Probabilidad y Estadística – Licenciatura en Computación y Profesorados

Guía N°1: Estadística Descriptiva

Problema 1: Dada una muestra de n valores $\{x_i, i = 1, \ldots, n\}$,

a) Demostrar que si a todos los valores de la muestra se le suma una constante r: $y_i = x_i + r$, entonces la media de los datos transformados es igual la media de los datos originales más r: $\overline{y} = \overline{x} + r$; mientras que la desviación estándar de los datos transformados es igual a la desviación estándar de los datos originales: $s_y = s_x$.

b) Demostrar que si todos los valores de la muestra se multiplican por una constante positiva k: $y_i = k\, x_i$, $k > 0$, entonces se cumple que: $\overline{y} = k\, \overline{x}$ y $s_y = k\, s_x$. ¿Qué pasa si la constante es negativa ($k < 0$)?

c) Una muestra de temperaturas para iniciar cierta reacción química, medidas en grados Celcius (^0C), generó un promedio muestral igual a 87.3^0C y una desviación estándar muestral igual a 1.04^0C.
¿Cuánto vale el promedio y el desvío estándar muestral de los datos expresados en grados Fahrenheit (^0F)?
Ayuda: La relación que existe entre ambas escalas de temperaturas está dada por: $T_F = 9/5\, T_C + 32$.

Problema 2: Sean $v_1, v_2, \ldots, v_n$ las observaciones obtenidas de un experimento. La media muestral resulta $\overline{v}$ y el correspondiente desvió estándar resulta s_v.

a) Si $u_i = v_i - \bar{v}$ para $i = 1, \ldots, n$. ¿Qué relación existe entre las medias muestrales y los desvíos estándar muestrales de u y v?

b) Si $z_i = (v_i - \overline{v})/s_v$ para $i = 1, \ldots, n$. ¿Qué relación existe entre las medias muestrales y los desvíos estándar muestrales de z y v?

Problema 3: En un artículo publicado en la revista Technometrics se reportan los resultados de experimentos en los que se registran las precipitaciones pluviales (expresados en mm) correspondientes a nubes naturales y "sembradas" artificialmente con centros de condensación. Los resultados reportados se transcriben en la tabla ordenados de mayor a menor.

a) Para los valores de lluvia, de cada grupo, calcular el máximo, el mínimo, el rango o amplitud, la media, la mediana, el desvío estándar y los cuartiles.

b) Confeccionar para cada grupo de mediciones un diagrama de caja (box plot) y comparar.

Lluvia de nubes de control			Lluvia de nubes "sembradas"		
1202.6	87.0	26.1	2745.6	274.7	115.3
830.1	81.2	24.4	1697.8	274.7	92.4
372.4	68.5	21.7	1656.0	255.0	40.6
345.5	47.3	17.3	978.0	242.5	32.7
321.2	41.1	11.5	703.4	200.7	31.4
244.3	36.6	4.9	489.1	198.6	17.5
163.0	29.0	4.9	430.0	129.6	7.7
147.8	28.6	1.0	334.1	119.0	4.1
95.0	26.3		302.8	118.3	

Problema 4: En la publicación "Time lapse cinematographic analysis of Beryllium lung fibroblast interactions"se reportaron los resultados del comportamiento de ciertas células individuales expuestas al berilio. Una característica importante de tales células es su tiempo de interdivisión (IDT). Los resultado obtenidos para el IDT se consignan en la tabla.

a) Para los valores muestrales de IDT y ln(IDT) (siendo ln(IDT) el logaritmo natural de IDT), calcular el máximo, el mínimo, el rango o amplitud, la media, la mediana y el desvío estándar.

b) Construir una tabla con las frecuencias acumuladas y relativas para los datos de IDT, utilizando intervalos de clase de longitud 10 y comenzando en el valor 10.

c) Con la tabla obtenida en (b), construir el histograma de frecuencias relativas.

d) Construir la correspondiente tabla con las frecuencias acumuladas y relativas de los datos de ln(IDT), utilizando en este caso intervalos de clase de longitud 0.4 y comenzando en 2.5.

e) Con la tabla obtenida en (d), construir el correspondiente histograma de frecuencias relativas.

f) Realizar los diagramas de caja (box plot) para los datos de IDT y de ln(IDT). ¿Qué información puede extraerse de ellos? Confrontar lo observado con los correspondientes histogramas.

g) Calcular la proporción de datos que se encuentran en el intervalo $\overline{y} \pm k\, s_y$, para $k = 1,\ 2,\ 3$, donde $\bar{y}$ y s_y son el promedio y desvío estándar muestral, respectivamente, en cada uno de los casos considerados (IDT y ln(IDT)).

IDT				ln(IDT)			
13.70	21.10	28.10	40.90	2.62	3.05	3.34	3.71
15.50	21.40	28.90	43.50	2.74	3.06	3.36	3.77
16.80	21.40	30.60	46.00	2.82	3.06	3.42	3.83
17.40	22.30	31.20	48.90	2.86	3.10	3.44	3.89
17.90	23.70	31.90	52.10	2.88	3.17	3.46	3.95
18.60	25.50	32.00	55.60	2.92	3.24	3.47	4.02
19.10	25.80	34.80	57.30	2.95	3.25	3.55	4.05
19.50	26.20	36.30	60.10	2.97	3.27	3.59	4.10
20.70	26.60	38.40	62.30	3.03	3.28	3.65	4.13
21.00	28.00	38.80	72.80	3.04	3.33	3.66	4.29

• Guía para la construcción de un diagrama de caja (box plot) •

Los histogramas nos dan una información cualitativa sobre el comportamiento de los datos, dado que presentan de forma resumida como se distribuyen los datos respecto de la media y permiten visualizar cual es el valor (o valores) mas frecuentes. En 1977, Tukey presentó un simple método gráfico-cuantitativo que resume varias de las características más destacadas de un conjunto de datos. Tal método se conoce con el nombre de *diagrama de caja* o *box plot*.

Las características de los datos incorporadas por este diagrama son:

a) centro o posición del valor mas representativo,

b) dispersión,

c) naturaleza y magnitud de cualquier desviación de la simetría e

d) identificación de los puntos no usuales o atípicos, o sea puntos marcadamente alejados de la masa principal de datos.

La presencia de datos atípicos producen cambios drásticos en la media muestral ($\overline{y}$) y la desviación estándar muestral (s), no así en otras medidas que son más *resistentes* o *robustas*, como lo son la mediana muestral ($\hat{y}$) y una medida de dispersión llamada rango intercuartil (RIQ). Estos parámetros los definimos a continuación. Sean $y_1, y_2, \ldots, y_n$ un conjunto de n datos. Con $y_{(i)}$ se denota la observación que ocupa el lugar i-ésimo, después de ser ordenados los datos de menor a mayor. Es decir, $y_{(1)}$ es la menor observación, $y_{(2)}$ la segunda más pequeña y así sucesivamente, siendo $y_{(n)}$ el mayor valor observado. Resulta de esta manera:

$$\hat{y} = \begin{cases} \dfrac{y_{(n/2)} + y_{((n/2)+1)}}{2}, & \text{si } n \text{ es par} \\ y_{((n+1)/2)}, & \text{si } n \text{ es impar} \end{cases}$$

$$\text{Cuartil Inferior} = \begin{cases} \text{mediana}\{y_{(i)}, 1 \leq i \leq \frac{n}{2}\}, & \text{si } n \text{ es par} \\ \text{mediana}\{y_{(i)}, 1 \leq i \leq \frac{n+1}{2}\}, & \text{si } n \text{ es impar} \end{cases}$$

$$\text{Cuartil Superior} = \begin{cases} \text{mediana}\{y_{(i)}, \frac{n}{2} \leq i \leq n\}, & \text{si } n \text{ es par} \\ \text{mediana}\{y_{(i)}, \frac{n+1}{2} \leq i \leq n\}, & \text{si } n \text{ es impar} \end{cases}$$

De esta manera el *Cuartil Inferior* (*Superior*) muestral es la mediana de la mitad más pequeña (más grande) de los datos. Notar que si n es impar $\hat{y}$ está incluída en ambas mitades.

Una medida de dispersión robusta a los puntos atípicos es el *Rango intercuartil* (*RIQ*) definido como:

RIQ = Cuartil Superior - Cuartil Inferior.

Con estas definiciones en mente, los pasos a seguir para la construcción del *box plot* son:

Paso 1: Ordenar los datos de menor a mayor.

Paso 2: Calcular $\hat{y}$, el cuartil superior, el cuartil inferior y el *RIQ*.

Paso 3: Sobre un eje horizontal o vertical marcar los valores extremos ($y_{(1)}$ y $y_{(n)}$) y los cuartiles inferior y superior.

Paso 4: Sobre este eje, dibujar una caja cuyo borde izquierdo sea el cuartil inferior y el borde derecho el cuartil superior.

Paso 5: Dentro de la caja trazar un segmento perpendicular ubicado sobre la mediana.

Paso 6: Trazar segmentos desde cada extremo de la caja hasta las observaciones más alejadas, que no superen $1{,}5 \times$ *RIQ)* de los bordes correspondiendes.

Paso 7: Marcar con circunferencias aquellos puntos comprendidos entre $1{,}5 \times$ *RIQ* y $3 \times$ *RIQ* respecto del borde más cercano, estos puntos se llaman *puntos anómalos suaves*, y con asteriscos aquellos puntos que superen los $3 \times$ *RIQ* respecto de los bordes más cercanos, estos puntos se llaman *puntos anómalos extremos.*

Fa.M.A.F ©2002