

Estadística Aplicada I (ESMA3101)

Prof. Pedro A. Torres Saavedra

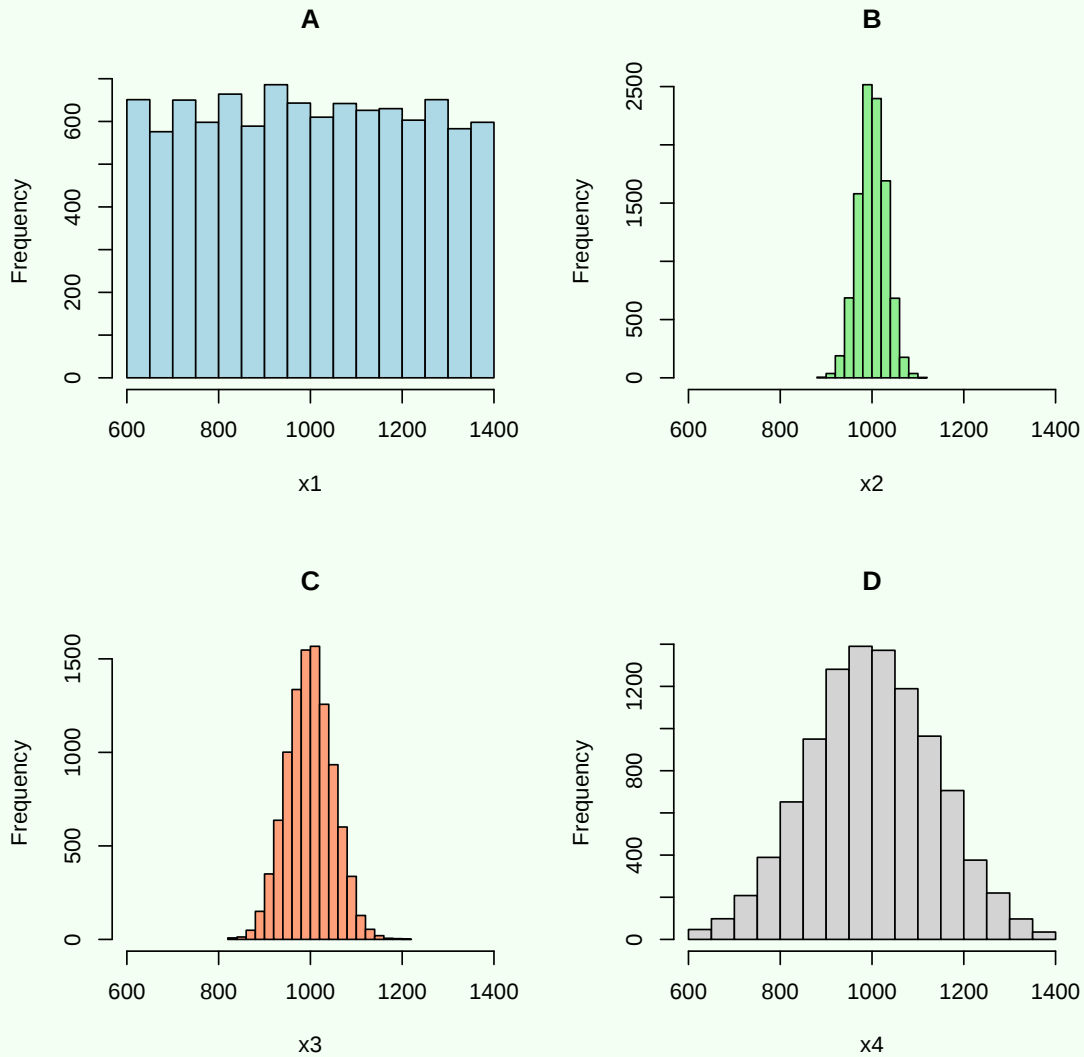
March 6, 2014

Lección 10-11: Medidas de Dispersión

Las medidas de tendencia central buscan encontrar un valor numérico para representar un individuo "típico" en la población (centro de la distribución). Las medidas de dispersión en cambio buscan determinar que tan dispersos están los datos. En otras palabras, que tanta variabilidad hay en los datos.

Ejemplo

Antes de introducir las medidas de dispersión veamos cómo podemos ver que tan dispersos están los datos usando histogramas. ¿Cuál de los siguientes histogramas muestra mayor dispersión en los datos? Ordenelos de mayor a menor dispersión.



§

Amplitud (Rango)

La amplitud o el rango se define como la diferencia entre el máximo valor y el mínimo valor. Básicamente indica sobre cuántas están dispersos los datos.

Ejemplo

Calcule a mano la amplitud (rango) de los siguientes datos {6, 3, 8, 6, 4}. Repita el cálculo usando R. Verifique que las dos cantidades sean iguales.

§

```
# Calcula la amplitud de los datos
misdatos = c(6,3,8,6,4)
max(misdatos) - min(misdatos)

[1] 5
```

Este valor nos dice que todos los datos caen dentro de 5 unidades.

Ahora vamos a utilizar la misma base de datos del estudio observacional sobre nutrición con $n = 315$ individuos (ver notas de la Lección 5).

```
# Lee la base de datos
install.packages("RCurl", repos="http://cran.us.r-project.org")
library(RCurl)
tt = getForm("https://docs.google.com/spreadsheet/pub",
             hl = "en_US",
             key = "0AhpzM-gDQ-UcdHB2TnVwakd4ZHJQVU4yMDdoWDdVWEE",
             output = "csv",
             .opts = list(followlocation = TRUE, verbose = TRUE,
                          ssl.verifypeer = FALSE))
dieta <- read.csv(textConnection(tt))
View(dieta)
attach(dieta)
```

Ejercicio

Calcule la amplitud de las siguientes variables de la base de datos de nutrición: BETADIET, CALORIES y RETDIET.

▶

```
# Calcula la amplitud de los datos
min(BETADIET)

[1] 214

max(BETADIET)

[1] 9642
```

```
# Amplitud
max(BETADIET) - min(BETADIET)

[1] 9428

min(CALORIES)

[1] 445.2

max(CALORIES)

[1] 6662

# Amplitud
max(CALORIES) - min(CALORIES)

[1] 6217

min(RETDIET)

[1] 30

max(RETDIET)

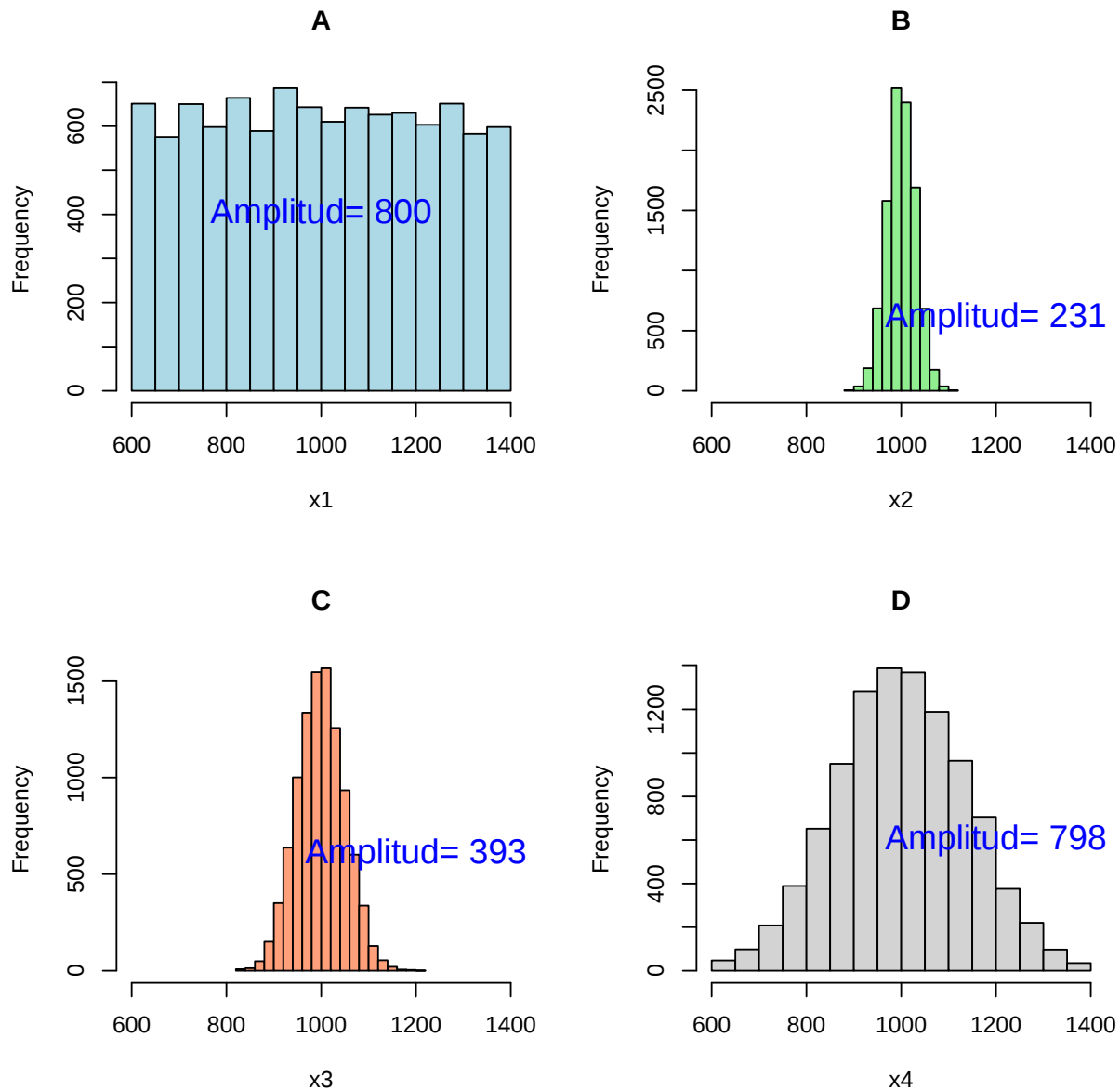
[1] 6901

# Amplitud
max(RETDIET) - min(RETDIET)

[1] 6871
```

Por ejemplo, podemos decir que los individuos reportaron consumos de calorías en una amplitud de 6,217 calorías (el mínimo valor reportado es 445.2 y el máximo es 6,662).

Calculemos el rango de los datos correspondientes a los histogramas A, B, C y D.



Varianza Muestral

Suponga que tenemos los siguientes n datos $\{X_1, X_2, \dots, X_n\}$, donde X representa la variable de interés. La varianza muestral de X , denotada por s^2 , se define como:

$$s^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n - 1} = \frac{\text{Suma de desviaciones al cuadrado}}{\text{Total de datos menos uno}}.$$

La cantidad $(X_i - \bar{X})$ es la desviación del dato X_i con respecto a la media $\bar{X}$.

Ejemplo

Calcule a mano la varianza de los siguientes datos {6, 3, 8, 6, 4}. Repita el cálculo usando R. Verifique que las dos cantidades sean iguales.

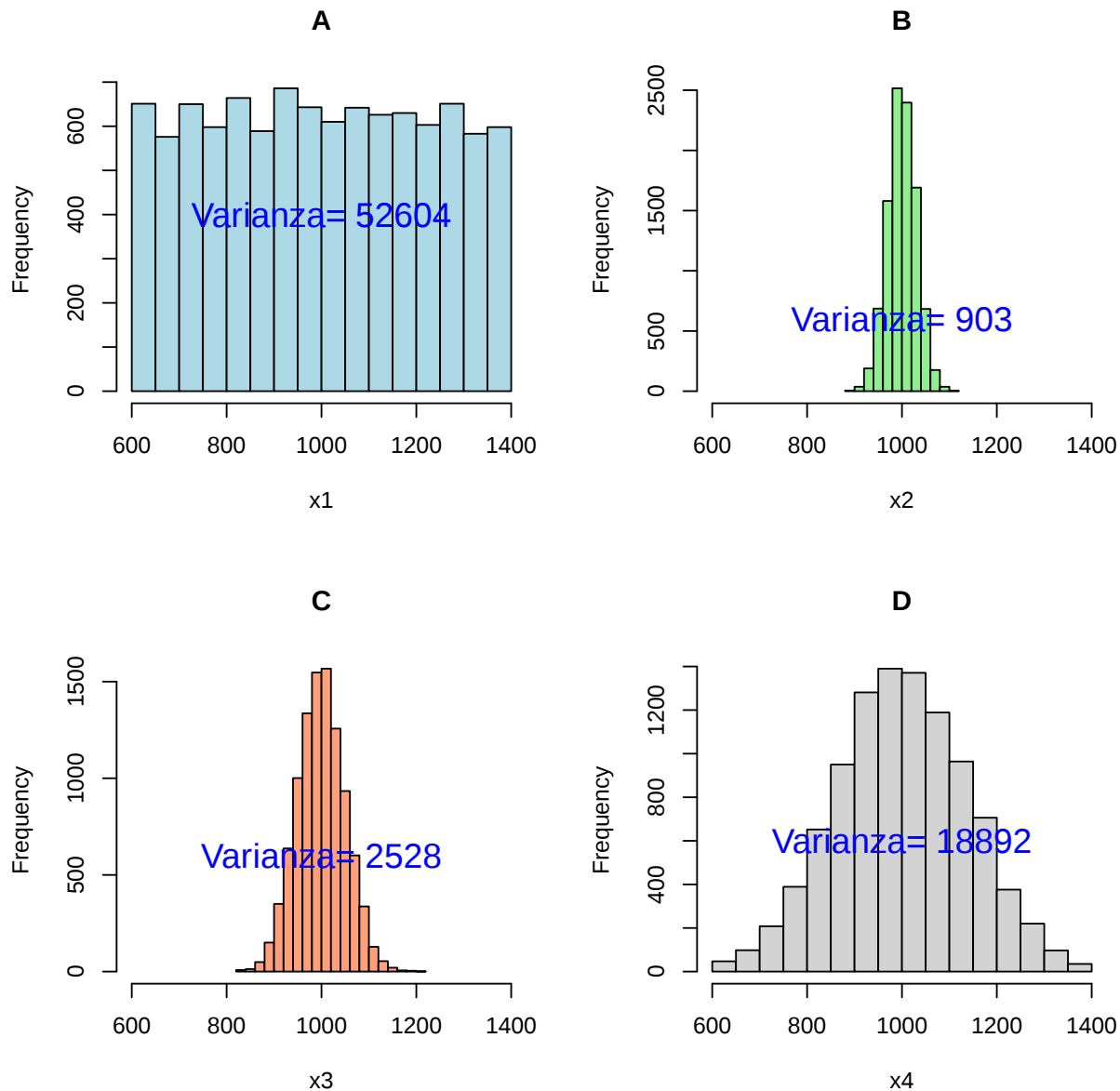
§

```
# Calcula la varianza de los datos
misdatos = c(6,3,8,6,4)
var(misdatos)

[1] 3.8
```

Por ahora sólo podemos decir que la varianza de los datos es igual a 3.8 (*unidades²*).

Calculemos la varianza de los datos correspondientes a los histogramas A, B, C y D vistos previamente.



Claramente la varianza de los datos en el histograma A es mayor que las varianzas de los demás conjuntos de datos. Es decir, los datos en el histograma A están más dispersos alrededor de la media. Note que el rango nos estaba diciendo que la variabilidad de los datos en el histograma A y D era similar, lo cuál no es cierto si las varianzas son usadas. **Es decir, la varianza mide mejor la dispersión de los datos alrededor de la media.**

Ejercicio

Calcule la varianza de las siguientes variables de la base de datos de nutrición: BETADIET, CALORIES y RETDIET.



```
# Calcula la varianza de los datos
```

```
var(BETADIET)
```

```
[1] 2172342
```

```
var(CALORIES)
```

```
[1] 462873
```

```
var(RETDIET)
```

```
[1] 347262
```

Por ejemplo, la varianza del consumo diario de retinol es igual a 347,262 *microgramos*² (note que las unidades de la variable están elevadas al cuadrado!).

Desviación Estándar Muestral

En la sección introducimos la varianza como una medida de dispersión pero mencionamos que la varianza es expresada en *unidades*². Por esta razón, se define otra medida de dispersión conocida como desviación estándar muestral, denotada por s , como

$$s = \sqrt{s^2}.$$

Ejemplo

Calcule a mano la varianza de los siguientes datos {6, 3, 8, 6, 4}. Repita el cálculo usando R. Verifique que las dos cantidades sean iguales.

§

```
# Calcula la desviación estándar
```

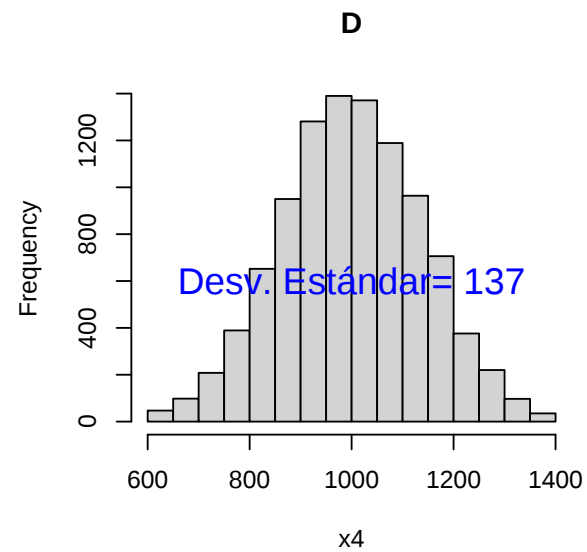
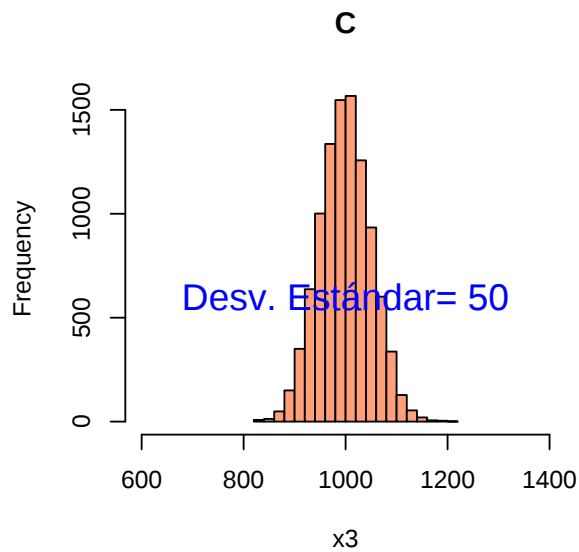
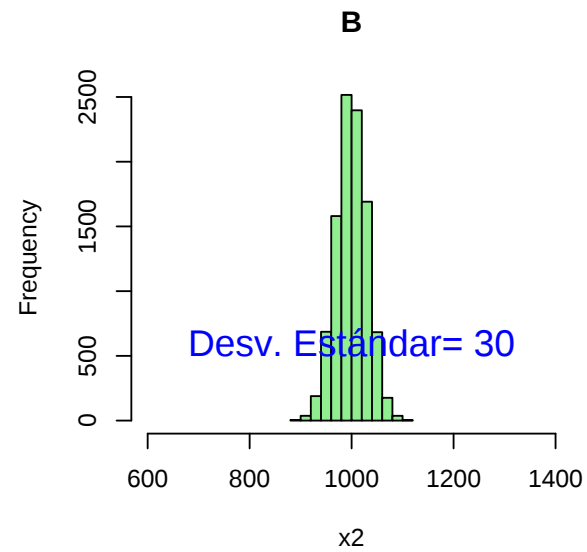
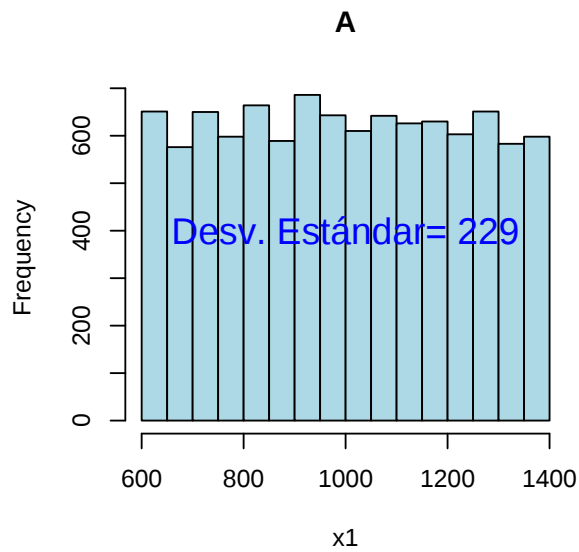
```
misdatos = c(6,3,8,6,4)
```

```
sd(misdatos)
```

```
[1] 1.949
```

Por ahora sólo podemos decir que la desviación estándar de los datos es $s = 1.949$. Note que $s = \sqrt{s^2} = \sqrt{3.8}$.

Calculemos la desviación estándar de los datos correspondientes a los histogramas A, B, C y D vistos previamente.



Ejercicio

Calcule la desviación estándar de las siguientes variables de la base de datos de nutrición: BETADIET, CALORIES y RETDIET.

```
# Calcula la desviación estándar
sd(BETADIET)

[1] 1474

sd(CALORIES)
```

```
[1] 680.3  
  
sd(RETDIET)  
  
[1] 589.3
```

Por ejemplo, la desviación estándar del consumo diario de retinol es igual a 6,871 *microgramos* (note que las unidades de la variable están en la escala original-microgramos).

Comparando la Dispersión (Variabilidad) de Dos o Más Grupos

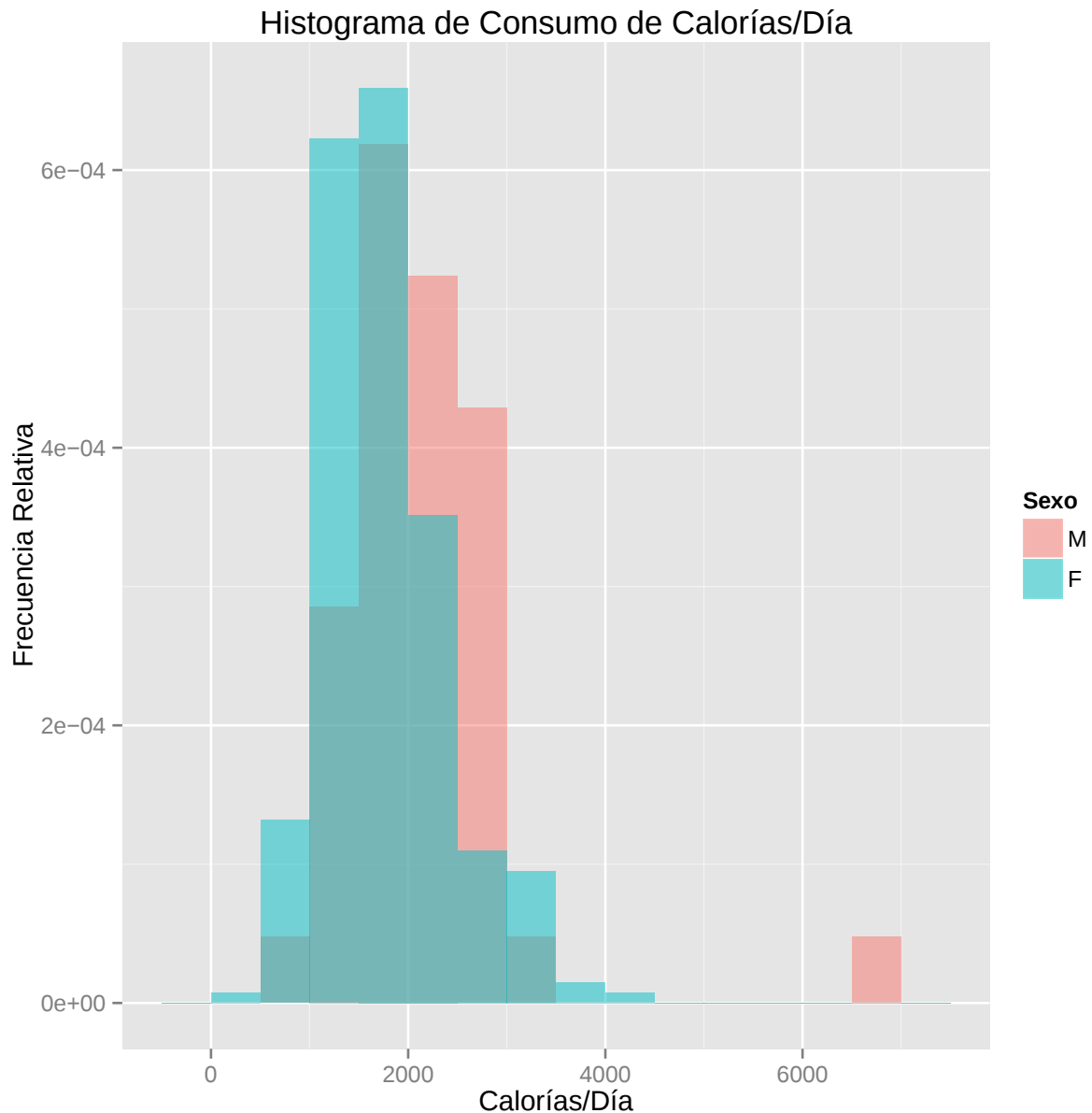
La desviación estándar es muy útil para comparar la variabilidad entre dos o más grupos. Es decir, podemos concluir en que grupos los valores de la variable son más parecidos entre sí (homogéneos).

Calculemos la desviación estándar para el consumo de calorías para hombre y mujeres.

```
# Calcula la desviación estándar por sexo  
SEX = factor(SEX, levels=c(1,2), labels=c("Masculino", "Femenino"))  
by(CALORIES, SEX, sd)  
  
SEX: Masculino  
[1] 916.6  
-----  
SEX: Femenino  
[1] 620.3
```

La desviación estándar de las calorías en mujeres y hombres es igual a 620.3 y 916.6, respectivamente. Así, el consumo de calorías en los hombres muestra mayor variabilidad (dispersión) alrededor de la media. Veamos si la comparación de los grupos es más evidente cuando usamos histogramas.

```
library(ggplot2)  
ggplot(dieta, aes(CALORIES, fill = factor(SEX, labels=c("M", "F")))) +  
  geom_histogram(alpha = 0.5, aes(y = ..density..),  
                 position = 'identity', binwidth = 500) +  
  labs(title="Histograma de Consumo de Calorías/Día",  
       y="Frecuencia Relativa", x="Calorías/Día") +  
  scale_fill_discrete("Sexo")
```



El histograma de los hombres se extiende más a la derecha, indicando mayor dispersión que en las mujeres. Es más, hay un hombre que reportó consumir más de 6,000 calorías en un día. Veamos que sucede cuando recalculamos las desviaciones estándar y hacemos los histograma eliminando ese individuo del análisis.

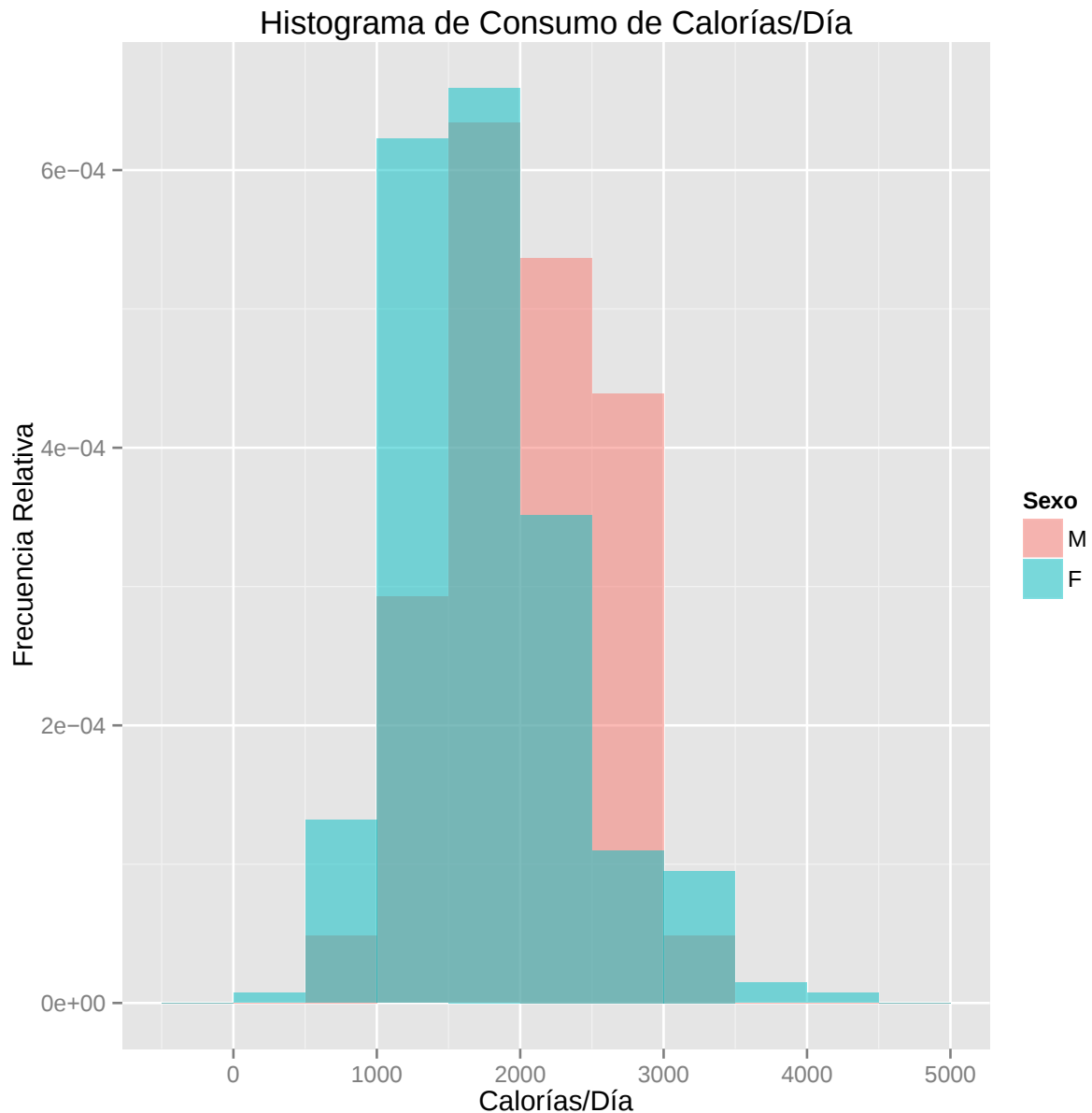
```
# Calcula la desviación estándar por sexo
# incluyendo solamente individuos con calorías<6000
by(CALORIES[CALORIES<6000], SEX[CALORIES<6000], sd)

SEX[CALORIES < 6000]: Masculino
[1] 584
```

```
SEX[CALORIES < 6000]: Femenino  
[1] 620.3
```

Note que la desviación estándar para hombres es ahora igual a 584 calorías, menor que la desviación estándar de mujeres 620.3 calorías. La conclusión es que no importa la medida de tendencia central o dispersión que usemos, un análisis gráfico de los datos es siempre recomendado. En este caso un sólo individuo (outlier?) está cambiando la conclusión a la que estamos llegando.

```
library(ggplot2)  
ggplot(dieta,  
       aes(CALORIES,  
           fill = factor(SEX, labels=c("M", "F")))) +  
  geom_histogram(data=subset(dieta, CALORIES<6000), alpha = 0.5,  
                aes(y = ..density..),  
                position = 'identity', binwidth = 500) +  
  labs(title="Histograma de Consumo de Calorías/Día",  
        y="Frecuencia Relativa", x="Calorías/Día") +  
  scale_fill_discrete("Sexo")
```



Los histogramas apoyan la conclusión que hay más variabilidad en mujeres que en hombres. Veamos que tan diferentes son las medias.

```
# Calcula la media por sexo
# incluyendo solamente individuos con calorías<6000
by(CALORIES[CALORIES<6000], SEX[CALORIES<6000], mean)

SEX[CALORIES < 6000]: Masculino
[1] 2046

-----
SEX[CALORIES < 6000]: Femenino
[1] 1741
```

En promedio, los hombres consumen 2,046 calorías al día mientras que las mujeres consume 1,741 calorías al día. El consumo de calorías promedio si el individuo que consumió más de 6,000 calorías no hubiese sido eliminado es igual a 2,156!.

```
# Calcula la desviación estándar por sexo
by(CALORIES, SEX, mean)
```

```
SEX: Masculino
[1] 2156
```

```
-----
SEX: Femenino
[1] 1741
```

Un aspecto importante en la discusión anterior es que el tamaño de muestra en hombres es pequeño. De acuerdo con la salida mostrada abajo hay 42 hombres. Así, que **los datos atípicos ("outliers") tienden a afectar más el promedio y la desviación cuando el tamaño de muestra es pequeño.**

```
# Frecuencias por sexo
table(SEX)
```

```
SEX
Masculino  Femenino
         42         273
```

Recursos Adicionales

- Sección 3.2 del libro de texto.