Medidas de dispersión

Manuel Villalobos Cid

Contents

Introducción

Las **medidas de dipersión** caracterizan la distribución de frecuencia de una variable estadística en una población o muestra permitiendo extraer información sobre su **variación**. Ejemplos de estas métricas son el **rango** (recorrido), la **varianza** y la **desviación estándar**.

Las medidas de dispersión complementan las **medidas de localización** permitiendo tener una noción sobre su representatividad en el conjunto estudiado, por ejemplo, usando el **coeficiente de variación**.

Esta actividad se centra en el uso de **medidas de dispersión**, las que serán estudiadas por medio de un conjunto de datos real proveniente del **área seguridad social** en nuestro país.

Conjunto de datos

Para esta actividad se usarán datos (descargar) publicados por el **Instituto Nacional de Estadísticas** (INE) provenientes del registro de **Carabineros de Chile**. Estos corresponden al **número de delitos** registrados por región (columna 1) durante el año 2017 (columna 2), divididos en aquellos que presentan o no detenidos (columnas 3 y 4). Ambas variables han sido **normalizadas** por la **población** de cada región (columna 5), presentándose como tasa de incidencia delictual cada 10,000 habitantes (columnas 6 y 7).

```
library("pander")
datos=read.csv("datos.csv",header = T,sep = ",")#Cargar datos
pander(head(datos,5),caption = "Resumen de conjunto de datos") #Mostrar datos
```

Table 1: Resumen de conjunto de datos (continued below)

Region	Casos.policiales	Casos.sin.detenidos
ARICA Y PARINACOTA	68529	62551
TARAPACA	106989	91056
ANTOFAGASTA	152404	128833
ATACAMA	54529	44229
COQUIMBO	168525	152937

Table 2: Table continues below

Casos..con.detenidos	Poblacion.Censo.2017	Tasa.sin.detenidos
5978	226068	2767
15933	330558	2755
23571	607534	2121
10300	286168	1546
15588	757586	2019

Tasa.con.detenidos
264
482
388
360
206

Medidas de localización

En la guía anterior se usaron diferentes funciones para el cálculo de las medidas de dispersión: **mean()**, **median()**, **quantile()**, entre otros. En esta oportunidad, tanto las medidas de localización y dispersión serán calculadas empleando la función **describe()** perteneciente a la biblioteca **psych()**.

La tasa de delitos con y sin incluir detenidos puede ser caracterizados como:

```
library("psych")
tabla_sindet=describe(datos$Tasa.sin.detenidos,skew=F,IQR=T,quant=c(0.25,0.5,0.75))
pander(tabla_sindet,caption = "Medidas de localización y dispersión
       para delitos sin detenidos")
```

Table 4: Medidas de localización y dispersión para delitos sin detenidos

vars	n	mean	sd	min	max	range	se	IQR	Q0.25	Q0.5	Q0.75
1	15	1728	503.6	1058	2767	1709	130	505.5	1436	1544	1942

```
tabla_condet=describe(datos$Tasa.con.detenidos,skew=F,IQR=T,quant=c(0.25,0.5,0.75))
pander(tabla_condet,caption = "Medidas de localización y dispersión
       para delitos con detenidos")
```

Table 5: Medidas de localización y dispersión para delitos con detenidos

vars	n	mean	sd	min	max	range	se	IQR	Q0.25	Q0.5	Q0.75
1	15	259.2	85.82	187	482	295	22.16	66.5	197.5	228	264

En este caso **n** corresponde al número de regiones, **mean** es el promedio de delitos, **min** es el valor mínimo registrado, **max** equivale al valor máximo, y **Q** representa los percentiles 25, 50 (mediana) y 75.

Los datos indican que en Chile en **promedio** hay **1,728 delitos** informados cada 10,000 habitantes que no involucran detenidos, y sólo **259** que sí los considera. Al estudiar los percentiles se puede apreciar que la brecha entre el primer y segundo cuartil es menor que la diferencia entre el segundo y tercero, lo que implica que la distribución de los datos no está centrada uniformemente sobre la mediana, sino concentrada entre los dos primeros cuartiles. Esto puede ser comprobado por medio de sus histogramas:

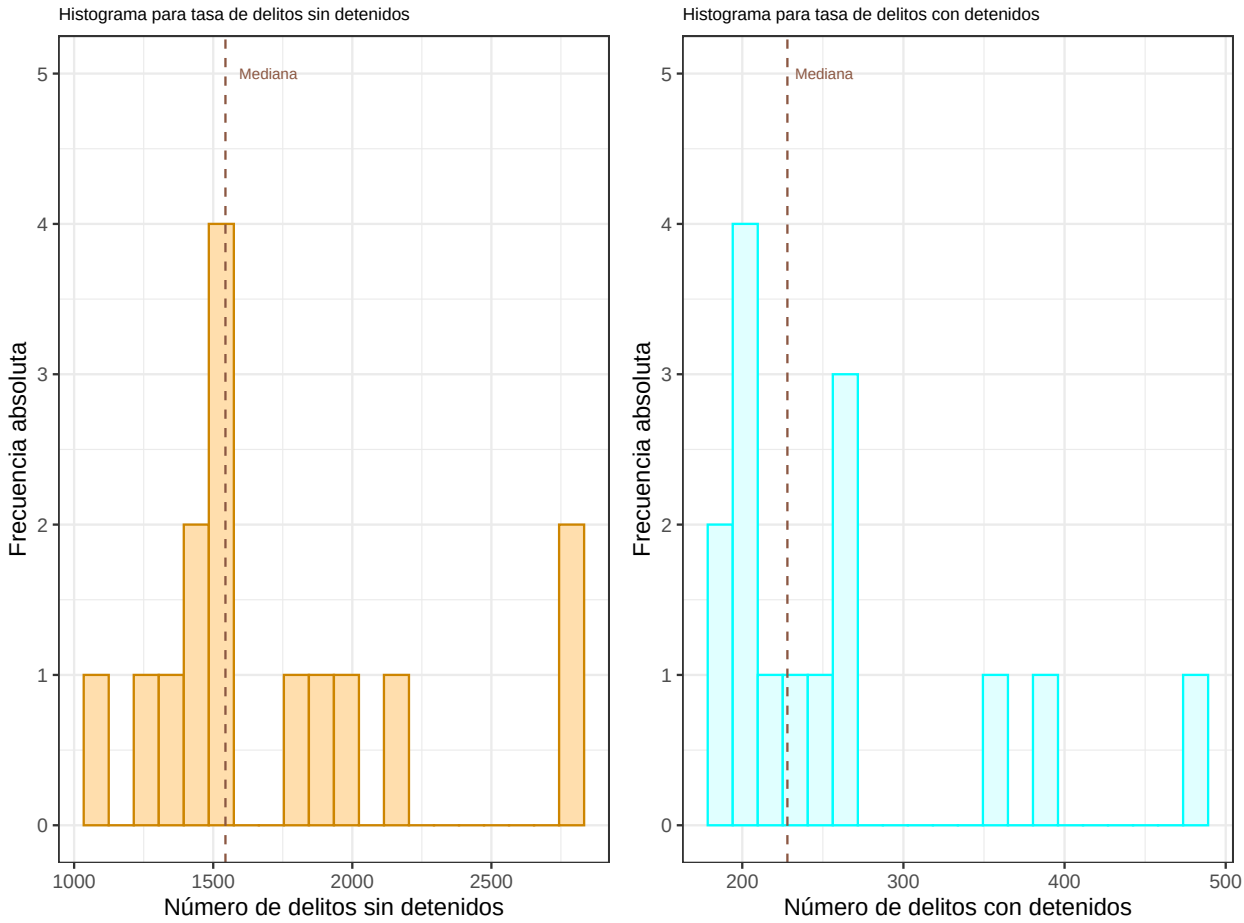
```
#install.packages('ggplot2')
library("ggplot2",warn.conflicts = F)
#library("easyGgplot2",warn.conflicts = F)


#Gráfico y mediana sin detenidos
grafico_sin=ggplot(datos,aes(datos$Tasa.sin.detenidos)) # Gráfico y datos base
grafico_sin = grafico_sin + geom_histogram(bins=20,fill="navajowhite",color="orange3")
grafico_sin = grafico_sin + theme_bw() # Visualización estándar en blanco y negro
grafico_sin = grafico_sin + ylab("Frecuencia absoluta") + xlab("Número de delitos sin detenidos")
grafico_sin = grafico_sin + ggtitle("Histograma para tasa de delitos sin detenidos")
grafico_sin=grafico_sin+geom_vline(xintercept =tabla_sindet$Q0.5,color="lightsalmon4",
                           linetype="dashed", show.legend = T)
grafico_sin=grafico_sin+annotate("text", x = (tabla_sindet$Q0.5+(tabla_sindet$Q0.5*0.1)),
                                 y = 5, label = "Mediana",color="lightsalmon4",cex=2.5)
grafico_sin=grafico_sin+theme(plot.title = element_text(size=8))

#GrÃ¡fico y mediana con detenidos
grafico_con=ggplot(datos,aes(datos$Tasa.con.detenidos)) # Gráfico y datos base
grafico_con = grafico_con + geom_histogram(bins=20,fill="lightcyan",color="cyan")
grafico_con = grafico_con + theme_bw() # Visualización estándar en blanco y negro
grafico_con = grafico_con + ylab("Frecuencia absoluta") + xlab("Número de delitos con detenidos")
grafico_con = grafico_con + ggtitle("Histograma para tasa de delitos con detenidos")
grafico_con=grafico_con+geom_vline(xintercept =tabla_condet$Q0.5,color="lightsalmon4",
                           linetype="dashed", show.legend = T)
grafico_con=grafico_con+annotate("text", x = (tabla_condet$Q0.5+(tabla_condet$Q0.5*0.1)),
                                 y = 5, label = "Mediana",color="lightsalmon4",cex=2.5)
grafico_con=grafico_con+theme(plot.title = element_text(size=8))

#Combinación de histogramas
multiplot(grafico_sin,grafico_con,cols=2)
```

```
## Loading required package: grid
```



Medidas de dispersión

Además de las medidas de localización, las Tablas 2 y 3 también incluyen medidas de dispersión: **rango** y **desviación estándar**. Otras medidas como la **varianza** o el **coeficiente de variación** pueden ser calculadas manualmente.

Rango

Esta métrica equivale al **intervalo** entre el **valor máximo** y el **valor mínimo** de una variable estadística. Para el caso de los delitos sin detenidos esta medida corresponde a **1,709** y en el caso de los delitos con detención es **295** (variable **range** en Tablas 2 y 3). Si bien esta métrica da cuenta de la dispersión en relación a las medias (1,727 y 259 respectivamente) es una magnitud que por sí sola no entrega una información acabada sobre la distribución de los datos.

Varianza y desviación estándar

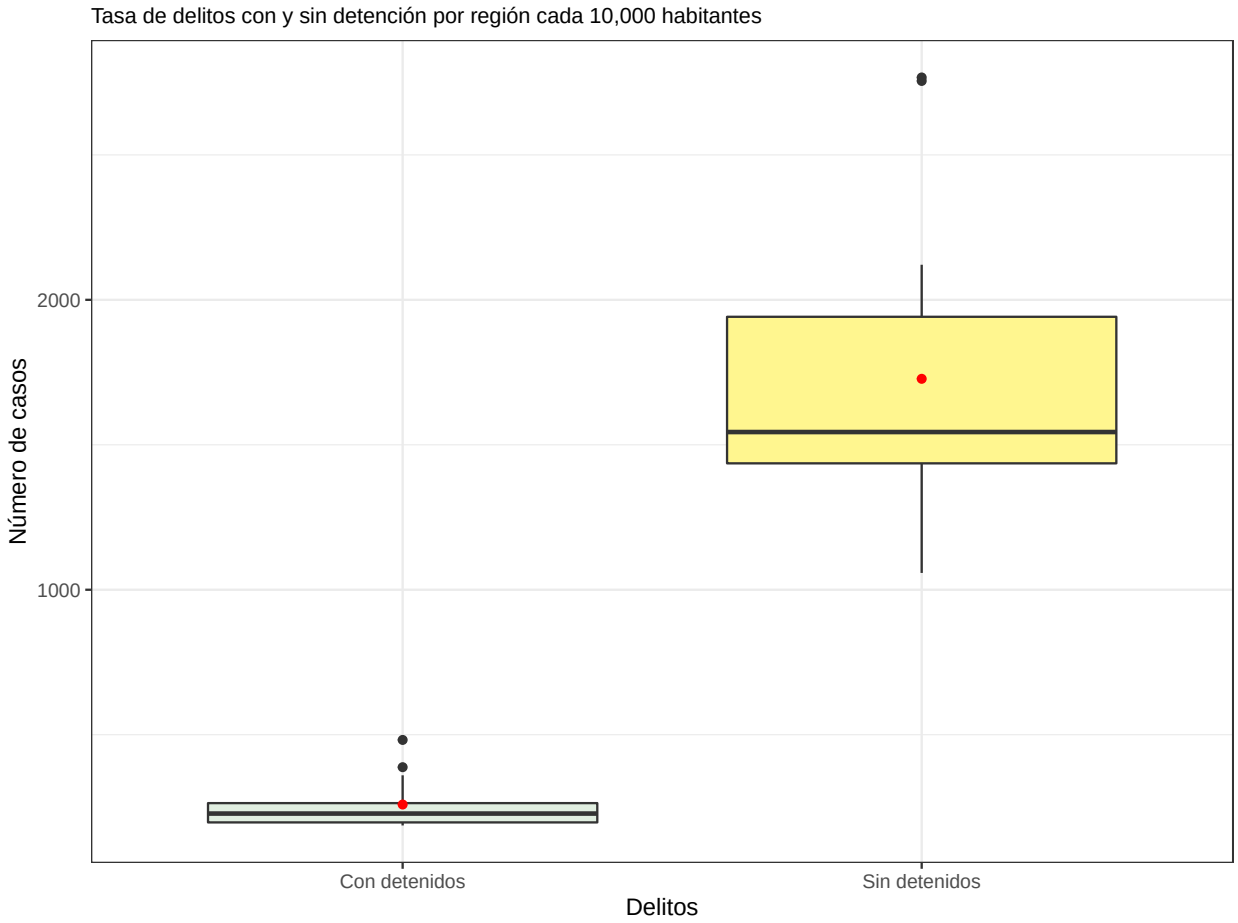
La **varianza** mide cómo los valores de los datos pueden diferir de su media. Se puede definir como la media aritmética de los cuadrados de las diferencias de los valores individuales de la media. Esta elevación al cuadrado asegura que los valores positivos y negativos no se anulan mutuamente, no obstante, la unidad de medida de la variable estadística también es elevada al cuadrado. Dejar la unidad de medida en una escala equivalente a la media requiere la aplicación de la raíz cuadrada sobre la **varianza**, esta métrica es conocida como **desviación estándar**. Esta medida es **85.82** para delitos con detención y **503.6** cuando no tienen, columna **sd** en Tablas 2 y 3.

Representación gráfica: diagramas de caja

Los gráficos o **diagramas de cajas** permiten estudiar visualmente la distribución de una variable estadística. La línea central en cada caja representa la mediana de la variable estudiada sobre un grupo, las líneas inferior y superior corresponden al primer y segundo cuartil, mientras que las líneas verticales son los límites que determinan los **valores atípicos** (círculos negros). Con ello nos es posible hacer una idea de la distribución de los datos, por ejemplo, el número de delitos sin detenidos varía en mayor medida entre regiones en comparación a los que involucran detenidos, teniendo una leve concentración entre los dos primeros cuartiles.

Con objetivo de diferenciar la media de la mediana, se incluirá esta métrica usando la función **stat_summary()** aplicando color rojo.

```
#install.packages('ggplot2')
datos_ajustados=c(datos$Tasa.sin.detenidos,datos$Tasa.con.detenidos)
clase=c(rep("Sin detenidos",length(datos$Tasa.sin.detenidos)),
        rep("Con detenidos",length(datos$Tasa.con.detenidos)))
datos_ajustados=data.frame(clase,datos_ajustados)
grafico_caja=ggplot(datos_ajustados,aes(x=clase,y=datos_ajustados)) +
  geom_boxplot(fill=c("honeydew2","khaki1"))
grafico_caja = grafico_caja + stat_summary(fun.y=mean, geom="point",
                                           show.legend = F,color="red")
grafico_caja= grafico_caja + theme_bw() + xlab("Delitos") +
  ylab("Número de casos") + ggtitle("Tasa de delitos con y sin detención por región cada 10,000 habitan
grafico_caja = grafico_caja+theme(plot.title = element_text(size=10))
#grafico_caja = grafico_caja + geom_jitter(shape=16, position=position_jitter(0.2))
plot(grafico_caja)
```



Valores atípicos

Un **valor atípico** es una observación que es numéricamente distante del resto de los datos. Para identificarlos se puede usar el siguiente criterio, donde Q_1 y Q_3 son el primer y tercer cuartil y el RIQ es el rango intercuartil $(Q_3 - Q_1)$. Un valor q será atípico si:

$$q < Q_1 - \alpha \times RIQ$$

, o,

$$q > Q_3 + \alpha \times RIQ$$

Si α se define como 1.5 se denomina valor atípico leve y si se usa 3, los valores q resultantes se denominan valores atípicos extremos. Según el gráfico de caja, hay algunas regiones que poseen una tasa de delitos que escapa del resto (valore atípicos en puntos negros). Para identificarlas, según el criterio leve, se puede hacer lo siguiente:

```
#Sin detenidos
ualto=tabla_sindet$Q0.75+1.5*(tabla_sindet$Q0.75-tabla_sindet$Q0.25)
print(as.character(datos[[1]][which(datos$Tasa.sin.detenidos>ualto)]))
```

```
## [1] "ARICA Y PARINACOTA" "TARAPACA"
```

```
#Sin detenidos
ualto=tabla_condet$Q0.75+1.5*(tabla_condet$Q0.75-tabla_condet$Q0.25)
print(as.character(datos[[1]][which(datos$Tasa.con.detenidos>ualto)]))
```

```
## [1] "TARAPACA"    "ANTOFAGASTA"
```

Los resultados indican que las regiones del norte de nuestro país (Arica y Parinacota, Tarapacá y Antofagasta) tienen una tasa de delitos que escapa de la del resto de las regiones en relación al número de habitantes.

Actividades

- Averiguar qué son los **valores atípicos** (superiores e inferiores) y cómo identificarlos.
- Repetir esta actividad para su conjunto de datos e incluya un análisis usando los conceptos estudiados en las clases anteriores.