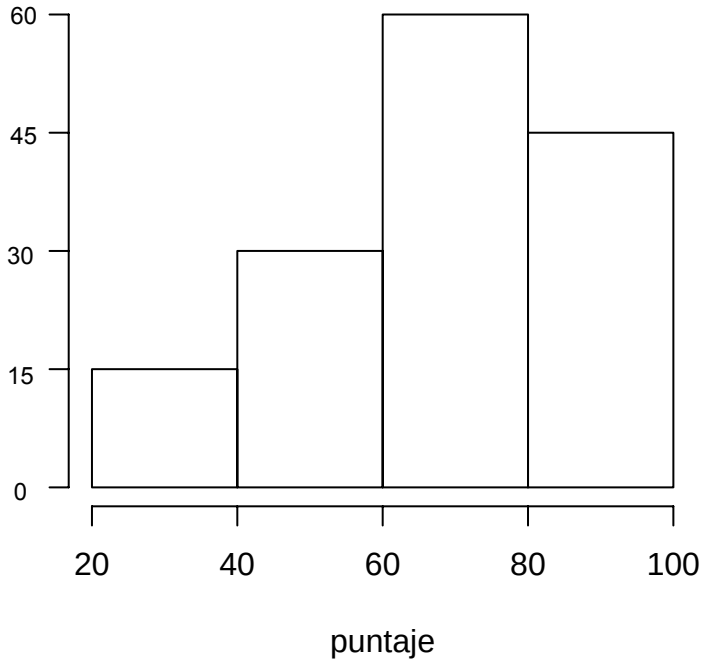
Probabilidad y Estadística
Profesorados y Licenciatura en Computación

Guía N°1: Estadística descriptiva

Ejercicio 1:
El histograma que presentamos a continuación muestra la distribución de notas finales obtenidas por los alumnos de una materia. Observar y responder.
a) ¿Obtuvo algún alumno una puntuación inferior a 20 puntos?
b) El histograma muestra que un 10% obtuvo notas entre 20 y 40 puntos, ¿qué porcentaje obtuvo notas entre 40 y 60 puntos?
c) ¿Qué porcentaje de alumnos tuvieron notas entre 60 y 80 puntos?
d) ¿Qué porcentaje de alumnos superó los 60 puntos?



Ejercicio 2:
Los datos provienen de un estudio del área de la salud. Las categorías registradas fueron:

O	O	N	J	T	F	B	B	F	O	J	O	O	M
O	F	F	O	O	N	O	N	J	F	J	B	O	T
J	O	J	J	F	N	O	B	M	O	J	M	O	B
O	F	J	O	O	B	N	T	O	O	O	M	B	F
J	O	F	N										

donde: J = articulación hinchada, F = fatiga, B = dolor de espalda, M = debilidad muscular, T = tos, N = nariz con flujo o irritación, O = otro. Obtenga las frecuencias absolutas y relativas para las distintas categorías y realice un gráfico para estos datos.

Ejercicio 3:
Dados los siguientes conjuntos de datos:

i) 0, 20, 40, 50, 60, 80, 100 ii) 0, 48, 49, 50, 51, 52, 100 iii) 0, 1, 2, 50, 98, 99, 100

a) ¿Cuál es el promedio en cada uno de ellos?

b) ¿Cuál de ellos presenta una mayor dispersión respecto de la media?

c) ¿Cuál de ellos presenta una menor dispersión respecto de la media?

Ejercicio 4:

Un profesor toma un examen con tres preguntas y asigna un punto a cada una de ellas. Un 30% de la clase consigue 3 puntos, un 50% 2 puntos, un 10% 1 punto y el 10% restante 0 puntos.

a) Si la clase se compone de 10 alumnos, ¿cuál es la nota promedio?

b) Si la clase se compone de 20 alumnos, ¿cuál es la nota promedio?

c) ¿Podría decir el valor de la nota promedio sin saber cuántos alumnos hay en la clase?

Ejercicio 5:

Tres profesores están comparando las notas de los exámenes finales que han tomado. Cada profesor tiene 99 alumnos. En la clase A un estudiante obtuvo 1 punto, otro obtuvo 99 puntos y el resto obtuvieron 50 puntos. En la clase B, 49 estudiantes lograron una puntuación de 1, un estudiante logró 50 puntos y 49 alumnos llegaron a los 99 puntos. En la clase C, un estudiante logró 1 punto, otro estudiante 2 puntos, otro estudiante 3 puntos y así sucesivamente hasta llegar al estudiante número 99 que logró 99 puntos.

a) ¿Hay alguna clase que tenga un promedio de notas superior a las demás o todos son iguales?

b) ¿Hay alguna clase que tenga una desviación estándar de notas superior a las demás o todas son iguales?

Ejercicio 6:

El artículo científico "The Pedaling Technique of Elite Endurance Cyclists" (*Int. J. Sport Biomechanics*, 1991, pp. 29-53) reportó los siguientes datos sobre fuerza en una sola pierna para carga pesada de trabajo:

160	174	176	177	179	180	180	181	183	187
191	194	200	202	204	205	207	211	211	254

a) Calcule e interprete la media y mediana muestrales.

b) Suponga que la última observación fue 211 en vez de 254. ¿Cómo cambiarían la media y la mediana?

c) La media muestral α-recortada consiste en calcular el promedio muestral excluyendo de la muestra el $\alpha\%$ de las observaciones más pequeñas y el $\alpha\%$ de las más grandes.

Calcule una media recortada eliminando las observaciones mínima y máxima de la muestra. ¿Cuál es el porcentaje correspondiente de recorte?

d) Calcule la media muestral 20% recortada.

e) También, el artículo presenta valores de la potencia de una sola pierna para el ejercicio liviano. La media muestral para $n = 19$ observaciones fue 119.8, y la vigésima observación, algo apartada, fue 159. ¿Cuál es el valor de la media muestral para toda la muestra?

Ejercicio 7:

En una fábrica automotriz se quiere evaluar la resistencia de los automóviles a una prueba de choque a una velocidad de 5 millas por hora. Se considera un éxito (E) si el automóvil no presenta daños visibles después de la prueba, y fracaso (F) en caso contrario. Se somete a la prueba a diez automóviles y se obtienen los siguientes resultados: E E F E E E F F E E

a) Si denotamos con x el número de éxitos en la muestra, ¿Cuál es el valor de la proporción muestral $\frac{x}{n}$?

b) Sustituya cada E con un 1 y cada F con un 0. Luego, calcule $\overline{x}$ para esta muestra numericamente codificada. ¿Cómo se compara $\overline{x}$ con $\frac{x}{n}$?

c) Suponga que se decide incluir 15 automóviles más en el experimento. ¿Cuántos de éstos deberían ser E para obtener $\frac{x}{n} = 0.80$ para toda la muestra de 25 automóviles?

Ejercicio 8:

a) Sean a y b dos constantes y sea $y_i = ax_i + b$ para $i = 1, 2, ..., n$. ¿Cuáles son las relaciones entre $\overline{x}$ e $\overline{y}$? y ¿entre s_x^2 y s_y^2?

b) El artículo científico "Penicillin in the treatment of meningitis" reportó las temperaturas del cuerpo, en grados Fahrenheit, de pacientes hospitalizados con meningitis. Diez de las observaciones fueron: 104.0,

104.8, 101.6, 108, 103.8, 100.8, 104.2, 100.2, 102.4 y 101.4.
i) Calcular el promedio y desvío estándar muestral.
ii) ¿Cuánto vale el promedio y el desvío estándar muestral de los datos expresados en grados Celcius (^0C)?
Ayuda: La relación que existe entre ambas escalas de temperaturas está dada por: $T_C = (T_F - 32)5/9$.
iii) 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?

Ejercicio 9:
Decir si son verdaderas o falsas las siguientes afirmaciones y justifique su respuesta.
a) Si suma 7 a cada uno de los datos de un conjunto entonces el promedio aumenta 7 puntos.
b) Si suma 7 a cada uno de los datos de un conjunto entonces la desviación estándar aumenta 7 puntos.
c) Si dobla el valor de cada dato de un conjunto entonces el promedio dobla su valor.
d) Si dobla el valor de cada dato de un conjunto entonces la desviación estándar dobla su valor.
e) Si cambia el signo de cada dato de un conjunto entonces el promedio cambia de signo.
f) Si cambia el signo de cada dato de un conjunto entonces el desvío estándar cambia el signo.

Ejercicio 10:
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 $u_i = v_i - \overline{v}$ 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?

Ejercicio 11:
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, hallar 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	

Ejercicio 12:
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), hallar el máximo, el mínimo, el rango o amplitud, la media, la mediana y el desvío estándar.
b) Construir una tabla con las frecuencias absolutas 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 absolutas 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 cómo se distribuyen los datos respecto de la media y permiten visualizar cuál es el valor (o valores) más 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
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 ($\tilde{y}$) 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:

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

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

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

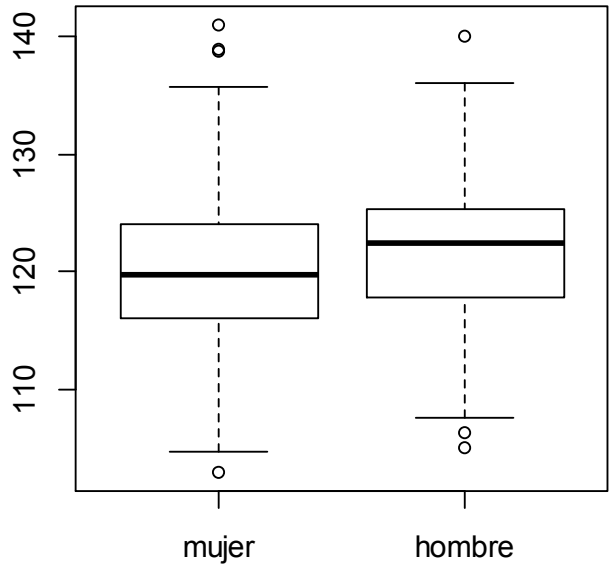
De esta manera el *Cuartil Inferior* (*Superior*) muestral es la mediana de la mitad más pequeña (más grande) de los datos. Notar que si n es impar $\tilde{y}$ está incluída en ambas mitades.
Una medida de dispersión robusta a los puntos atípicos es el *Rango intercuartil* (RIQ) definido como:
RIQ = 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 $\overline{y}$, $\tilde{y}$, el cuartil superior, el cuartil inferior y el RIQ.
Paso 3: Sobre este eje, dibujar una caja cuyo borde izquierdo sea el cuartil inferior y el borde derecho el cuartil superior.
Paso 4: Dentro de la caja marcar con un punto la posición del promedio o media muestral y trazar un segmento perpendicular ubicado sobre la mediana.
Paso 5: 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 6: 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*.

Continuación Guía 1

Ejercicio 13:

En un estudio, se midió la presión sanguínea sistólica (en mm Hg) en 350 mujeres y 350 hombres, entre 15 y 69 años. Los datos fueron cargados en R, creando los vectores `mujeres` y `hombres`.
A continuación se presentan medidas descriptivas de resumen y gráficos de caja para este conjunto de datos. Compare la presión sanguínea por sexo según lo que observa.

```
> summary(mujeres)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  102.9   116.1   119.8   119.8   124.0   141.1
> summary(hombres)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  105.1   117.8   122.4   121.8   125.4   140.1
> sd(mujeres)
[1] 6.450651
> sd(hombres)
[1] 5.711367
> sexo<-gl(2,350,labels=c("mujer","hombre"))
> plot(sexo,c(mujeres,hombres))
```



Ejercicio 14:

Los gráficos que se presentan a continuación corresponden a las notas del primer parcial de los alumnos de un curso de Estadística, en una comisión con clases en horario matutino y otra en horario nocturno. Comente respecto a la distribución de las calificaciones en los distintos horarios.

```
> boxplot(matutina,ylab="Calificación",main="Boxplot comisión
matutina")
> hist(matutina,xlab="Calificación",ylab="Frecuencia
absoluta",main="Histograma comisión matutina")
> boxplot(nocturna,ylab="Calificación",main="Boxplot comisión
nocturna")
> hist(nocturna,xlab="Calificación",ylab="Frecuencia
absoluta",main="Histograma comisión nocturna")
```