-values (valores p)

Definición

- Se llama p-value a una medida de la probabilidad de obtener los resultados observados, o resultados más extremos, asumiendo que la hipótesis nula es cierta. Es decir, el p-value indica la probabilidad de que los datos observados sean consistentes con la hipótesis nula.

- Un p-value bajo (por ejemplo, $p < 0.05$) indica que los datos observados son poco probables bajo la hipótesis nula, lo que sugiere que la hipótesis nula es poco plausible y se debe rechazar en favor de la hipótesis alternativa.

- Un p-value alto (por ejemplo, $p > 0.05$) indica que los datos observados son bastante probables bajo la hipótesis nula, lo que sugiere que no hay suficiente evidencia para rechazar la hipótesis nula.

Inferencia sobre la media y la proporción

Inferencia sobre la media

Intervalos de confianza

Utilizando esencialmente los mismos conceptos podemos hacer inferencias respecto de la media de una población. Considerando una muestra de tamaño n y un nivel de significancia α determinado, podemos establecer un **intervalo de confianza** para la media de la población a partir de la media muestral, mediante la fórmula

$$\left(\overline{x} - z_{\frac{\alpha}{2}} \cdot \frac{S}{\sqrt{n}}, \overline{x} + z_{\frac{\alpha}{2}} \cdot \frac{S}{\sqrt{n}}\right)$$

donde $z_{\frac{\alpha}{2}}$ corresponde a un valor crítico dado por la distribución de probabilidad que se esté utilizando (existen diferentes métodos para determinarlo).

Inferencia sobre la media

Intervalos de confianza - interpretación

Si llamamos error E a $z_{\frac{\alpha}{2}} \cdot \frac{S}{\sqrt{n}}$, y hemos escogido un nivel de significancia α = 5%, por ejemplo, el intervalo de confianza

$$\left(\overline{x} - z_{\frac{\alpha}{2}} \cdot \frac{S}{\sqrt{n}}, \overline{x} + z_{\frac{\alpha}{2}} \cdot \frac{S}{\sqrt{n}}\right)$$

se puede interpretar como

"podemos afirmar, con un 95% de certeza, que la media poblacional se encuentra entre $\overline{x} - E$ y $\overline{x} + E$ "

Inferencia sobre la media

Intervalos de confianza - interpretación II

Si estamos analizando la estatura promedio de una población y obtenemos que es 162 cm, con un error de 3 cm y significancia del 5%, podemos decir que:

La media de la población es 162 cm, con un margen de error de 3 cm y una confiabilidad del 95%

Inferencia sobre la media

Intervalos de confianza - tamaño de la muestra

A partir de lo anterior, podemos determinar el tamaño de la muestra necesario utilizando la fórmula

$$n = \left(\frac{z_{\frac{\alpha}{2}} \cdot \sigma}{E} \right)^2$$

Inferencia sobre la proporción

Intervalo de confianza

Para estimar la proporción p de la población que presenta una característica determinada, podemos utilizar la proporción muestral $\hat{p}$ con el intervalo de confianza

$$\left(\hat{p} - z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$$

Inferencia sobre la proporción

Intervalo de confianza - tamaño de la muestra

Y de la misma manera, podemos calcular el tamaño de la muestra necesaria mediante la fórmula

$$n = \left(\frac{z_{\frac{\alpha}{2}} \cdot \widehat{p} \cdot \left(1 - \widehat{p}\right)}{E} \right)^2$$

Prueba de hipótesis para muestras independientes

Prueba de hipótesis para muestras independientes

¿En qué consiste?

Una **prueba de hipótesis para muestras independientes** es un tipo de análisis estadístico que se utiliza para comparar las diferencias entre dos grupos o muestras de datos que son independientes entre sí.

Se utiliza cuando se quieren comparar dos grupos distintos, como dos poblaciones diferentes o dos tratamientos diferentes, y se desea determinar si las diferencias observadas entre los grupos son estadísticamente significativas o simplemente el resultado del azar.

Desafío “Inferencia e Hipótesis”

{desafío} latam_

Desafío

"Inferencia e hipótesis"

- Descarga el archivo "Desafío".
- Tiempo de desarrollo asincrónico: desde 2 horas.
- Tipo de desafío: individual.

¡AHORA TE TOCA A TI! 💪

{desafío} latam_

Ideas fuerza

La **estadística inferencial** busca averiguar, a partir de los datos de una **muestra**, las características de una **población**.

Para lograrlo, se utilizan **estimadores** para construir **estadísticos de prueba**, que nos permiten realizar **pruebas de hipótesis.**

Las técnicas de estadística inferencial permiten **estimar la media y la proporción** en una población a partir de una muestra, con **márgenes de error** y confiabilidad, utilizando **intervalos de confianza.**

¿Qué aprendí hoy, que me resulte importante en mi área de trabajo?

Recursos asincrónicos

¡No olvides revisarlos!

Para esta semana deberás revisar:

- Guía de estudio.
- Desafío “Inferencia e hipótesis”.

{desafío} latam_

Próxima sesión...

- *Aplicar los conceptos de correlación y regresión para modelar situaciones.*

{desafío}
latam_

Academia de
talentos digitales