Estadística y Método Científico

Hugo S. Salinas

Fuente: http://dta.utalca.cl/estadistica/

Estadística y Método Científico

Podemos definir Estadística como la ciencia de los datos. La palabra ciencia viene del latín "scientia" que significa conocimiento. El método científico es un conjunto de principios yprocedimientos para la búsqueda sistemática del conocimiento.

El método científico está compuesto por los siguientes pasos:

1. **Formula una teoría** (problema).
2. Recoger **datos** para probar la teoría.
3. **Analizar** los datos.
4. **Interpretar** los resultados y **tomar una decisión**.

El método científico es un procedimiento iterativo de aprendizaje. No podemos tener certeza de la veracidad de las teorías que probemos usando el método científico, eso es inherente a la ciencia. La ciencia está, por lo tanto, todo el tiempo revisando sus teorías.

La Estadística no es un conjunto de diferentes técnicas aisladas unas de otras, sino que la Estadística, en conjunto con el método científico, nos entrega un procedimiento analítico para tomar decisiones.

El lenguaje de toma de decisiones en Estadística

Constantemente buscamos información para tomar decisiones. Al levantarnos en la mañana observamos como está el tiempo para decidir como vestirnos o averiguamos si va haber mucho sol para decidir si llevamos ropa ligera, etc. Inconscientemente aplicamos la lógica estadística para tomar ese tipo de decisiones. Además veremos como podemos usar la estadística para situaciones tan diversas como:

- Analizar si el tabaquismo se asocia al cáncer al pulmón.
- Analizar si la reforma educacional mejoró la calidad de la educación en Chile.
- Predecir los resultados de las próximas elecciones.
- Predecir si ocurrirá una situación de emergencia ambiental en Santiago.

Aprender estadística es como aprender un nuevo idioma.

Definiciones:

Una **población** es el grupo de objetos o individuos bajo estudio, acerca de los cuales queremos obtener información.

Una **muestra** es una parte de la población de la cual se obtiene información.

Inferencia estadística es el proceso de sacar conclusiones acerca de una población basados en información de una muestra de esa población.

¿Debo tomar Aspirina o Migranol para el dolor de cabeza?

Laboratorios A me dice que tome Aspirina Existe teoría (antigua) de que lo mejor es Aspirina

Laboratorios B me dice que tome Migranol Existe teoría (nueva) de que lo mejor es Migranol

Tenemos dos teorías que compiten. En estadística las vamos a llamar **HIPOTESIS**.

El lenguaje de toma de decisiones en Estadística cont.

Definiciones:

La **hipótesis nula**, denotada por H0, es el "status quo", lo convencional, lo que sabemos de la población, lo aceptado hasta el momento.

La **hipótesis alternativa**, denotada por H1, es una alternativa a la hipótesis nula, implica cambio, es lo que el investigador espera que sea cierto.

H0: El nuevo medicamento es tan efectivo como el antiguo.

H1: El nuevo medicamento es más efectivo que el antiguo.

Durante el curso revisaremos los **pasos para tomar una decisión en estadística**, por ahora los podemos resumir en:

1. Tenemos una hipótesis, que la asumimos cierta.
2. Obtenemos datos de un experimento relativo a la hipótesis.
3. Tomamos una decisión acerca de la hipótesis a partir de cuán probable son esos datos proviniendo de la hipótesis.

Definición:

Los datos que obtenemos de una muestra serán **estadísticamente significativos**, si las observaciones son muy poco probables si H0 es verdadera. Si los datos son estadísticamente significativos, entonces rechazamos H0.

El lenguaje de toma de decisiones en Estadística cont.

¿Qué errores podemos cometer?

El sistema de justicia tiene como principio básico que un inculpado en un juicio debe ser considerado inocente hasta que se pruebe lo contrario (que es culpable). ¿Cuáles serán las hipótesis nula y alternativa en el contexto de un juicio?

H0: El inculpado es

H1: El inculpado es

Los abogados defensores y el fiscal presentan sus casos. El juez debe ponderar la evidencia presentada y decidir si es suficiente para declarar culpable al inculpado. Pero el sistema de justicia no es perfecto. Si el juez decide que es culpable y en realidad es inocente, ocurre un error. Si el juez decide que la duda es razonable y lo declara inocente y el inculpado era en realidad culpable, ocurre un error.

Definición:

Error Tipo I: es el error que se comete cuando rechazamos la hipótesis nula (H0) en circunstancia que es la hipótesis verdadera.

Error tipo II: es el error que se comete cuando aceptamos la hipótesis nula (H0) en circunstancia que es la hipótesis falsa.

El lenguaje de toma de decisiones en Estadística cont.

	La hipótesis verdadera	
Su decisión basada en los datos:	H_0 Verdadera	H_1 Verdadera
Rechazar H_0		
No rechazar H_0		

Podemos cometer un Error Tipo I sólo si rechazamos H0 .
Podemos cometer un Error Tipo II sólo **si no** rechazamos H0 .

Ejemplo: **Probando un test o prueba para detectar el cáncer:**

H0: El sujeto está sano
H1: El sujeto padece cáncer

¿Cuáles son los dos tipos que errores que se pueden cometer al decidir entre las dos hipótesis?
Error Tipo I:
Error Tipo II:
¿Cuáles son las consecuencias del error tipo I?
¿Cuáles son las consecuencias del error tipo II?
¿Cuál error es más grave desde un punto de vista ético?

Generalmente queremos proteger lo que ya conocemos reduciendo la probabilidad de cometer el Error Tipo I.

Si el Error Tipo I es considerado grave, ¿por qué no hacerlo cero?.

Una posible manera de no cometer el error tipo I sería no rechazar nunca la hipótesis nula. Pero si hacemos esto nunca vamos a aceptar una teoría nueva. Tenemos que correr un pequeño riesgo.

El lenguaje de toma de decisiones en Estadística cont.

Conceptos claves:

En estadística, la probabilidad de cometer el error tipo I se llama **nivel de significancia** o **nivel de significación** y se denota por la letra griega alfa, α.

La probabilidad de cometer el error tipo II se denota por la letra griega beta, β.

Indicar si son verdaderos o falsos las siguientes proposiciones:

1. Un nivel de significación del 5% significa que, en promedio, 5 de cada 100 veces que rechacemos la hipótesis nula estaremos equivocados.
2. Un nivel de significación del 5% significa que, en promedio, 5 de cada 100 veces que la hipótesis nula es cierta la rechazaremos.

El lenguaje de toma de decisiones en Estadística cont.

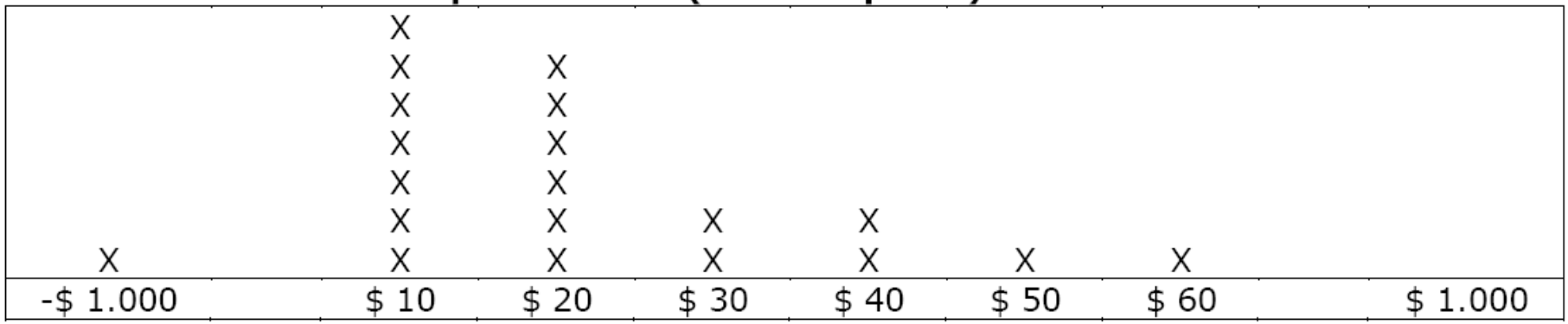
¿Qué hay en la bolsa?

Se tienen dos bolsas idénticas **Bolsa A** y **Bolsa B**. Cada bolsa contiene 20 vales. El contenido de cada bolsa es (en miles de pesos):

Bolsa A	
Valor	Número de vales
-1.000	1
10	7
20	6
30	2
40	2
50	1
60	1

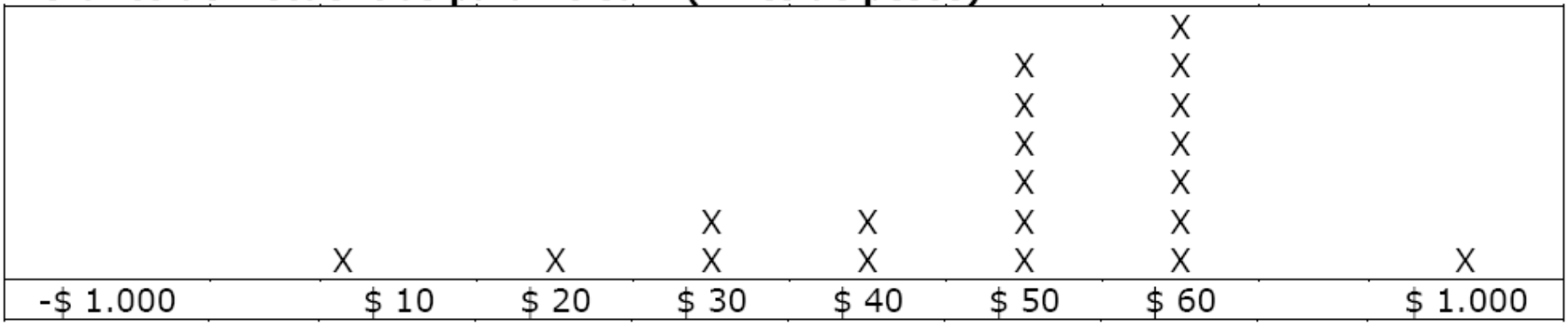
Bolsa B	
Valor	Número de vales
10	1
20	1
30	2
40	2
50	6
60	7
1.000	1

Gráfico de frecuencias para Bolsa A (miles de pesos)



El lenguaje de toma de decisiones en Estadística cont.

Gráfico de frecuencias para Bolsa B (miles de pesos)



BOLSA A TIENE EN TOTAL - $ 560 000

BOLSA B TIENE EN TOTAL + $ 1 890 000

Saquemos un vale al azar de la bolsa que nos muestran. Basándose en una observación (n=1), usted tendrá que decidir entre:

H_0 : *La bolsa que me están mostrando es la Bolsa A*
H_1 : *La bolsa que me están mostrando es la Bolsa B*

¿Cuál es el Error tipo I y Error tipo II?

¿Cómo tomamos la decisión?

Discuta con su vecino las posibles decisiones. ¿En qué situaciones sería fácil tomar una decisión?

El lenguaje de toma de decisiones en Estadística cont.

¿Qué decisión tomaría si el vale es de $60 mil?, ¿Por qué?

¿Qué decisión tomaría si el vale es de $10 mil?, ¿Por qué?

Definición:

Una **regla de decisión** es una regla formal que establece cuando rechazar H_0, basados en los datos.

Examinemos las proporciones en que se encuentran los vales:

Valor	Proporciones en la bolsa A	Proporciones en la bolsa B
- 1.000.000	1/20	0
10.000	7/20	1/20
20.000	6/20	1/20
30.000	2/20	2/20
40.000	2/20	2/20
50.000	1/20	6/20
60.000	1/20	7/20
1.000.000	0	1/20

El lenguaje de toma de decisiones en Estadística cont.

Definición:

La **dirección del extremo** corresponde a la posición de los valores que son más probables bajo H_1 que bajo H_0.

Si los valores más grandes son más probables bajo H_1 que bajo H_0, entonces la dirección del extremo es hacia la **derecha.**

Veremos que no siempre la dirección es hacia la derecha.

Para determinar la regla de decisión usamos primero el valor más extremo.

En nuestro ejemplo es el vale de $60 mil.

Regla de decisión #1:

Rechazar H_0 si el vale seleccionado es de $ 60 mil o más.

Rechazar H_0 si el vale seleccionado es $\geq$ $ 60 mil.

A toda regla de decisión le corresponde una zona de rechazo.

El lenguaje de toma de decisiones en Estadística cont.

Definición:

Una **región de rechazo** es un conjunto de valores para los cuales rechazamos H_0. Cuando el valor no se encuentra en la región de rechazo, decimos que no podemos rechazar H_0.

Un **valor crítico** es el valor que marca el punto inicial del conjunto de valores de la región de rechazo.

Revisemos el ejemplo del juicio. ¿Qué significa rechazar H_0?, ¿Qué significa que no podemos rechazar H_0?

Definición:

Una región de rechazo se llama **unilateral** o de **una cola** si el conjunto de valores extremos están todos en una dirección, ya sea a la derecha (cola superior) o a la izquierda (cola inferior).

Una región de rechazo se llama **bilateral** o de **dos colas** si el conjunto de valores extremos están en las dos direcciones derecha e izquierda.

El lenguaje de toma de decisiones en Estadística cont.

Calculemos ahora α y β para la regla de decisión número 1.

Tabla resumen para α y β resultantes de las tres reglas de decisión:

Regla de Decisión	Región de Rechazo	α	β
# 1: Rechazar H_0 si el vale seleccionado es de \$60 mil o más	\$60 mil o más	0,05	0,60
# 2: Rechazar H_0 si el vale seleccionado es de \$50 mil o más	\$50 mil o más		
# 3: Rechazar H_0 si el vale seleccionado es de \$40 mil o más	\$40 mil o más		

El *valor-p* o ¿cuán raros son los datos?

- Nos vamos a enfocar en lo que los datos observados nos dicen.
- Partimos con el supuesto de que H_0 es verdadera.

Preguntamos:
Si H_0 es verdadera (la Bolsa es la A), ¿cuán probable es obtener el valor observado en la muestra o uno más extremo?

=> Esta probabilidad se llama ***valor-p***.

El lenguaje de toma de decisiones en Estadística cont.

Definición:

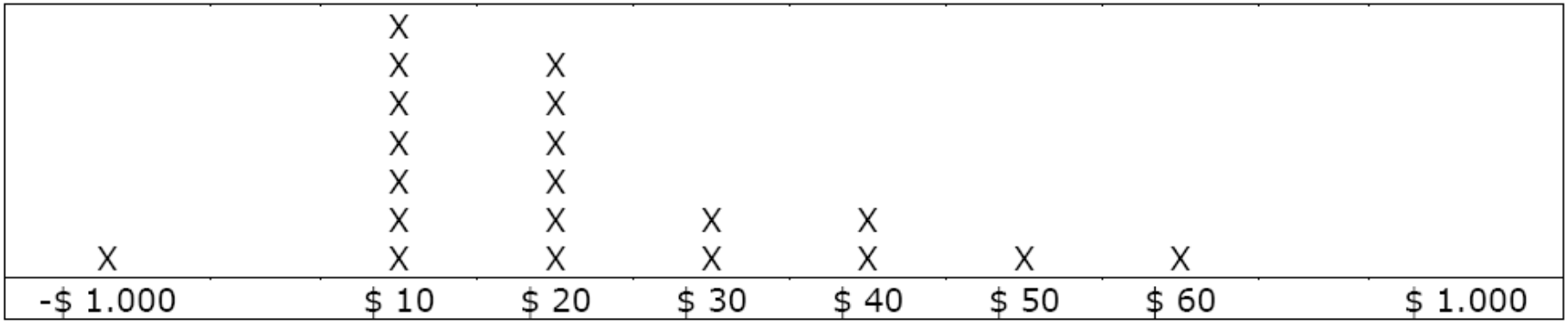
El ***valor-p*** es la probabilidad, calculada bajo el supuesto que H_0 es verdadera, de obtener el valor observado o uno más extremo.

Entendemos que a **menor *valor-p***,
mayor es la evidencia de los datos
en contra de la hipótesis nula H_0 .

Bolsa A versus Bolsa B:

Suponga que usted selecciona un vale de $30 mil y la regla de decisión es la #2. Calcule el *valor-p*.

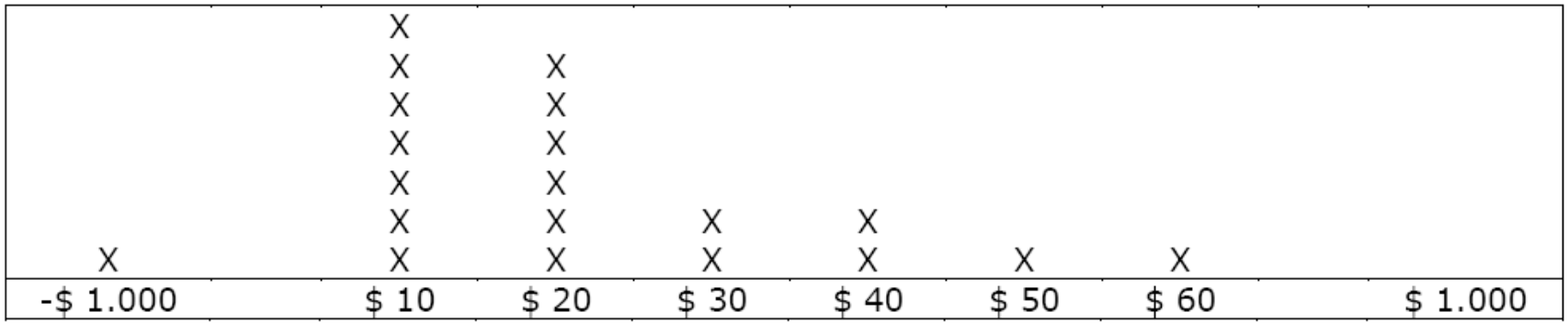
Bolsa A:



El lenguaje de toma de decisiones en Estadística cont.

Suponga que usted selecciona un vale de $60 mil y la regla de decisión es la #2. Calcule el *valor-p*.

Bolsa A:



Relación entre *valor-p* y el nivel de significancia α:

Si el *valor-p* es $\leq \alpha$ ==> rechazamos H_0

Si el *valor-p* es $> \alpha$ ==> no podemos rechazar H_0

El lenguaje de toma de decisiones en Estadística cont.

El nivel de significancia es α = 0,05, probabilidad de cometer el error Tipo I.

La regla de decisión correspondiente es: Rechazar H_0 si el vale seleccionado es \$60 mil o más.

Se selecciona un vale y es \$60 mil. Su decisión es rechazar la hipótesis nula y concluir que los datos son estadísticamente significativos al 5%.

Al rechazar H_0, ¿Puede haber cometido un error?

¿Cuál es la probabilidad de haber cometido ese error?