Notas de clase

Este material está sujeto a correcciones, comentarios y demostraciones adicionales durante el dictado de las clases, no se recomienda su uso a aquellos alumnos que no concurran a las mismas

Prof. Nora Arnesi

Estadística

La estadística es el arte y la ciencia que se dedica al diseño de estudios y al análisis de los datos que dichos estudios producen. Su objetivo final es traducir los datos en conocimiento y así entender el mundo que nos rodea.

En síntesis, la estadística es *el arte y la ciencia de aprender a partir de los datos.*

(Agresti y Franklin, 2013)

Etapas

1. **Recolección de los datos**: es la base sobre la cual se asienta el resto del análisis. Es por esto que es esencial un planeamiento cuidadoso previo a la recolección de los datos. Los métodos de recolección son muy diversos y el investigador en cada caso deberá elegir cuál utilizar de acuerdo a sus objetivos, disponibilidad de la información, etc.

2. **Presentación de los datos**: los datos recolectados pueden ser presentado en forma tabular o gráfica, de manera de obtener una primera descripción de las características que presentan.

3. **Análisis de los datos**: los datos deben ser analizados cuidadosamente para la realización de ***inferencias***.

4. **Interpretación de los datos**: la etapa final es la elaboración de conclusiones a partir del análisis realizado de manera de poder obtener una conclusión válida que dé respuesta al problema que originó dicho análisis.

La estadística puede dividirse en dos grandes ramas…

- **estadística descriptiva** ⟶ se concluye exclusivamente sobre el conjunto de datos analizados

- **estadística inferencial** ⟶ a partir de los datos analizados (muestra) se concluye respecto a la población a partir de la cual fueron obtenidos los datos

Estadística descriptiva y estadística inferencial

La descripción en el análisis estadístico

La **estadística descriptiva** abarca los métodos para resumir la información recolectada. Al resumir, se debe intentar perder la menor cantidad de información posible. El análisis descriptivo combina gráficos y medidas de resumen numéricas como promedios y porcentajes.

La inferencia en el análisis estadístico

La **estadística inferencial** abarca los métodos para tomar decisiones o realizar predicciones acerca de una población, en base a los datos obtenidos a partir de una muestra de tal población.

- La estadística descriptiva se utiliza tanto en **poblaciones** como en **muestras**.

- La estadística inferencial se utiliza cuando los datos sólo están disponibles para una muestra pero necesitamos sacar conclusiones sobre la población total.

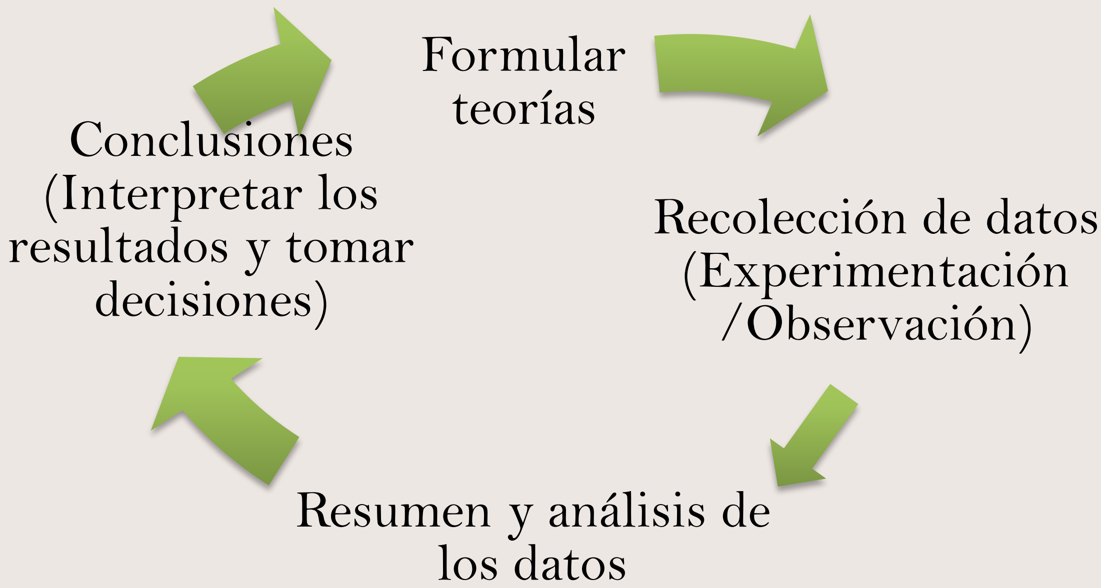
La **población** es el conjunto de todas las unidades en las que estamos interesados, de las cuales queremos obtener conclusiones (en la próxima unidad ampliaremos esta definición)

Una **muestra** es un subconjunto de la población para el cual tenemos (o planeamos tener) datos.

La **unidad de análisis** es el elemento mínimo de una población. Se refiere a qué o quién es objeto de interés en una investigación.

Estadística y el método científico

A partir de las definiciones presentadas surge de manera natural la vinculación con el **método científico**, el cual comprende un conjunto de principios y procedimientos para la obtención sistemática del conocimiento .



Estadística y el método científico

La **estadística** y el **método científico** nos proveen un conjunto de principios y procedimientos para obtener y resumir la información para la **toma de decisiones**.

“Planteo de una hipótesis”

Un problema que plantea la investigación científica es el de decidir la “veracidad” o “falsedad” de una hipótesis.

“*…una hipótesis es una conjetura, una afirmación de carácter incierto que se propone sin conocimiento de su verdad o falsedad…*” (Klimovsky, 1997).

La estadística provee métodos adecuados para establecer la “veracidad” o “falsedad” de una hipótesis y es por este motivo que está tan vinculada al método científico.

¿Por qué usamos comillas?

En toda prueba estadística de hipótesis, se plantean 2 teorías que compiten.
Los test de hipótesis corresponden a procedimientos de inferencia estadística que enfrentan un problema de decisión en base a la información parcial brindada por una muestra.

Hipótesis Estadísticas:
H_0) **Hipótesis nula**: es la afirmación de que nada está sucediendo, que no existe diferencia, que no hay cambios en la población.
H_1) **Hipótesis alternativa**: es la afirmación que el investigador espera que sea cierta; representa el cambio en la población que el investigador está buscando.

Resuelve!!! 1.1 - ¿Dados regulares?

En un famoso experimento de dados de 315.672 tiradas, 106.656 resultaron 5 ó 6.
Si los dados son regulares la verdadera proporción de 5 ó 6 es 2/6.
Sin embargo, al analizar los datos se observa que los puntos son hechos por pequeñas muescas o depresiones en la cara del dado.
Los lados con 5 y 6 tienen más muescas que las otras caras y así estos lados son más livianos que las otras caras, lo cual sugiere que la verdadera proporción de 5 ó 6 puede ser más alta que el valor regular 1/3.

- Establezca las hipótesis H_0 y H_1 apropiadas.

En general la hipótesis nula es construida para refutarla, con el objetivo de apoyar la hipótesis alternativa.

Una manera simple de entender este proceso es considerando que la hipótesis nula es puesta bajo juicio.

"Tal como un acusado se presume inocente hasta que se demuestra que es culpable, la hipótesis nula se asume correcta (inocente) hasta que es rechazada (hasta que se prueba que es culpable) a favor de alguna alternativa más plausible.

La carga de prueba corresponde al investigador, como el fiscal debe demostrar la culpabilidad de la nula contra una comunidad científica escéptica, el jurado. El investigador es también un detective."

(Rodgers, 2010).

"Demostración o refutación"

¿Cómo tomar una decisión con estadística?

Para conocer cuál de las dos teorías parece más razonable, recabamos información, la analizamos y nos preguntamos:

¿son estos datos más probables de ser observados si la 1º teoría es cierta o si la 2º teoría es cierta?

Una teoría será rechazada si se puede demostrar estadísticamente que los datos que observamos son muy poco probables de ocurrir si fuera cierta.

Si los datos son poco probables de ocurrir cuando la primer teoría es cierta

⇩

Rechazamos esta teoría (...y sustentamos la otra)

Tener en cuenta que el **no rechazo** de la H_0 **no implica** necesariamente que esta teoría sea cierta.

La lógica detrás de la toma de decisión esta basada en el concepto de **suceso raro**. Dado que la H_0 es en general, el *statu quo*, comenzamos **suponiendo que la H_0 es cierta.**

En diversas fuentes frecuentemente se leen frases como:

"Los resultados no fueron estadísticamente significativos",

o bien,

"No hay diferencias estadísticamente significativas entre los grupos".

Los datos observados conducen a un test **estadísticamente significativo** si ellos son poco probables de ser observados bajo el supuesto de que H_0 es cierta.

Es decir, si rechazamos H_0, decimos que los resultados son *estadísticamente significativos.*

Ejemplo

Tratamiento de una enfermedad mental

Supongamos que se realiza un estudio para analizar si cierta terapia para tratar depresión mental es efectiva. Si se determinara que su efecto es beneficioso, se comenzarían a tratar a los pacientes con dicha terapia.

Consideremos las hipótesis:

H_0) La terapia no tiene ningún efecto.
H_1) La terapia tiene un efecto beneficioso.

Supongamos que al finalizar el estudio se concluye que los datos son estadísticamente significativos.

- ¿Qué hipótesis fue rechazada?
- ¿Qué acción se tomó?

¿Qué errores podemos cometer?

Un principio del sistema de la justicia es que

"El acusado será considerado inocente hasta que se pruebe su culpabilidad"

En el contexto de un juicio criminal ¿cómo juegan las hipótesis nula y alternativa?

La H_0 es el *statu quo*, el acusado es inocente. Se deberán presentar las evidencias y se evaluarán.
Si hay suficientes pruebas contra el acusado, será declarado "culpable".

Si se dictamina un veredicto "culpable" y el acusado es "inocente" ocurrirá un ***ERROR***.

Si se dictamina un veredicto "inocente" y el acusado es "culpable" ocurrirá un ***ERROR***.

En términos estadísticos...

Error tipo I (e_I): error que se comete cuando se rechaza la Ho siendo cierta.

Error tipo II (e_{II}): error que se comete cuando no se rechaza la H_0 siendo cierta la H_1.

Decisión basada en los datos	H_0 es cierta	H_1 es cierta
No rechazar H_0	No hay error	e_{II}
Rechazar H_0	e_I	No hay error

Ejemplo

Tratamiento de una enfermedad mental

- ¿Cuáles son los errores que se pueden cometer cuando se decide entre las dos hipótesis planteadas?
- ¿Cuáles son sus consecuencias?

A cada uno de los errores posibles, se les puede asignar una probabilidad de cometerlos…

- $P(e_I)$: probabilidad de rechazar H_0 siendo cierta

  Se lee como: "*probabilidad de error de tipo I*" y se simboliza con la letra griega α.

- $P(e_{II})$: probabilidad de no rechazar H_0 siendo cierta H_1

  Se lee como: "*probabilidad de error de tipo II*" y se simboliza con la letra griega β.

En un mundo ideal, los errores no ocurrirían.

Sin embargo, ya sea en test de hipótesis o en situaciones tales como juicios o en diagnósticos médicos, los errores ocurren.

Dado que ambas, $P(e_I)$ y $P(e_{II})$, representan probabilidades de cometer un error, idealmente deseamos que sean lo más pequeñas posible.

Importante!

la probabilidad es un valor entre 0 y 1 !!!

Para lograr un valor 0 en la probabilidad de cometer un e_I, nunca rechazaríamos la H_0, nunca sustentaríamos una teoría nueva o alternativa. Por lo tanto debemos aceptar una pequeña probabilidad de cometer un error.

Este razonamiento refleja una relación fundamental entre la $P(e_I)$ y la $P(e_{II})$:

para un tamaño de muestra fijo (n),
mientras $P(e_I)$ disminuye, la $P(e_{II})$ aumenta.

Ambas probabilidades están inversamente relacionadas.

Si *n* se incrementa, aportamos mayor información y ambas probabilidades disminuyen.

Se desea encontrar un test de manera tal que las $P(e_I)$ y $P(e_{II})$ sean mínimas.

Se acostumbra asignar un límite a la

$$P(\text{rechazar } H_0/H_0 \text{ es cierta})$$

Se lee como "*probabilidad de rechazar H_0 dado que H_0 es cierta*"

y minimizar la otra probabilidad.

Es decir, seleccionamos un número $0 < \alpha < 1$, llamado **nivel de significación** e imponemos la condición:

$$P(\text{rechazar } H_0/H_0 \text{ es cierta}) \leq \alpha.$$

Resuelve!!! 1.5 – Suplemento de cromo

En un estudio realizado por el Colegio Americano de Medicina Deportiva se monitorearon 16 hombres; a la mitad se les administró suplemento de cromo y a la otra mitad un falso sustituto (placebo).

El objetivo era analizar si los hombres que entrenaban mientras tomaban suplementos de cromo aumentaban su fuerza en mayor medida que aquellos que entrenan sin ingerir tal suplemento.

Al final del programa de entrenamiento, el estudio concluyó:

"*Ambos grupos ganaron en fuerza en forma similar; no hubo diferencias significativas en el incremento de la fuerza*".

Resuelve!!! 1.5 – Suplemento de cromo

Considere las siguientes hipótesis:

H_0) no hay diferencias en el aumento de fuerza entre los dos grupos
H_1) el grupo con suplemento de cromo presenta mayor aumento en fuerza

Basándose en los resultados:

- ¿qué hipótesis es la sustentada por los datos?
- ¿qué puede decir acerca de la magnitud del *p-value* en función de la conclusión a la que se arribó?
- ¿se pudo haber cometido algún tipo de error?
- explique las posibles consecuencias del mismo.

Algunas definiciones...

- **Región de rechazo** o **región crítica**: conjunto de valores para los cuales rechazaríamos H_0 (valores que son contradictorios a la H_0 y favorables a la H_1).

- **Región de aceptación**: conjunto de valores para los cuales no rechazaríamos la H_0.

- **Valor crítico**: aquel que marca el "punto inicial" del conjunto de valores que comprenden la región de rechazo.

- **Potencia del test**: $1-\beta$.

Un experimento simple para ejemplificar…

Supongamos que se nos presentan 2 bolsas, y se nos pide que elijamos una.

- La bolsa A tiene un total de $-560, es decir que si la elegimos, perdemos dinero.
- La bolsa B tiene un total de $1890, es decir que si la elegimos, ganamos dinero.

El experimento consiste en recibir una bolsa, extraer un billete y decidir si el mismo proviene de la bolsa A o de la B.

Lógicamente, si concluimos a través de la información dada por el billete que la bolsa recibida es la A, la cambiaremos.

Gráfico de frecuencias de la bolsa A:

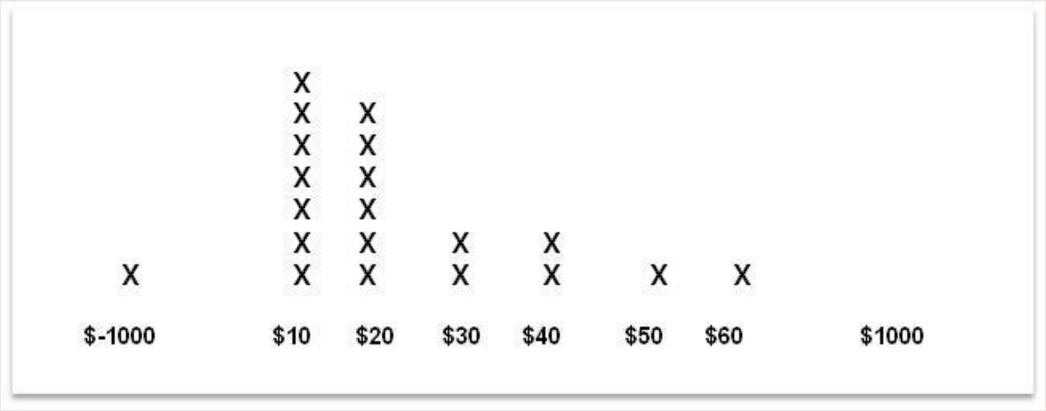
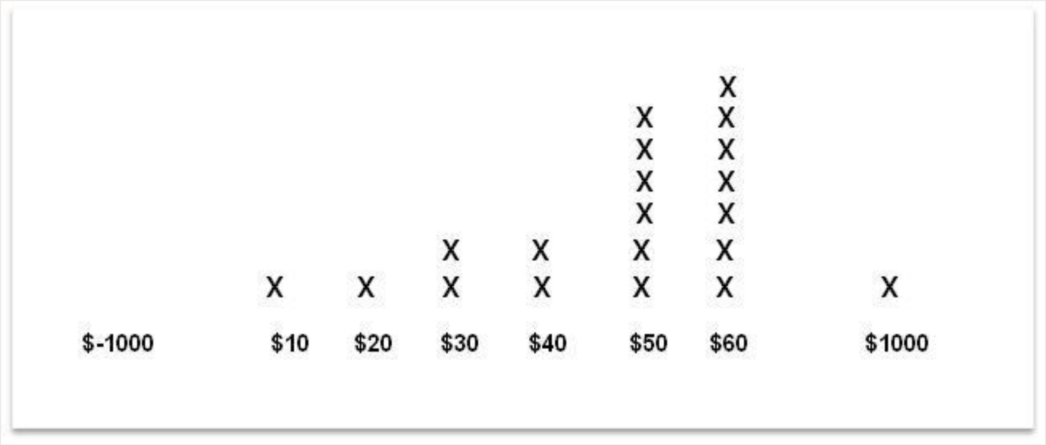


Gráfico de frecuencias de la bolsa B:



H_0) La bolsa mostrada es la A.
H_1) La bolsa mostrada no es la A.

¿Cómo establecemos la regla de decisión
en base a una observación (n=1)?

Consideremos dos casos obvios:

- Si el billete seleccionado vale $-1000, sabremos que la bolsa es la A, por lo tanto no rechazaremos H_0.

- Si el billete seleccionado vale $1000, sabremos que la bolsa es la B, por lo tanto rechazaremos H_0.

Pero, ¿qué pasa si el billete vale $60? O si vale $10?

Regla de decisión	Región de rechazo	Valor crítico	$P(e_I)$	$P(e_{II})$
#1 - Rechazar H_0 si el valor elegido es ≥\$60.	Valores mayores o iguales que \$60.	\$60		
#2 - Rechazar H_0 si el valor elegido es ≥\$50.	Valores mayores o iguales que \$50.	\$50		
#3 - Rechazar H_0 si el valor elegido es ≥\$40.	Valores mayores o iguales que \$40.	\$40		

Dada una regla de decisión, podemos hallar los niveles de α y β exactamente.

También podemos ir al revés:

Fijar el nivel de significación α y a partir de él, determinar la regla de decisión.

¿Cuán raros son los datos si H_0 es cierta?

El valor de la **probabilidad asociada** a un resultado (*p*) es la probabilidad de obtener el resultado observado o uno más extremo (en dirección a la H_1), suponiendo que la H_0 es la verdadera.

Cuanto menor es el valor *p*, mayor es la evidencia provista por los datos en contra de H_0.

El valor *p* se compara con el nivel de significación, α, requerido para la toma de decisiones.

Relación entre la probabilidad asociada y el nivel de significación α

- Si $\mathbf{p} \leq \boldsymbol{\alpha} \Rightarrow$ se **rechaza** H_0 y se dice que los resultados *son estadísticamente significativos.*

- Si $\mathbf{p} > \boldsymbol{\alpha} \Rightarrow$ no se **rechaza** H_0 y se dice que los resultados *no son estadísticamente significativos.*

Un experimento simple para ejemplificar…

Supongamos que se extrae un billete y su valor es $50.
Si se considera la regla #1:

- ¿Cuál es el valor de la probabilidad asociada a este resultado?
- ¿Cuál es la conclusión a partir de este valor?
- ¿Puede estar cometiendo un error al tomar tal decisión?
- ¿Qué tipo de error puede cometer?
- ¿ Cuál es la probabilidad de haber cometido ese error?

Es importante distinguir entre:

- establecer una regla de decisión: "*antes de observar los datos*"

  y

- tomar una decisión "*luego de observar los datos*".

Una vez que se ha tomado una decisión, ésta será correcta o errónea y la probabilidad de cometer una equivocación será igual a 0 ó 1.

✓ Resuelve!!! 1.6 – Tres estudios

Estudio	H_0	H_1	*p*
A	La verdadera duración de vida promedio es ≥ 54 meses.	La verdadera duración de vida promedio es < 54 meses.	0,0251
B	La experiencia de supervivencia bajo el tratamiento A es igual a la obtenida con el tratamiento B.	La experiencia de supervivencia bajo el tratamiento A es diferente a la obtenida con el tratamiento B.	0,0018
C	La verdadera proporción de personas que tienen dos empleos es ≤ 0,33.	La verdadera proporción de personas que tienen dos empleos es > 0,33.	0,3590

- Para cuál de los estudios los resultados muestran mayor soporte para la H_0? Explique.

- Suponga que en el estudio A se concluyó de manera equivocada, en el lenguaje estadístico, éste es un error tipo I o un error tipo II?

- En el estudio C, cuál es la hipótesis sustentada?

Acerca de la dirección del extremo

En el ejemplo de las bolsas A y B, la dirección de los valores extremos fue a la derecha → **Test unilateral por derecha**

Pueden presentarse otros casos, dependiendo de la posición de los valores que son más probables bajo H_1 que bajo H_0:

- **Test unilateral por izquierda**
- **Test bilateral**

Gráfico de frecuencias de la bolsa A

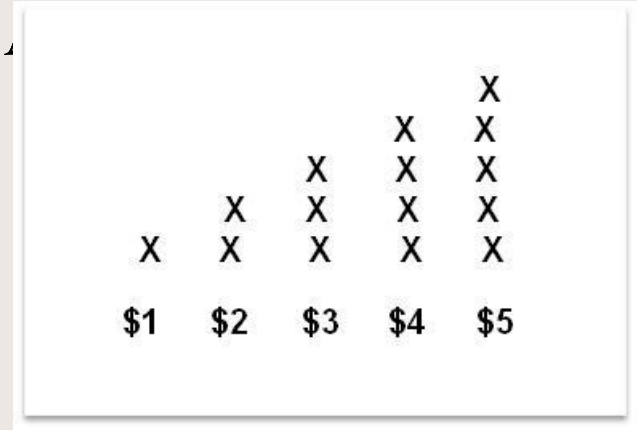
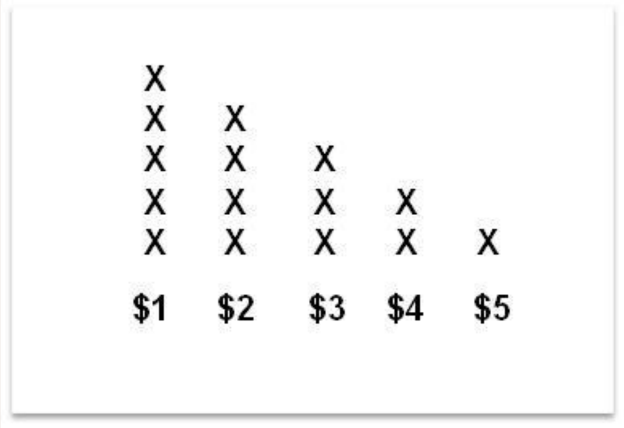


Gráfico de frecuencias de la bolsa B:



Ejercicio:

Se deja propuesto trabajar de manera análoga a lo resuelto para el caso unilateral a la derecha.

Gráfico de frecuencias de la bolsa A:

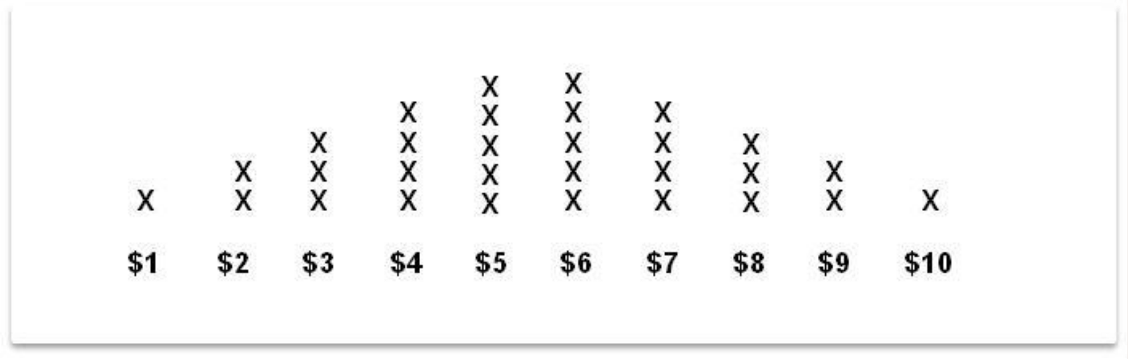
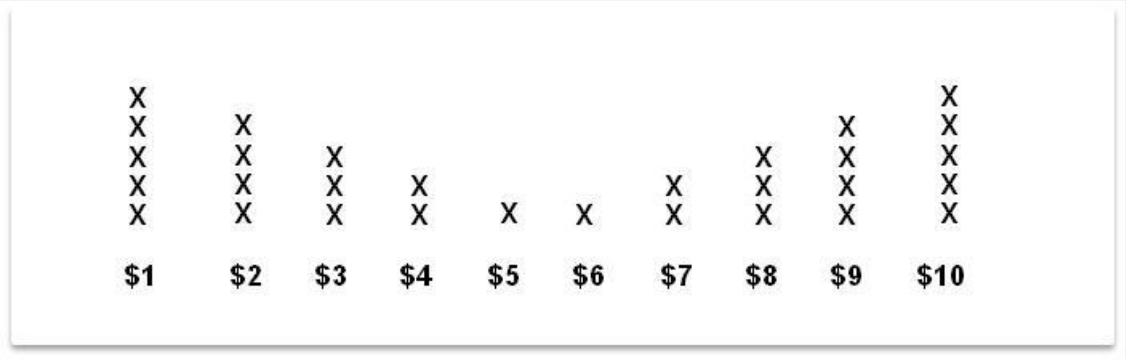


Gráfico de frecuencias de la bolsa B:



En este caso las reglas de decisión incluyen valores extremos en ambas direcciones.

Regla de decisión	Región de rechazo	Valor crítico	$P(e_I)$	$P(e_{II})$
#1 - Rechazar H_0 si el valor elegido es ≤$1 o bien ≥$10.	Valores menores o iguales que $1 y valores mayores o iguales que $10.	$1 y $10		
#2 - Rechazar H_0 si el valor elegido es ≤$2 o bien ≥$9.	Valores menores o iguales que $2 y valores mayores o iguales que $9.	$2 y $9		

- Si se extrae un billete de $10, ¿cuál es su probabilidad asociada?

BIBLIOGRAFÍA

- Agresti A., Franklin C. (2009) The art and science of learning from data. 2º Ed. New Jersey. Pearson Prentice Hall.
- Cerda J. L. y Valdivia C. (2007) John Snow, la epidemia de cólera y el nacimiento de la epidemiología moderna. https://scielo.conicyt.cl/scielo.php?script=sci_arttext&pid=S0716-10182007000400014 Consultado: marzo 2019.
- Patiño Herrera R. (2002). Introducción al análisis estadístico. Instituto Tecnológico de Celaya. http://www.iqcelaya.itc.mx/~roosph/pye/U1/eu1t1.pdf. Consultado: julio 2014.
- Rodgers J.L. (2010). The Epistemology of Mathematical and Statistical Modeling. *A Quiet Methodological Revolution. University of Oklahoma.* American PsychologistAssociation. Vol. 65, No. 1, 1–12 DOI: 10.1037/a0018326.
- Ruggieri M. (2010) Métodos estadísticos I. Reimpresión. Rosario. UNR editora.
- Varalakshmi V., Suseela N., GnanaSundaram G., Ezhilarasi S. Indrani B. (2004). Statistics Higher Secondary – First Year. 1st Ed. India: Government of Tamilnadu. Tamilnadu Textbook Corporation.