Regularidad estadística.

En virtud de la gran variabilidad de muchos procesos, se recurre al estudio del comportamiento en grandes conjuntos de elementos.

- Se busca captar los aspectos sistemáticos y no los aleatorios.

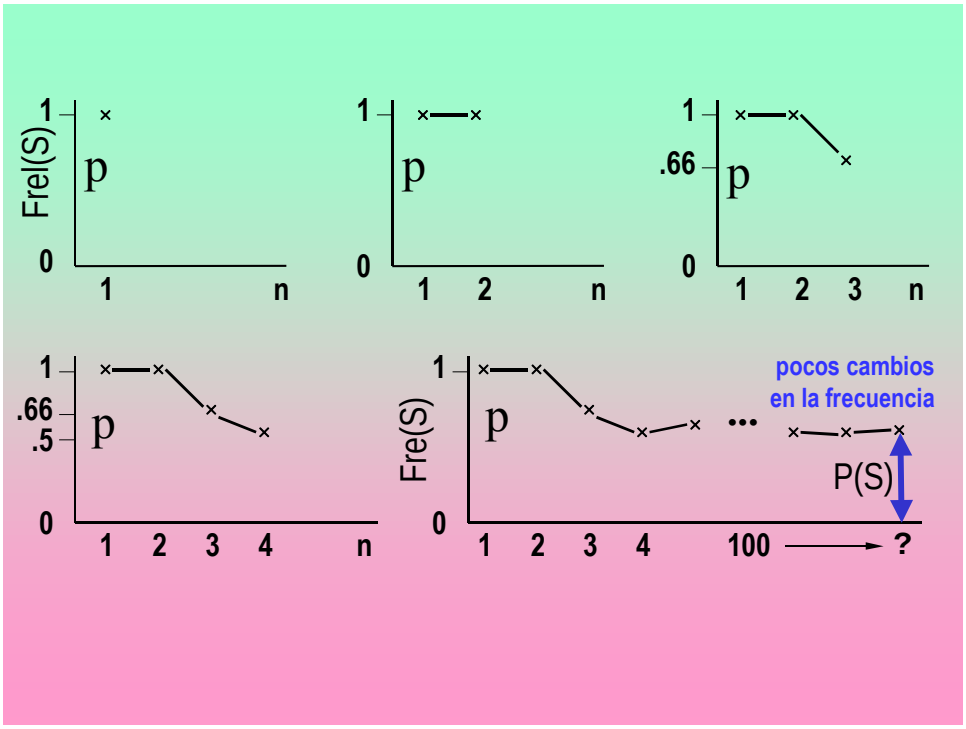
Se pretende determinar lo que ocurrirá casi con seguridad en grandes grupos de elementos, aunque sea impredecible la ocurrencia de un resultado particular. Esto es posible en virtud de la llamada regularidad estadística

La regularidad estadística consiste en el hecho universalmente observado, que funciona como un supuesto muy apoyado, es que: al estudiar un número grande de veces un fenómeno en condiciones (casi) constantes, las proporciones en las que ocurren los posibles resultados son muy estables.

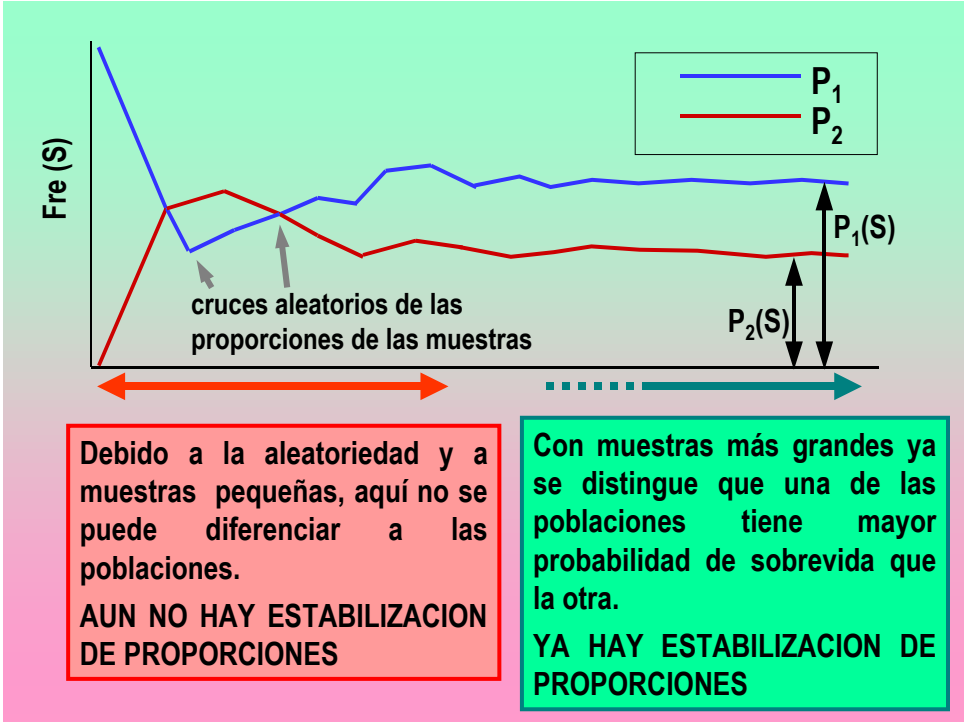
El valor en el que se estabilizan las proporciones se le conceptualiza como la **probabilidad**

Consideremos una operación de corazón abierto y que registramos la sobrevida a más de 5 años del paciente:
El resultado de un paciente o grupos pequeños de ellos es **impredecible**, sin embargo al estudiar muchos pacientes semejantes, con la misma técnica quirúrgica **la proporción** de sobrevida a 5 años **casi no cambia.**

Sea Frel(s) la frecuencia relativa de sobrevida es decir el cociente del número de casos con sobrevida, entre el número de estudiados.



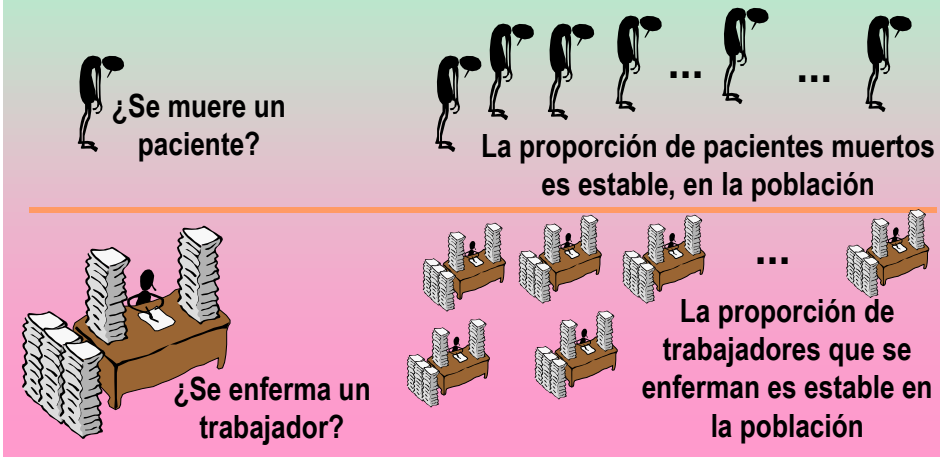
Considérese ahora dos poblaciones de pacientes que difieren en un factor de riesgo (edad, padecimientos agregados, etc.) o bien dos técnicas diferentes. ¿Cambia la probabilidad de sobrevida?



REGULARIDAD ESTADÍSTICA

Al estudiar un fenómeno aleatorio muchas veces, en condiciones casi constantes (población), los diferentes resultados ocurren con una proporción estable.

A esa proporción le llamamos **probabilidad** de cada resultado.



Regularidad estadística, base de la probabilidad frecuentista

- Al estudiar un fenómeno muchas veces en condiciones constantes o casi (la población), la frecuencia de los posibles resultados es muy estable.
- La definición de los resultados de interés (espacio muestral) y las condiciones de estudio (población) es subjetiva, sin embargo, los valores en los que se estabilizan las frecuencias relativas o probabilidades son objetivos.
- Para entender, describir y predecir fenómenos aleatorios, se pretende conocer esas probabilidades

Uso de modelos en la regularidad estadística

Para describir, entender y predecir los fenómenos aleatorios, frecuentemente se recure a postular modelos probabilísticos.

Estos pueden haber surgido por tres vías:

1. Experiencias empíricas previas.
2. Consideraciones teóricas sobre la naturaleza del fenómeno estudiado, y
3. Combinaciones de las dos anteriores.

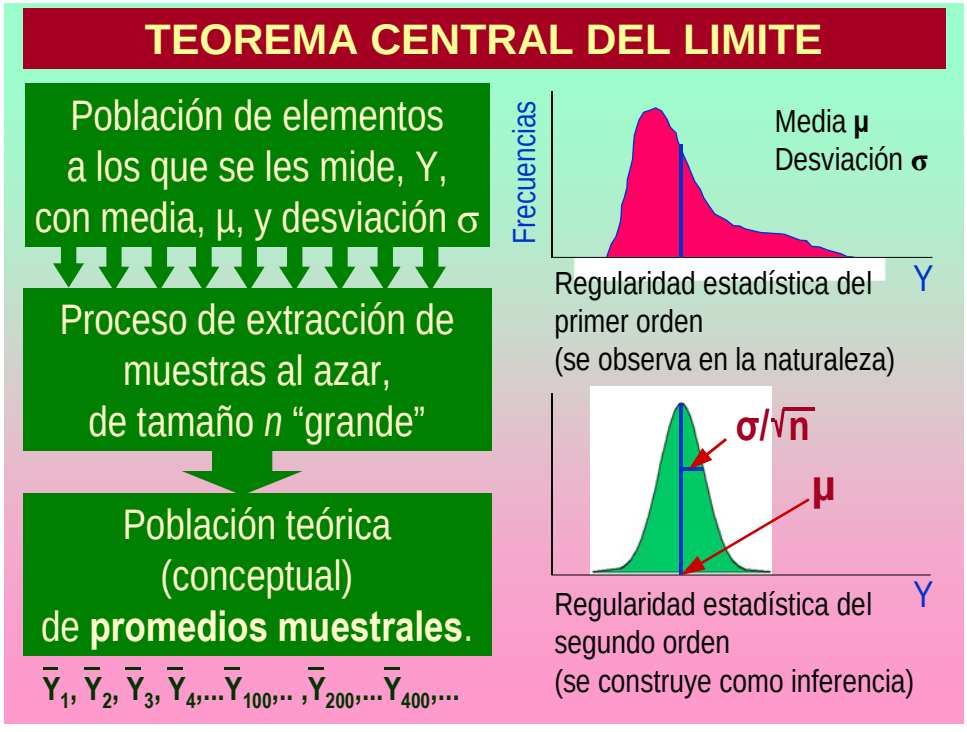
•**INFERENCIA ESTADISTICA**• Se usan **modelos matemáticos** para describir (modelar) la regularidad estadística de los resultados, es decir las probabilidades de ellos. Los más comunes son: binomial, **normal**, poisson, etc. Estos modelos quedan caracterizados por **parámetros** como la media (μ), la proporción (p), o desviación estándar poblacional (σ).

TEOREMA CENTRAL DEL LIMITE

Hay varias versiones matemáticas del teorema, sin embargo, se enuncia la mas sencilla.
Si se tiene una población de elementos a los que se les mide una característica numérica: "y"; ***no se requiere que la regularidad estadística de "y", en la población se pueda modelar con la normal.*** Al considerar la posibilidad de repetir la toma de muestras en las mismas condiciones y del mismo tamaño, n, se tendría un número muy grande de muestras distintas y en cada una su media muestral.

El teorema dice que si el tamaño de muestra es grande, entonces la regularidad estadística de la población de medias muestrales, es normal con la misma media poblacional, **μ**, de la población original, y con una varianza que es la varianza original dividida entre el tamaño de muestra n (o equivalentemente con una desviación estándar que es la desviación estándar original dividida entre la raíz cuadrada del tamaño de muestra).

$$\bar{y} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$



Datos cuantitativos.

Medidas descriptivas:

(1) ***Medidas de tendencia central.*** Ayudan a encontrar el centro de la distribución de frecuencias relativas.
(2) ***Medidas de variación.*** Miden la dispersión.
(3) ***Medidas de posición relativa.*** Describen la posición relativa de una observación dentro de un conjunto de datos.

1. Medidas de tendencia central.

Media (promedio)	Mediana	Moda
Sea $y_1, y_2, ..., y_n$ un conjunto de n mediciones, definimos la media muestral como: $\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$	Es el número de en medio cuando las n observaciones se acomodan en orden. Denotamos como $y_{(i)}$ el i-ésimo valor de y cuando la muestra de n observaciones se acomoda en orden ascendente. $mediana = \begin{cases} y_{\left(\frac{n+1}{2}\right)} & para\ n\ impar \\ \frac{y_{\left(\frac{n}{2}\right)} + y_{\left(\frac{n}{2}+1\right)}}{2} & para\ n\ par \end{cases}$	Es el valor de y que ocurre con mayor frecuencia

La media es sensible a observaciones muy pequeñas o grandes (determinaciones extremas) y puede, en este caso, ser engañosa. La mediana es una medida resistente a la influencia de determinaciones extremas y representa mejor el centro de la distribución de datos; pero la media tiene propiedades matemáticas muy útiles.

2. Medidas de variación.

- Varianza de una muestra de n observaciones $y_1, y_2, ..., y_n$

$$s^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1} = \frac{\sum_{i=1}^{n}\left(y_i^2 - 2y_i\bar{y} + \bar{y}^2\right)}{n-1} = \frac{\sum_{i=1}^{n} y_i^2 - 2\bar{y}\sum_{i=1}^{n} y_i + \sum_{i=1}^{n}\bar{y}^2}{n-1} = \frac{\sum_{i=1}^{n} y_i^2 - 2n\bar{y}\frac{\sum_{i=1}^{n} y_i}{n} + n\bar{y}^2}{n-1}$$

$$= \frac{\sum_{i=1}^{n} y_i^2 - 2n\bar{y}^2 + n\bar{y}^2}{n-1} = \frac{\sum_{i=1}^{n} y_i^2 - n\bar{y}^2}{n-1} = \frac{\sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}}{n-1}$$

desviación estándar de la muestra: $s = \sqrt{s^2}$

- Varianza poblacional: $\sigma^2 = \dfrac{\sum_{i=1}^{n}(y_i - \mu)^2}{n}$, *donde* μ *es la media poblacional*

desviación estándar poblacional: $\sigma = \sqrt{\sigma^2}$

La varianza tiene importancia teórica, pero es difícil de interpretar debido a que da unidades cuadradas, lo que hace que la desviación estándar sea más fácil de interpretar.

Calcular la varianza de la siguiente muestra, 1, 3, 2, 2, 4

$$\sum_{i=1}^{n} y_i = 1+3+2+2+4 = 12$$

$$\sum_{i=1}^{n} y_i^2 = 1+9+4+4+16 = 34$$

$$s^2 = \frac{\sum_{i=1}^{n} y_i^2 - \dfrac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}}{n-1} = \frac{34 - \dfrac{(12)^2}{5}}{4} = 1.3$$

Regla empírica: si un conjunto de datos tiene distribución con forma aproximada de "campana", tenemos que:
Aproximadamente 68% de los datos quedan a una desviación estándar (tanto a la izquierda como a derecha) de la media
Aproximadamente 95% de los datos quedan a dos desviaciones estándar (tanto a la izquierda como a derecha) de la media
Casi todos los datos quedan a tres desviaciones estándar de la media.

2. Medidas de posición relativa.

Percentiles:

Un percentil de a %, $P_{a\%}$ es aquel valor tal que un a % de los datos es menor a él y un (1-a)% de ellos es mayor a él.

- Cuartiles: tres puntos Q_1, Q_2, Q_3, que dividen el total de frecuencias acumuladas en 25, 50 y 75 % respectivamente
- Deciles: nueve puntos D_1, D_2, . . . , D_9 que dividen el total de frecuencias acumuladas en porciones de 10%

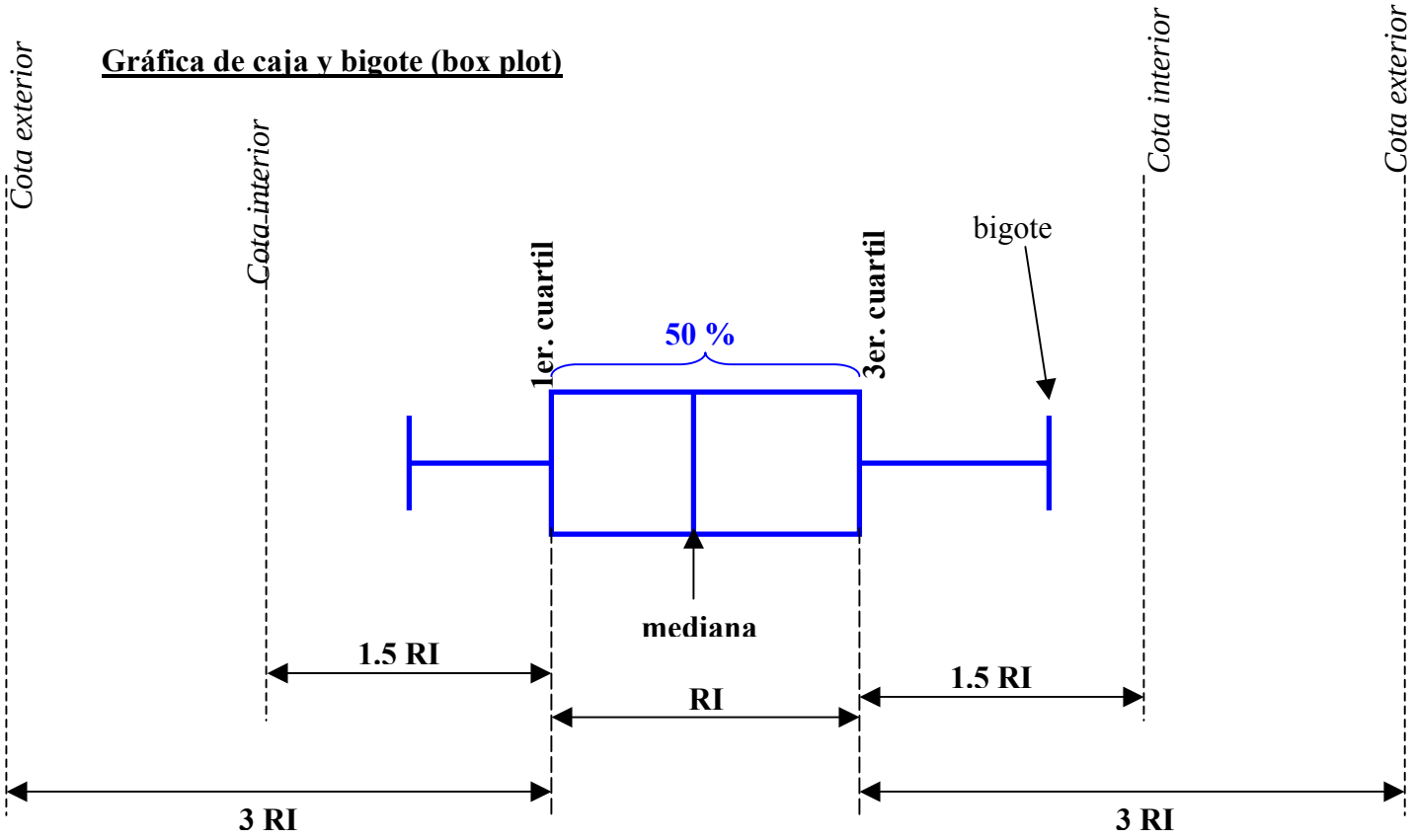
El percentil cincuenta $P_{50} = Q_2$ = segundo cuartil = mediana
El percentil setenta y cinco $P_{75} = Q_3$ = tercer cuartil

La posición (i) del p-ésimo percentil se calcula como: $(i) = \frac{p(n+1)}{100}$ y luego se redondea al entero más cercano.

En los casos del

- Primer cuartil $(i) = \frac{1}{4}(n+1)$ si cae entre dos enteros redondear hacia arriba.
- Tercer cuartil $(i) = \frac{3}{4}(n+1)$ si cae entre dos enteros redondear hacia abajo.

Gráfica de caja y bigote (box plot)



RI = rango intercuartilico = 3er cuartil – 1er cuartil = $Q_3 - Q_1$

Los bigotes representan la distancia de la mayor y la menor de las observaciones que están a menos de 1.5RI de la caja (valores más cercanos a las cotas interiores por dentro)

Valores fuera de las cotas.

Fuera de las cotas interiores: casos atípicos (representados por un asterisco*****)
Fuera de las cotas exteriores: valores extremos (representados por un circulo ***o***)

Ejemplo: tiempos en segundos de uso de la unidad central de proceso CPU

n = 25

1.17	1.61	1.16	1.38	3.53
1.23	3.76	1.94	0.96	4.75
0.19	2.41	0.71	0.02	1.59
0.15	0.82	0.47	2.16	2.01
0.92	0.75	2.59	3.07	1.40

Ordenando los datos mediante un diagrama de tallo y hoja:

Tallo (unidades)	Hojas (decimales)								
0	02	15	19	47	71	75	82	92	96
1	16	17	23	38	40	59	61	94	
2	01	16	41	59					
3	07	53	76						
4	75								

Ya se observa que los datos presentan asimetría.

Determinando los tres cuartiles tenemos:

- Primer cuartil (25 percentil):
  $(i) = \frac{1}{4}(n+1) = \frac{26}{4} = 6.5 \ (\textit{redondeando hacia arriba}) \rightarrow 7 \Rightarrow Q_1 = 0.82$

- Segundo cuartil (mediana): $(i) = \frac{1}{2}(n+1) = \frac{26}{2} = 13 \Rightarrow Q_2 = 1.38$

- Tercer cuartil (75 percentil):
  $(i) = \frac{3}{4}(n+1) = \frac{3}{4}(26) = 19.5 \ (\textit{redondeando hacia abajo}) \rightarrow 19 \Rightarrow Q_3 = 2.16$

RI = 2.16 – 0.82 =1.34

Cotas interiores:

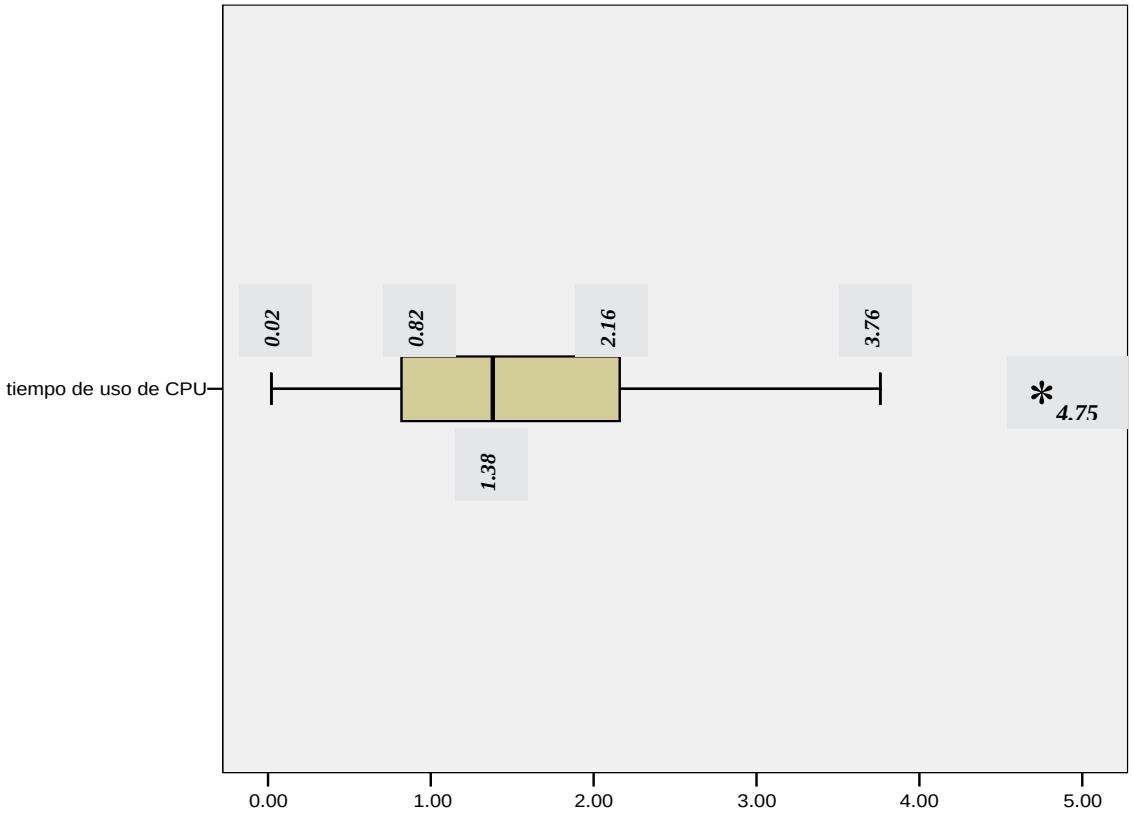
- Q_1 – 1.5RI = 0.82 – 1.5(1.34) = – 1.19
- Q_3 + 1.5RI = 2.16 + 1.5(1.34) = 4.17

valores más cercanos a las cotas interiores por dentro (bigotes): 0.02 y 3.76

Cotas exteriores:

- Q_1 – 3RI = 0.82 – 3(1.34) = – 3.2
- Q_3 + 3RI = 2.16 + 3(1.34) = 6.18

Valores fuera de las cotas: 4.75 (caso atípico)



- Tenemos asimetría hacia la izquierda
- Existe un valor atípico de 4.75 segundos
- La mediana es de 1.38 segundos
- El 50% de los tiempos de uso de la unidad central de proceso CPU esta en el intervalo de 0.82 a 2.16 segundos

Definiciones:

Una estadística. Medida descriptiva numérica calculada a partir de datos de la muestra.
Un parámetro θ**.** Medida descriptiva numérica de una población
Estimador $\hat{\theta}$**.** Una estadística que "pretende" darnos una idea del valor del parámetro.

Para estimar un parámetro, muestreamos una población y luego utilizamos la estadística de la muestra para inferir acerca del valor del parámetro de la población.

Dado que un estimador puntual se calcula a partir de una muestra posee una distribución de muestreo que describe por completo sus propiedades.

Propiedades deseables de los estimadores:

- Que el estimador sea insesgado

Un estimador es insesgado si $E\left(\hat{\theta}\right)=\theta$; si $E\left(\hat{\theta}\right)\neq\theta$ se dice que el estimador está sesgado. El sesgo B de un estimador $\hat{\theta}$ es $B=E\left(\hat{\theta}\right)-\theta$

- Que la distribución de muestreo de un estimador sea de varianza mínima.

El estimador insesgado de varianza mínima (MVUE), es el estimador que tiene la varianza más pequeña de entre todos los estimadores insesgados.

Hay ocasiones en las que no podemos lograr la falta de sesgo y también de varianza mínima en el mismo estimador. En un caso así preferimos el estimador que minimiza el error cuadrático medio (ECM):

$$ECM=E\left[\left(\hat{\theta}-\theta\right)^2\right]=V\left(\hat{\theta}\right)+B^2$$

Ejercicio:

Sea $y_1, y_2, ..., y_n$ una muestra aleatoria de n observaciones. Se desconoce la distribución de la población muestreada.
Demuestre que la varianza de la muestra s^2 es un estimador insesgado de la varianza de la población σ^2

Por demostrar que $E\left(s^2\right)=\sigma^2$

Dem.

$$s^2=\frac{\sum_{i=1}^{n}y_i^2-\frac{\left(\sum_{i=1}^{n}y_i\right)^2}{n}}{n-1}=\frac{\sum_{i=1}^{n}y_i^2-n\bar{y}^2}{n-1}$$

$$E\left(s^2\right)=E\left\{\frac{1}{n-1}\left[\sum_{i=1}^{n}y_i^2-n\bar{y}^2\right]\right\}=\frac{1}{n-1}\left\{E\left[\sum_{i=1}^{n}y_i^2\right]-E\left[n\bar{y}^2\right]\right\}=\frac{1}{n-1}\left\{\sum_{i=1}^{n}E\left(y_i^2\right)-nE\left(\bar{y}^2\right)\right\}$$

Pero:

$V(y)=E\left(y^2\right)-\left[E(y)\right]^2\Rightarrow\sigma^2=E\left(y^2\right)-\mu^2\Rightarrow E\left(y^2\right)=\sigma^2+\mu^2$

cada valor de *i* se escogió al azar de una población con media μ y varianza σ^2, entonces:

$E\left(y_i^2\right)=\sigma^2+\mu^2\qquad(i=1,2,...,n)$

y $E\left(\bar{y}^2\right)=\sigma_{\bar{y}}^2+\left(\mu_{\bar{y}}\right)^2=\frac{\sigma^2}{n}+\mu^2\rightarrow$ la última igualdad es debido al Teorema Central del límite $\bar{y}\sim N\left(\mu,\frac{\sigma^2}{n}\right)$

Sustituyendo

$$E\left(s^2\right)=\frac{1}{n-1}\left\{\sum_{i=1}^{n}E\left(y_i^2\right)-nE\left(\bar{y}^2\right)\right\}=\frac{1}{n-1}\left\{\sum_{i=1}^{n}\left(\sigma^2+\mu^2\right)-n\left(\frac{\sigma^2}{n}+\mu^2\right)\right\}=\frac{1}{n-1}\left(n\left(\sigma^2+\mu^2\right)-\sigma^2-n\mu^2\right)$$

$$E\left(s^2\right)=\frac{1}{n-1}\left(n\sigma^2+n\mu^2-\sigma^2-n\mu^2\right)=\frac{1}{n-1}\left(n\sigma^2-\sigma^2\right)=\frac{1}{n-1}(n-1)\sigma^2=\sigma^2$$

$\therefore E\left(s^2\right)=\sigma^2\quad Q.E.D$

s^2 es un estimador insesgado de la varianza de la población σ^2