

Ejercicio resuelto 7.1

Los siguientes datos se corresponden con los retrasos (en minutos) de una muestra de 30 vuelos de cierta compañía aérea por semana y 35 durante los fines de semana.

Por semana														
12	25	3	16	11	7	21	54	19	6	11	23	21	7	13
14	187	15	21	7	5	17	21	13	26	21	19	8	24	16

Fines de semana														
16	11	3	12	21	13	17	12	24	21	14	18	7	11	62
13	17	19	48	11	9	13	21	126	11	17	8	23	15	52
11	14	17	71	12										

- A grandes rasgos, ¿se parecen las distribuciones de retrasos por semana y en fin de semana?, ¿crees que hay datos atípicos?
- ¿Alrededor de qué valor dirías que oscilan los retrasos?
- Compara la dispersión de los retrasos por semana y los fines de semana.
- ¿Se parecen las distribuciones a la normal? Si se consideran únicamente el 90 % de los datos más bajos en cada caso, ¿qué distribución se parece más a la normal en cuanto a su forma?

Solución:

El **objetivo del Apartado a)** es determinar si las distribuciones de retrasos por semana y en fin de semana se parecen a grandes rasgos y si hay datos atípicos.

Planteamiento: el experimento consiste en seleccionar vuelos (individuos) y observar su retraso en minutos (variable). Se distinguen dos poblaciones: los vuelos por semana y los vuelos de fin de semana. Se tiene una muestra de 30 vuelos por semana y 35 de fin de semana. Los posibles valores de la variable son $\{9, 15, 18, \dots\}$. En principio podrían ser todos los números, luego sería una variable cardinal continua. Como la precisión excluye los decimales podría pensarse que está discretizada, pero al haber muchos valores distintos, en la práctica, está actuando como continua de todas formas. Se tienen los datos de la muestra completa.

Método y justificación: como se trata de comparar la distribución a grandes rasgos lo mejor es hacer un gráfico. Como la variable es continua, el

gráfico más adecuado es el histograma, aunque como también hay que explorar los datos y localizar valores atípicos, se debe realizar primero un gráfico de cajas.

Cálculos: la Figura 7.4 muestra los gráficos de cajas de los retrasos durante la semana y durante los fines de semana en la misma escala (la mayor parte de los paquetes estadísticos permiten esta opción cuando se introducen los datos en el mismo fichero especificando el nivel del factor -en este caso semana o fin de semana- al que pertenece cada individuo). Los histogramas se representan en la Figura 7.5.

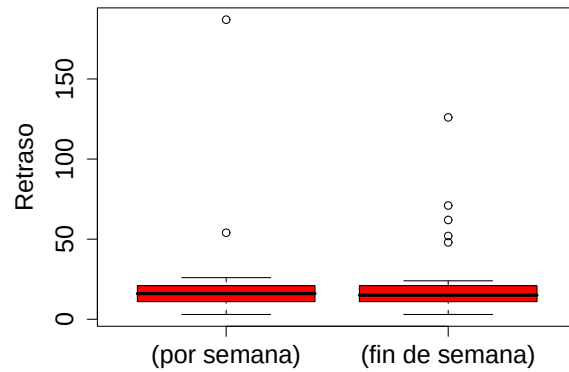


Figura 7.4: Gráfico de cajas.

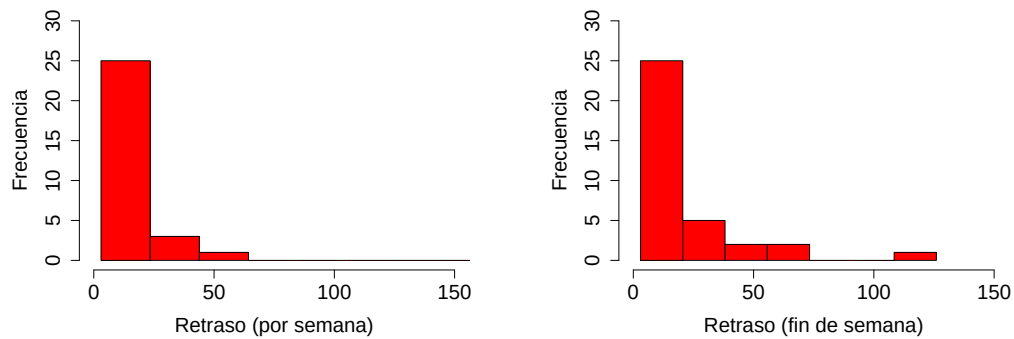


Figura 7.5: Histogramas.

Como los datos extremos dificultan un poco la visión general de la distribución del resto de los datos, se pueden representar, por ejemplo, los vuelos con retrasos inferiores a 30 para apreciar mejor la distribución en esa zona (ver Figuras 7.6 y 7.7).

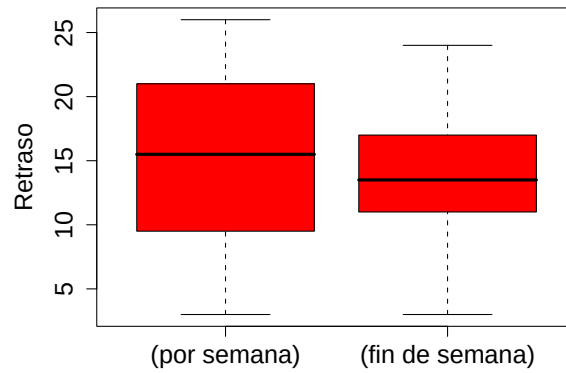


Figura 7.6: Gráfico de cajas sin valores atípicos.

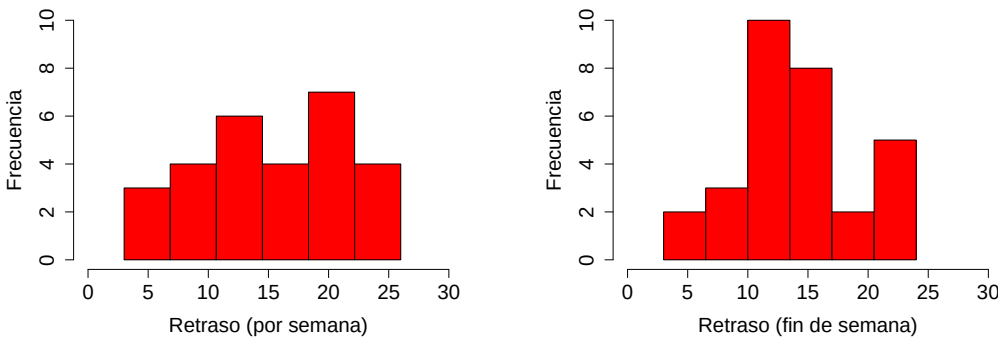


Figura 7.7: Histogramas sin valores atípicos.

Conclusión: en la Figura 7.4 se observan algunos valores atípicos, especialmente un retraso de casi 200 minutos en un vuelo durante la semana. Durante el fin de semana también hay algunos datos extremos, pero muestran una continuidad más natural que durante la semana. Esto también se observa en la Figura 7.5. Si bien en ambos histogramas se aprecia una concentración alta en retrasos pequeños, la distribución parece que se desplaza hacia valores altos con más frecuencia durante los fines de semana. En cualquier caso, la presencia de datos atípicos hace que la escala sea demasiado grande para poder apreciar bien el comportamiento en valores bajos.

Al considerar la distribución de los vuelos con retrasos en un rango en el que no hay datos atípicos (Figuras 7.6 y 7.7) se percibe mucho mejor la distribución. En la Figura 7.6 se observa que el retraso mediano es ligeramente superior por semana, pero también hay una variabilidad mayor (altura de la caja). La diferencia de variabilidad en los valores centrales también se observa en la Figura 7.7, ya que en estos histogramas se aprecia que los retrasos moderados por semana están más repartidos, mientras que en los fines de semana están más concentrados en el centro.

El **objetivo del Apartado b)** es determinar el valor alrededor del cual oscilan los retrasos. El **planteamiento** es el mismo que el del Apartado a).

Método y justificación: el valor alrededor del cual oscilan los retrasos se corresponderá con un valor central. Como la variable es cardinal, la medida más conveniente sería la media, si no hubiera datos extremos. En el Apartado a) se encontraron datos extremos que afectarían mucho al valor de la media, por lo que es aconsejable recurrir a las medidas robustas. Se podría calcular la mediana o la media recortada.

La mediana sólo tiene en cuenta el orden de los valores y no su magnitud, por lo que puede resultar poco informativa. La media recortada descarta sistemáticamente la información de los valores altos o bajos.

Como en esta ocasión se ha realizado un análisis exploratorio y se sabe que los valores atípicos son altos, se puede recortar sólo por esa zona. La muestra no es muy grande y para eliminar los valores extremos se necesita recortar alrededor del 10 %.

Para ilustrar las similitudes y diferencias en este caso se calcularán las 3 medidas (media, mediana y media recortada al 10 % **por arriba**).

Cálculos: aplicando los métodos vistos anteriormente o con ayuda de un programa estadístico es sencillo obtener las medidas de la Tabla 7.1.

Medida	Por Semana	Fin de semana
Media	22,1	22,5714
Mediana	16	15
Media recortada	14,667	16,594

Tabla 7.1: Medidas de centro.

Conclusión: en la Tabla 7.1 se observa una clara diferencia entre la media y las medidas robustas, debida a los datos atípicos. La media no es una buena referencia del centro. La mediana y la media recortada dan un valor similar. Cualquiera de las dos refleja bien el centro de ‘la mayor

parte de la distribución', es decir, de los retrasos habituales, sin tener en cuenta que en ocasiones se producen retrasos fuera de lo normal.

La media recortada da un valor algo más afinado, ya que tiene en cuenta las magnitudes. Como se ha considerado un recorte del 10 % en ambos casos y hay más valores concentrados en la zona alta del fin de semana (por semana se quitó incluso algún valor de la zona normal, en cambio se conservaron todos los bajos) el valor por semana resulta algo inferior.

Se puede concluir por tanto que los retrasos habituales oscilan alrededor de aproximadamente 14,667 minutos por semana y 16,594 minutos los fines de semana.

El **objetivo del Apartado c)** es comparar la dispersión de los retrasos por semana y los fines de semana. El **planteamiento** es el mismo que el del Apartado a).

Método y justificación: como en el Apartado b) se eligieron como medidas centrales la mediana o, preferiblemente, la media recortada, se necesita cuantificar la dispersión alrededor de esos puntos.

A modo ilustrativo se calcularán todas ellas, aunque para la elección realizada anteriormente es suficiente calcular la dispersión recortada.

Como los valores centrales son cantidades similares, se podría realizar la comparativa a grandes rasgos en términos absolutos (ya que las magnitudes pueden considerarse comparables), aunque en general conviene relativizar.

Cálculos: aplicando los métodos vistos anteriormente o con ayuda de un programa estadístico es sencillo obtener las medidas de dispersión absoluta de la Tabla 7.2 y las de dispersión relativa de la Tabla 7.3.

Medida	Por Semana	Fin de semana
AI	10	10
MEDA	5	4
Dispersión recortada	6,313	9,820

Tabla 7.2: Medidas de dispersión absolutas robustas.

Conclusión: en la Tabla 7.2 aunque la amplitud intercuartílica es mayor durante la semana cuando no se consideran los datos atípicos (ver Figura 7.6), resulta igual cuando se consideran todos los valores (ver Figura 7.4). Si se mide la dispersión alrededor de la mediana con la MEDA, el valor absoluto es ligeramente inferior para los fines de semana. Como las medianas eran similares, la conclusión en términos relativos también es

Medida	Por Semana	Fin de semana
AI/Me	0,625	0,667
MEDA/Me	0,3125	0,267
Dispersión recortada/ Media recortada	0,430	0,592

Tabla 7.3: Medidas de dispersión relativas robustas.

similar. La variabilidad relativa de los retrasos con respecto a la mediana es ligeramente inferior los fines de semana (26,7 % frente a 31,25 %). Esto sugiere que la MEDA capta mejor las diferencias que se apreciaban en la Figura 7.4.

Si se considera la media recortada, la situación cambia, por los motivos explicados en el Apartado b). Se concluiría que la dispersión relativa de los retrasos habituales es de aprox. un 43 % por semana frente a un 59 % los fines de semana.

Los valores relativos no pueden compararse entre sí para decir si es más representativa la mediana o la media recortada, porque están basadas en distintas metodologías y cuantifican la dispersión de formas no comparables.

El **objetivo del Apartado d)** es determinar si las distribuciones se parecen a la normal y cuál se parece más en cuanto a la forma si se consideran únicamente el 90 % de los datos más bajos en cada caso. El **planteamiento** es el mismo que el del Apartado a).

Método y justificación: como se aprecia en la Figura 7.5 en comparación con la Figura 7.1, las distribuciones no se parecen a la normal. Las variables con distribución normal no tienen valores excesivamente altos o bajos. Además, como se comentó anteriormente, las medidas de forma para comparar con la distribución normal tienen sentido únicamente si no hay datos extremos.

Si se considera únicamente el 90 % de los datos más bajos se eliminan los datos extremos y en este caso sí tiene sentido plantearse las medidas de forma para establecer la comparación. Se calculará primero la asimetría, para ver si la distribución por encima y por debajo de la media (recortada) es similar, como en la distribución normal. A continuación se calculará la curtosis para comprobar si las distribuciones se concentran en torno a la media (recortada) con la forma de campana típica de la distribución normal o, por el contrario, son más apuntadas o planas.

Cálculos: aplicando las fórmulas vistas anteriormente o con ayuda de un programa estadístico es sencillo obtener las medidas de forma de la Tabla 7.4.

Medida	Semana	Fin de semana
Asimetría	-0,143	2,326
Curtosis	-1,175	5,767

Tabla 7.4: Medidas de forma.

Conclusión: en la Tabla 7.4 se observa un coeficiente de asimetría para los retrasos durante la semana muy próximo a 0 pero es negativo, con lo que podría decirse que esta distribución es prácticamente simétrica, aunque ligeramente asimétrica a la izquierda. En relación con los retrasos durante el fin de semana el coeficiente está bastante más alejado de 0 y además es positivo lo que indica que la distribución es asimétrica a la derecha, es decir que los valores altos están más alejados de la media y los bajos se concentran más alrededor de ésta. Por otro lado, atendiendo al coeficiente de curtosis los valores obtenidos en ambos casos están muy alejados del 0. Para los retrasos por semana es negativo, lo que la distribución es menos apuntada que la normal, mientras que para los retrasos durante los fines de semana el coeficiente de curtosis es mayor que durante la semana y además es positivo, lo que nos indica que la distribución es más apuntada que la normal. Estas situaciones ya se habían observado gráficamente en la Figura 7.7. Atendiendo a ambos criterios se puede concluir que las distribuciones de los retrasos no se parecen a la distribución normal.