2.5 Cálculo de Medidas Estadisticas

Hay dos tipos principales de medidas Estadísticas: medidas de Tendencia Central y medidas de Variabilidad.

Las medidas de tendencia central dan una idea del centro de la distribución de los datos. Las principales medidas de este tipo son la media o promedio aritmético, la mediana, la moda y la media podada.

Las medidas de variabilidad expresan el grado de concentración o dispersión de los datos con respecto al centro de la distribución. Entre las principales medidas de este tipo están la varianza, la desviación estándar, el rango intercuartílico. Aparte también hay medidas de posición, como son los cuartiles, deciles y percentiles. Además, una medida de asimetría ("skewness") y una medida de aplanamiento ("kurtosis").

2.5.1 Medidas de Centralidad

La media o promedio se obtiene sumando todos los datos y dividiendo entre el número de datos. Es decir, si x_1, x_2,…,x_n, representan las observaciones de una variable X en una muestra de tamaño n, entonces la media de la variable X está dada por:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Medidas de Centralidad (cont.)

La media o promedio se obtiene sumando todos los datos y dividiendo entre el número de datos. Es decir, si x_1, x_2,…,x_n, representan las observaciones de una variable X en una muestra de tamaño n, entonces la media de la variable X está dada por:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Ejemplo 2.6. Supongamos que los siguientes datos representan el precio de 9 casas en miles.

74, 82, 107, 92, 125, 130, 118, 140, 153

Hallar el precio promedio de las casas.

$$\bar{x} = \frac{74+82+107+92+125+130+118+140+153}{9} = 113.4$$

Es decir, el costo promedio de una casa será 113,400.

Medidas de Centralidad (cont.)

La media es afectada por la asimetría de la distribución de los datos y por la presencia de “outliers” como se muestra en el siguiente ejemplo.

Ejemplo 2.7. Supongamos que en el ejemplo anterior se elige adicionalmente una casa cuyo precio es de 500,000.

Luego el promedio será:

$$\bar{x} = \frac{74 + 82 + 107 + 92 + 125 + 130 + 118 + 140 + 153 + 500}{10} = 152.1$$

En este caso la media da una idea errónea del centro de la distribución, la presencia del “outlier” ha afectado la media. Sólo dos de las 10 casas tienen precio promedio mayor de 152,100.

Medidas de Centralidad (cont.)

Propiedades de la media son:

A) El valor de la media debe estar entre el mayor dato y el menor dato.

B) Si a cada dato de la muestra se le suma (o resta) una constante entonces, la media queda sumada (o restada) por dicha constante.

C) Si a cada dato de la muestra se le multiplica (o divide) por una constante entonces, la media queda multiplicada (o dividida) por dicha constante.

La mediana es un valor que divide a la muestra en dos partes aproximadamente iguales. Es decir, como un 50 por ciento de los datos de la muestra serán menores o iguales que la mediana y el restante 50 por ciento son mayores o iguales que ella.

Para calcular la mediana primero se deben ordenar los datos de menor a mayor. Si el número de datos es impar, entonces la mediana será el valor central. Si el número de datos es par entonces, la mediana se obtiene promediando los dos valores centrales.

Medidas de Centralidad (cont.)

Ejemplo 2.8. Calcular la mediana de los datos del Ejemplo 3.6.

Ordenando los datos en forma ascendente, se tiene: 74, 82, 92, 107, 118, 125, 130, 140, 153. En este caso el número de datos es impar así que la mediana resulta ser 118 que es el quinto dato ordenado.

A diferencia de la media, la mediana no es afectada por la presencia de valores anormales. Asi, la mediana para los datos del ejemplo 3.7, donde hay

un número par de datos, la mediana resulta ser el promedio de los dos valores centrales: =121.5 y el dato anormal 500 no afecta el valor de la mediana.

Cuando la distribución es asimétrica hacia la derecha, la mediana es menor que la media. Si hay asimetría hacia la izquierda entonces la mediana es mayor que la media y cuando hay simetría, ambas son iguales.

Medidas de Centralidad (cont.)

La moda es el valor (o valores) que se repite con mayor frecuencia en la muestra. La Moda puede aplicarse tanto a datos cuantitativos como cualitativos.

Ejemplo 2.10. Los siguientes datos representan el número de veces que 11 personas van al cine mensualmente:

3, 4, 4, 5, 0, 2, 1, 5, 4, 5 y 4

Hallar la moda.

La Moda es 4. O sea que predominan más las personas que asisten 4 veces al mes al cine.

Ejemplo 2.11. Los siguientes datos representan tipos de sangre de 9 personas

A, O, B, O, AB, O, B, O, A

Hallar la Moda.

La Moda es el tipo de sangre O.

Medidas de Centralidad (cont.)

La media podada es una medida más resistente que la media a la presencia de valores anormales. Para calcular la Media Podada, primero se ordenan los datos en forma creciente y luego se elimina un cierto porcentaje de datos (redondear si no da entero) en cada extremo de la distribución, finalmente se promedian los valores restantes.

Ejemplo 2.12. Hallar la media podada del 5 por ciento para los datos del Ejemplo 2.7

El 5 por ciento de 10 datos es .5 que redondeando a 1 implica que hay que eliminar el mayor (500) y el menor (74) dato. Luego la media podada del 5 por ciento será

$$\bar{x} = \frac{82 + 107 + 92 + 125 + 130 + 118 + 140 + 153}{8} = 118.375$$

.

2.5.2 Medidas de Variabilidad

El rango o amplitud es la diferencia entre el mayor y menor valor de la muestra. Mientras mayor sea el rango existe mayor variabilidad.

Otra medida, para determinar el grado de concentración de los datos sería el promedio de las desviaciones con respecto a la media, es decir ,

$$\frac{\sum_{i=1}^{n}(x_i - \bar{x})}{n}$$

Sin embargo esto siempre daria CERO porque la suma de las desviaciones con respecto a la media es siempre CERO.

Medidas de Variabilidad (cont)

La **varianza** es una medida que da una idea del grado de concentración de los datos con respecto a la media.

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

Se divide por n-1 y no por n, porque se puede demostrar teóricamente que cuando se hace esto s^2 estima más eficientemente a la varianza poblacional

La desviación estándar es la raíz cuadrada positiva de la varianza y tiene la ventaja que está en las mismas unidades de medida que los datos. Se representa por s.

De por si sola la desviación estándar no permite concluir si la muestra es muy variable o poco variable. Al igual que la varianza es usada principalmente para comparar la variabilidad entre grupos.

Medidas de Variabilidad (cont)

Ejemplo 2.13. Las muestras siguientes:

muestra1

16 18 25 28 23 42 24 47 38 19 22 34

muestra2

116 118 125 128 123 142 124 147 138 119 122 134

tienen medias 28 y 128 respectivamente, e igual desviación estándar s = 10.018. O sea que se puede decir en términos absolutos que tienen igual variabilidad. Sin embargo comparándola con los datos tomados se puede concluir que la muestra 1 es bastante variable, mientras que la muestra 2 es poco variable.

El **coeficiente de variación** (CV) y que se calcula por $\frac{s}{|\bar{x}|}x100\%$. Si e l CV es mayor que 30% la muestra es muy variable y si CV es menor del 10% entonces no existe mucha variabilidad. Para el ejemplo el CV para la muestra 1 es 35.77 y para la muestra 2 es 7.82.

Medidas de Variabilidad (cont)

Ejemplo 2.13. Las muestras siguientes:

muestra1

16 18 25 28 23 42 24 47 38 19 22 34

muestra2

116 118 125 128 123 142 124 147 138 119 122 134

tienen medias 28 y 128 respectivamente, e igual desviación estándar s = 10.018. O sea que se puede decir en términos absolutos que tienen igual variabilidad. Sin embargo comparándola con los datos tomados se puede concluir que la muestra 1 es bastante variable, mientras que la muestra 2 es poco variable.

El **coeficiente de variación** (CV) y que se calcula por $\frac{s}{|\bar{x}|}x100\%$. Si e l CV es mayor que 30% la muestra es muy variable y si CV es menor del 10% entonces no existe mucha variabilidad. Para el ejemplo el CV para la muestra 1 es 35.77 y para la muestra 2 es 7.82.

Medidas de Variabilidad (cont)

Criterio para detectar "outliers".

Un primer criterio para identificar si un dato es un "outlier" es el siguiente:

Un dato que cae fuera del intervalo

$$(\bar{x} - 3s, \bar{x} + 3s)$$

puede ser considerado un "outlier". El criterio debe ser usado con precaución, puesto que la media, la varianza y la desviación estándar son afectadas por la presencia de "outliers".

Ejemplo 2.14. Dada la siguiente muestra

59, 62, 73, 79, 68, 77, 69, 71, 66, 98, 75

Determinar si 98 es un "outlier".

Solución:

Como 72.45 y s=10.43. Se tiene que si un dato cae fuera del intervalo (41.15, 103.75) será considerado un "outlier", 98 cae dentro de dicho intervalo por lo tanto no es "outlier".

Laboratorio 5

2.5.3. Medidas de Posición

Los Cuartiles: Son valores que dividen a la muestra en 4 partes aproximadamente iguales. El 25% de los datos son menores o iguales que el cuartil inferior o primer cuartil, representado por Q1. El siguiente 25 % de datos cae entre el cuartil inferior y la mediana, la cual es equivalente al segundo cuartil. El 75 % de los datos son menores o iguales que el cuartil superior o tercer cuartil, representado por Q3, y el restante 25% de datos son mayores o iguales que Q3.

Existen varios métodos para calcular los cuartiles pero la manera mas simple es ordenar los datos y considerar Q_1 como la mediana de la primera mitad, o sea aquella que va desde el menor valor hasta la mediana. Similarmente Q_3 es la mediana de la segunda mitad, o sea aquella que va desde la mediana hasta el mayor valor.

Medidas de Posición (cont)

Ejemplo 2.15. Calcular los cuartiles de las siguientes muestras:

a) 6, 8, 4, 12, 15, 17, 23, 18, 25, 11

Los datos ordenados serán: 4, 6, 8, 11, 12, 15, 17, 18, 23, 25

La primera mitad es: 4, 6, 8, 11, 12, luego $Q_1 = 8$

La segunda mitad es: 15, 17, 18, 23, 25, luego $Q_3 = 18$

b) 10, 22, 17, 13, 28, 40, 29, 18, 23, 39, 44

Los datos ordenados serán: 10, 13, 17, 18, 22, 23, 28, 29, 39, 40, 44

La primera mitad es: 10, 13, 17, 18, 22, 23, luego $Q_1 =$ $= 17.5$

La segunda mitad es: 23, 28, 29,39, 40, 44, luego $Q_3 =$ $= 34$

Una variante que se usa cuando el número de datos es impar es no usar la mediana, para cada mitad. Es decir en el ejemplo b), considerar que la primera mitad es 10, 13, 17, 18, y 22 y la segunda mitad es 28, 29, 39, 40, y 44. Así Q_1 sería 17 y Q_3 sería 39. R usa un proceso de interpolación para calcular los cuartiles

Medidas de Posición (cont)

A la diferencia de Q_3 y Q_1 se le llama **Rango Intercuartílico**, ésta es una medida de variabilidad que puede ser usada en lugar de la desviación estándar, cuando hay "outliers".

Los Deciles: Son valores que dividen a la muestra en 10 partes iguales

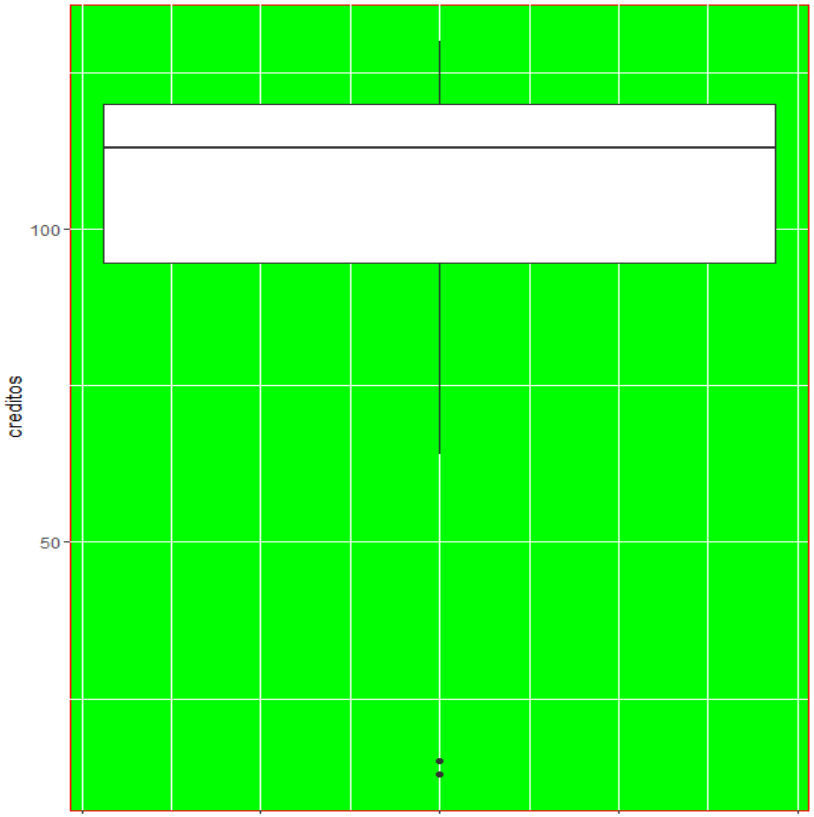
Los Percentiles: Dado un cierto porcentaje 100p, donde p varía entre 0 y 1, el percentil del 100p% es un valor tal que 100p% de los datos caen a la izquierda del percentil. En particular, la mediana y los cuartiles son percentiles. El primer cuartil es el percentil de 25%, la mediana es el percentil del 50% y el tercer cuartil es el percentil del 75%. Por ejemplo el puntaje minimo para que un estudiante este en el tercio superior de su clase es igual al puntaje de 66.67%.

•

2.6 Diagrama de caja ("Boxplot")

El "**Boxplot"** , al igual que el histograma y el "stem-and-leaf", permite tener una idea visual de la distribución de los datos. O sea, determinar si hay simetría, ver el grado de variabilidad existente y finalmente detectar "outliers". Pero además, el "Boxplot**"** es bien útil para comparar grupos, es una alternativa gráfica a la prueba estadística t de Student, si se comparan dos grupos o la prueba F del análisis de varianza si se comparan más de dos grupos. Todo lo anterior es posible debido a que se puede hacer múltiples boxplots en una misma gráfica, en cambio los histogramas y "stem-and- leaf" salen en secuencia uno por página.

En **R** hay varias maneras de obtener el "Boxplot" de un conjunto de datos, la primera es usando la funcion la bo***plot*** . Otra manera es usando la funcion geom_boxplot de la librería ggplot2



Interpretación: *La línea central de la caja representa la Mediana y los lados de la caja representan los cuartiles. Si la Mediana está bien al centro de la caja, entonces hay simetria. Si la Mediana está más cerca a Q_3 que a Q_1 entonces la asimetría es hacia la izquierda, de lo contrario la asimetría es hacia la derecha. Si la caja no es muy alargada entonces se dice que no hay mucha* variabilidad.

2.6 Diagrama de caja ("Boxplot") cont.

Si no hay "outliers" entonces las líneas laterales de la caja llegan hasta el valor mínimo por abajo, y hasta el valor máximo por arriba. Cuando hay "outliers" entonces éstos aparecen identificados en la figura y las lineas laterales llegan hasta los valores adyacentes a las fronteras interiores. Si las lineas laterales son bastantes alargadas entonces significa que los extremos de la distribución de los datos se acercan lentamente al eje X.

Las fronteras interiores se calculan como Q_1 - 1.5RIQ y Q_3 + 1.5RIQ respectivamente, donde RIQ = Q_3-Q_1 es el Rango Intercuartílico. Las fronteras exteriores se calculan por Q_1 - 3RIQ y Q_3 + 3RIQ. Si un valor cae más alla de las fronteras exteriores se dice que es **un "outlier" extremo**, en caso contrario el outlier es **moderado.** Un "outlier" moderado se representa por ***** y uno extremo por **0**.

Laboratorio 6