Práctica 4

Trabajaremos con el archivo `Mundo95.RData`, que se debe descargar de Moodle. Dicho archivo contiene datos de 109 países. Concretamente, contiene las siguientes variables:

- `país`: nombre del país al que pertenece el caso estudiado.
- `poblac`: población del país, en miles de habitantes.
- `densidad`: densidad de población (habitantes/Km^2).
- `urbana`: porcentaje de la población que habita en ciudades.
- `espvidaf`: esperanza de vida femenina, en años.
- `espvidam`: esperanza de vida masculina, en años.
- `alfabet`: tasa de alfabetización de la población.
- `pib_cap`: producto interior bruto per cápita.
- `región`: región mundial a la que pertenece el país (OCDE, Europa Oriental, Asia/Pacífico, África, Oriente Medio, América Latina).
- `calorías`: ingesta diaria media de calorías por habitante.
- `nac_def`: tasa nacimientos/defunciones.
- `fertilid`: tasa de fertilidad (número de nacimientos por cada mujer).
- `religión`: religión mayoritaria en cada país.
- `región_cod`: región codificada (OCDE: 1, Europa Oriental: 2, Asia/Pacífico:3, África:4, Oriente Medio:5, América Latina:6).

```
> summary(Datos)
             país         poblac           densidad          urbana          espvidaf
 Acerbaján       :  1   Min.   :    256   Min.   :   2.3   Min.   :  5.00   Min.   :43.00
 Afganistán      :  1   1st Qu.:   5100   1st Qu.:  29.0   1st Qu.: 40.75   1st Qu.:67.00
 Alemania        :  1   Median :  10400   Median :  64.0   Median : 60.00   Median :74.00
 Arabia Saudí    :  1   Mean   :  47724   Mean   : 203.4   Mean   : 56.53   Mean   :70.16
 Argentina       :  1   3rd Qu.:  35600   3rd Qu.: 126.0   3rd Qu.: 75.00   3rd Qu.:78.00
 Armenia         :  1   Max.   :1205200   Max.   :5494.0   Max.   :100.00   Max.   :82.00
 (Other)         :103                                      NA's   :1
    espvidam        alfabet          pib_cap                 región       calorías
 Min.   :41.00   Min.   : 18.00   Min.   :  122   OCDE           :21   Min.   :1667
 1st Qu.:61.00   1st Qu.: 63.00   1st Qu.: 1000   Europa Oriental:14   1st Qu.:2256
 Median :67.00   Median : 88.00   Median : 2995   Asia / Pacífico:17   Median :2653
 Mean   :64.92   Mean   : 78.34   Mean   : 5860   África          :19   Mean   :2754
 3rd Qu.:72.00   3rd Qu.: 98.00   3rd Qu.: 7467   Oriente Medio   :17   3rd Qu.:3226
 Max.   :76.00   Max.   :100.00   Max.   :23474   América Latina  :21   Max.   :3825
                 NA's   :2                                             NA's   :34
    nac_def          fertilid          religión    región_cod
 Min.   : 0.9231   Min.   :1.300   Católica :41   Min.   :1.00
 1st Qu.: 1.5417   1st Qu.:1.880   Musulmana:27   1st Qu.:2.00
 Median : 2.6667   Median :3.050   Protest. :16   Median :4.00
 Mean   : 3.2035   Mean   :3.563   Ortodoxa : 8   Mean   :3.55
 3rd Qu.: 4.1750   3rd Qu.:5.000   Budista  : 7   3rd Qu.:5.00
 Max.   :14.0000   Max.   :8.190   Animista : 4   Max.   :6.00
 NA's   :1         NA's   :2       (Other)  : 6
```

Figura 4.1: Mínimo, media y máximo de la tasa de alfabetización

Realizar un estudio descriptivo sobre el conjunto de datos, respondiendo a las siguientes cuestiones:

a) Calcular e interpretar las siguientes medidas descriptivas para la tasa de alfabetización: mínimo, máximo, media, percentil 30, coeficiente de asimetría y coeficiente de apuntamiento.

Del menú `Estadísticos`→`Resúmenes`→`Conjunto de datos activo`, obtenemos la salida de la figura 4.1. En la misma encontramos, entre otros, el valor de los siguientes estadísticos sobre la tasa de alfabetización:

- Mínimo: 18. En el país de la muestra con la menor tasa de alfabetización observada, esta es del 18 %.
- Máximo: 100. En el país de la muestra con la mayor tasa de alfabetización observada, esta es del 100 %.
- Media: 78.34. La media de las tasas de alfabetización de los países de la muestra es del 78.34 %.

Por otra parte, en el menú `Estadísticos`→`Resúmenes`→`Resúmenes numéricos` seleccionamos, en el marco de la pestaña `Datos`, la variable *alfabet*. En la pestaña `Estadísticos` seleccionamos los estadísticos *asimetría y apuntamiento (tipo 1)* e incluimos en el marco de los cuantiles el cuantil *.3*, que corresponde al percentil 30 (ver figura 4.2). Obtenemos entonces la salida de la figura 4.3, en la que encontramos el valor de los restantes estadísticos.

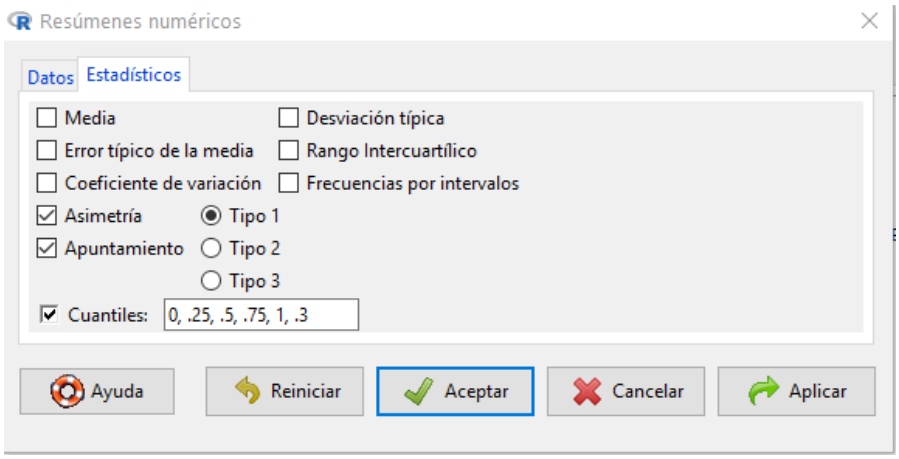

Figura 4.2: Selección de estadísticos

```
> numSummary(Datos[,"alfabet", drop=FALSE], statistics=c("quantiles", "skewness",
+     "kurtosis"),quantiles=c(0,.25,.5,.75,1,.3), type="1")
   skewness   kurtosis 0% 25% 50% 75% 100%  30%   n NA
-0.9802928 -0.2077413 18  63  88  98  100 72.8 107  2
```

Figura 4.3: Asimetría, apuntamiento y percentil 30 de la tasa de alfabetización

- El valor del coeficiente de asimetría es -0.9803. Esto significa que la distribución es asimétrica negativa o sesgada a la izquierda. En nuestro caso, que la distribución sea sesgada a la izquierda implica que hay mayoría de países que tienen tasa de alfabetización superior a la tasa de alfabetización media.
- El valor del coeficiente de apuntamiento es -0.2077. Al ser negativo, la distribución es platicúrtica (menos apuntada que la distribución normal).
- El valor del percentil 30 es 72.8. Esto significa que en la muestra hay un 30 % de países con una tasa de alfabetización igual o inferior al 72.8 %.

b) ¿Qué datos presentan una mayor dispersión relativa, los correspondientes a la esperanza de vida femenina o los datos de la tasa de alfabetización?

Para comparar la dispersión relativa de ambas variables calcularemos y compararemos sus *coeficientes de variación*. Para ello, seleccionamos el menú `Estadísticos`→`Resúmenes`→ `Resúmenes numéricos` y seleccionamos las variables *espvidaf* y *alfabet*. En la pestaña `Estadísticos`, marcamos la opción `Coeficiente de variación` y pulsamos el botón `Aceptar`.

En la salida observamos que el valor del coeficiente de variación es 0.2921 para la tasa de alfabetización y 0.1507 para la esperanza de vida femenina. Dado que el coeficiente de variación se expresa en tantos por ciento, podemos decir que la tasa de alfabetización presenta un coeficiente de variación del 29.21 % y la esperanza de vida, del 15.07 %. En consecuencia, la tasa de alfabetización presenta una mayor dispersión relativa.

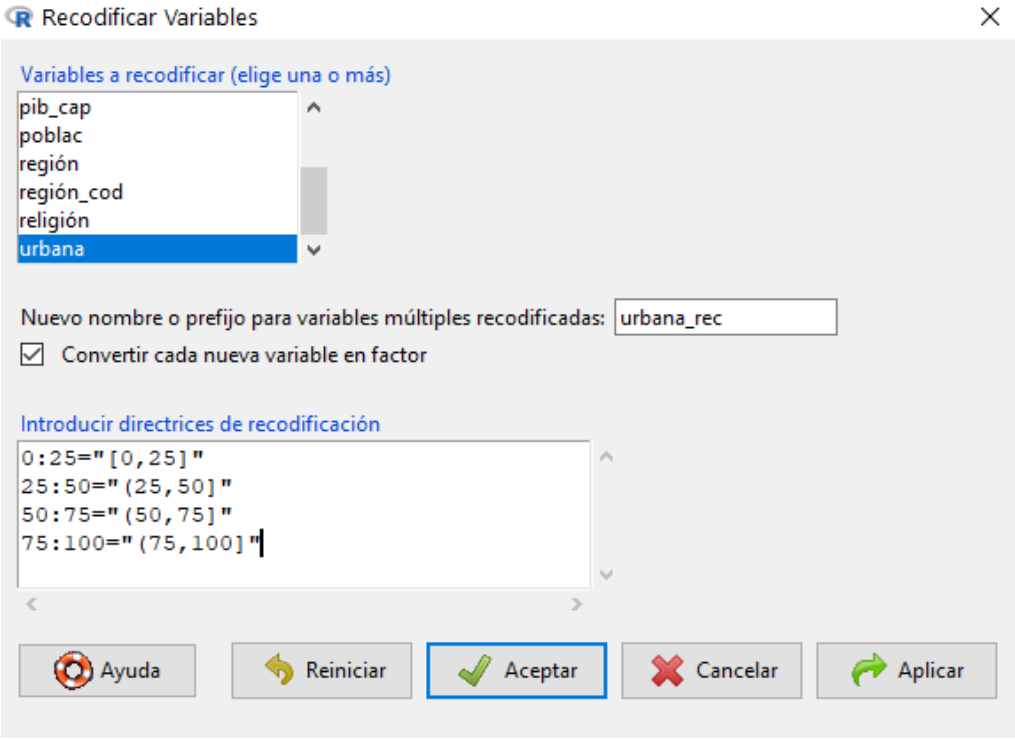

Figura 4.4: Recodificación de la variable *urbana* en cuatro intervalos

c) Recodificar la variable *urbana* en otra variable de nombre *urbana_rec*, mediante cuatro intervalos de igual amplitud, comenzando en 0 y terminando en 100. Construir tablas de contingencia adecuadas, para las variables *urbana_rec* y *región*, para responder a las siguientes preguntas.

Antes de resolver los distintos apartados procederemos a recodificar la variable urbana. Para ello seleccionamos el menú `Datos`→`Modificar variables del conjunto de datos activo`→`Recodificar variables`, seleccionando las opciones e introduciendo las directrices de recodificación que se indican en la figura 4.4.

En todos los apartados comenzaremos por construir una tabla de doble entrada desde el menú `Estadísticos`→`Tablas de contingencia`→`Tablas de doble entrada`, seleccionando la variable *urbana_rec* en las filas y *región* en las columnas. Adicionalmente será necesario realizar las acciones que se indican en cada apartado.

c1) ¿Qué porcentaje de países, de los que tienen un porcentaje en ciudades entre el 25 % y el 50 %, se encuentran en América Latina?

En la pestaña `Estadísticos`, de la ventana `Tabla de doble entrada`, marcamos la opción `Porcentaje por filas` y pulsamos el botón `Aceptar`. Encontramos que el porcentaje es del 25.9 %.

c2) ¿Cuál es el porcentaje de países de Oriente Medio en los que más del 50 % de la población vive en ciudades?

En la pestaña `Estadísticos`, de la ventana `Tabla de doble entrada`, marcamos la opción `Porcentaje por columnas` y pulsamos el botón `Aceptar`. Observamos que si nos restringimos a la región Oriente Medio, los países que tienen un porcentaje de habitantes en ciudades entre el 50 % y el 75 % suponen el 35.3 %. Del mismo modo, dentro de esta región, los países con un porcentaje en ciudades entre el 75 % y el 100 % suponen el 41.2 %. Por tanto, el porcentaje de países de Oriente medio en los que más del 50 % de la población vive en ciudades es del 76.5 %.

c3) ¿Qué porcentaje de países pertenecen a la OCDE y tienen un porcentaje de habitantes que habitan en ciudades entre el 25 % y el 75 %?

En la pestaña `Estadísticos`, de la ventana `Tabla de doble entrada`, marcamos la opción `Porcentajes totales` y pulsamos el botón `Aceptar`. Observamos que los países que pertenecen a la OCDE y tienen un porcentaje de habitantes en ciudades entre el 25 % y el 50 %, son el 0.9 % de todos los países de la muestra. Por otra parte, los países que pertenecen a la OCDE y tienen un porcentaje de habitantes en ciudades entre el 50 % y el 75 %, son el 8.3 % del total. En consecuencia, el porcentaje de países de la muestra que pertenecen a la OCDE y tienen un porcentaje de habitantes en ciudades, entre el 25 % y el 75 %, es del 9.2 %.

d) Construir tablas de estadísticas adecuadas, que tengan como factores las variables *región* y *urbana_rec*, que permitan responder a las siguientes cuestiones sobre la ingesta diaria de calorías:

En ambos apartados, seleccionamos el menú `Estadísticos`→`Resúmenes`→`Tablas de estadísticas`. Adicionalmente, realizamos en cada apartado, las acciones que se indican.

d1) ¿Cuál es la media de las ingestas diarias de calorías por habitante para los países de la OCDE en los que más del 75 % de los habitantes vive en ciudades?

Seleccionamos, en el marco `Factores`, las variables *urbana_rec* y *región*. En el marco `Variables explicadas`, seleccionamos la variable *Calorías*. Como `Estadístico` seleccionamos la opción `Media` y pulsamos el botón `Aceptar`. En la salida observamos que la ingesta media de calorías, en los países de la OCDE con un porcentaje de habitantes en ciudades superior al 75 %, es de 3329.22 calorías.

d2) Si consideramos los países en los que el porcentaje de habitantes en ciudades se encuentra entre el 25 % y el 50 %, cuál es la región en la que la ingesta diaria por habitante presenta una mayor variabilidad?

Para comparar la variabilidad de los datos de ambas regiones podemos utilizar la varianza de los mismos o, equivalentemente, su desviación típica. Mayor variabilidad

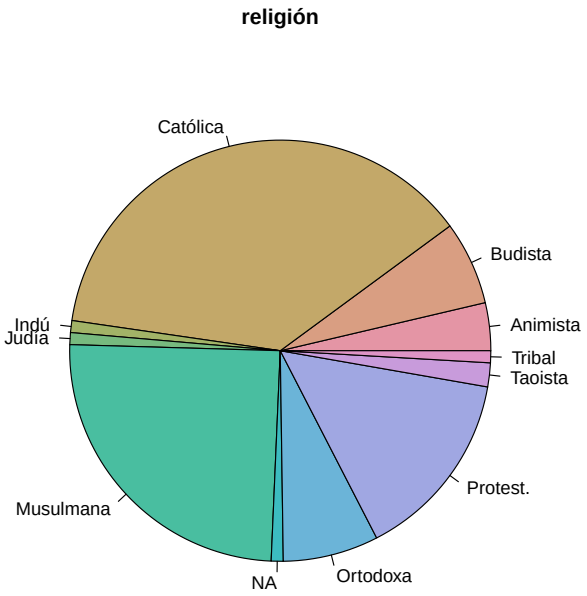

Figura 4.5: Gráfico de sectores de la variable religión

va asociada a mayor varianza/desviación típica. Por el contrario menor variabilidad va asociada a menor varianza/desviación típica.

Realizamos entonces los mismos pasos del apartado enterior pero como estadístico marcamos la opción `Desviación típica`. Si observamos la columna del intervalo (25,50], la región en la que los datos presentan una mayor desviación típica y, por lo tanto, mayor variabilidad es América Latina, con una desviación típica de 316.58 calorías.

Nótese también que, en la OCDE, Europa Oriental y Oriente medio, no se dispone del valor de la desviación típica de la ingesta diaria de calorías, dado que en estas regiones no hay países en los que el porcentaje de habitantes en ciudades se encuentra entre el 25 % y el 50 %.

e) Construir un gráfico de sectores para la variable *religión*. Determinar el porcentaje de países que corresponde a cada sector.

El gráfico de sectores se puede ver en la figura 4.5. Los porcentajes correspondientes a cada uno de los sectores los obtenemos a partir del menú `Estadísticos`→`Resúmenes`→`Distribución de frecuencias`, eligiendo la variable *religión*. Son los siguientes:

Animista: 3.67 %, Budista: 6.42 %, Católica: 37.61 %, Indú: 0.92 %, Musulmana: 24.77 %, Ortodoxa: 7.34 %, Protestante: 14.68 %, Taoista: 1.83 %, Tribal: 0.92 %. No se tienen datos sobre la religión mayoritaria en el 0.92 % de los países de la muestra.

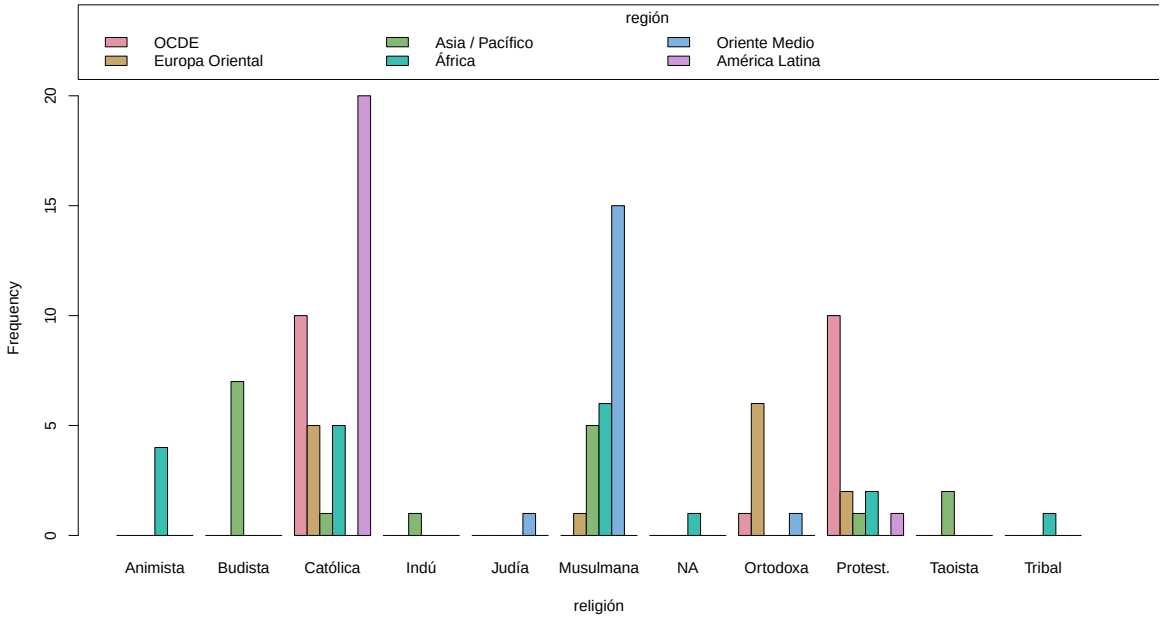

Figura 4.6: Gráfico de barras de la variable religión, agrupada por regiones

f) Construir un grafico de barras que muestre el número de países en los que cada religión es mayoritaria, agrupandolos según las distintas regiones. Elegir en la pestaña `Opciones`, de la ventana `Gráfica de barras`, la opción `Lado a lado` y comparar los resultados que se obtienen para los distintos grupos.

El gráfico obtenido es el representado en la figura 4.6. Como se puede ver en la leyenda del gráfico, el eje vertical representa la frecuencia absoluta, esto es, el número de casos observados. Por tanto, cuanto más alta sea la barra, mayor será el número de casos que representa. Por otra parte, el color de la barra indica a qué región corresponde la misma y la religión que define cada grupo se representa en el eje horizontal[1]. El recuento de casos se hace restringiéndonos a cada región. Por lo tanto, cada una de las barras del gráfico muestra el número de países en cada región que tienen como religión dominante la indicada por el color de la barra.

De este modo, se puede ver que las religiones Animista y Tribal están presentes, como religiones mayoritarias, únicamente en países de África y las religiones Budista, Indú y Taoista son religiones mayoritarias únicamente en países de Asia/Pacífico. Además, en estas regiones, no hay países con religiones mayoritarias distitas de las anteriores.

También observamos que en América Latina es donde más países hay con la religión Católica como mayoritaria. Por el contrario en Oriente Medio ningún país tiene la religión Católica como mayoritaria ya que la barra correspondiente tiene altura 0.

[1]En caso de que el gráfico no muestre las etiquetas de algunas de las religiones se debe aumentar la anchura de la ventana del gráfico, bien pinchando con el cursor en uno de sus lados y arrastrando éste para agrandar la ventana o bien maximizando la ventana para ponerla a pantalla completa

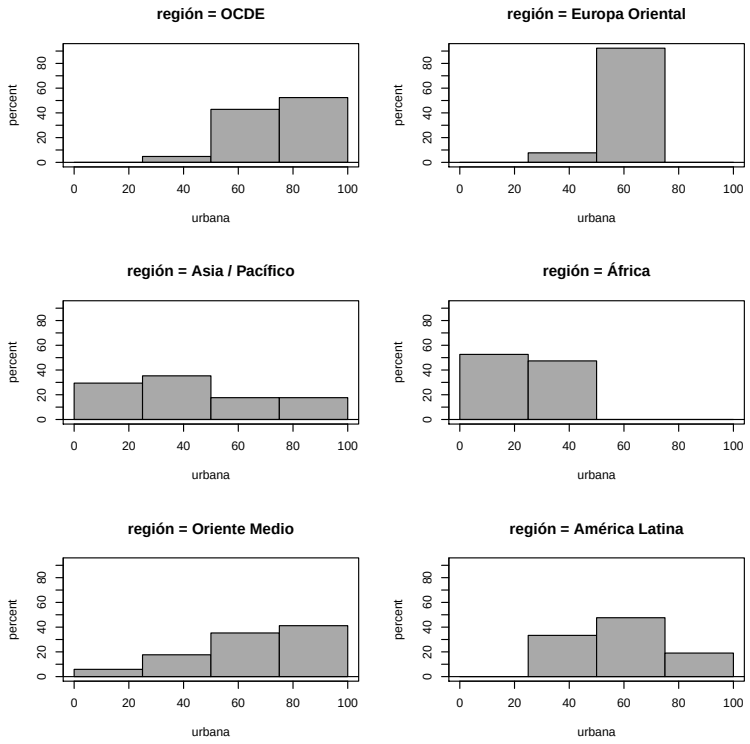

Figura 4.7: Gráfico de barras de la variable religión, agrupando por regiones

Igualmente, la región con mayor número de países de religión musulmana es Oriente medio y la región con mas países de religión Ortodoxa es Europa Oriental.

Conclusiones similares se pueden obtener para el resto de religiones y regiones.

g) Construir histogramas que permitan estudiar el porcentaje de habitantes en ciudades, agrupando los datos según las distintas regiones. Utilizar cuatro intervalos para la construcción de los histogramas que comiencen en 0 y terminen en 100.

Sugerencias: Construir el histograma usando el correspondiente menú de R-Commander y modificar la instrucción que aparece en la ventana `R Script`, indicando los intervalos deseados. En la pestaña `Opciones` del menú `Histograma`, seleccionar la opción porcentajes para que sea posible realizar la comparación entre los distintos histogramas.

La instrucción modificada que debemos ejecutar es `with(Datos, Hist(urbana, groups= región, scale="percent", breaks=c(0,25,50,75,100), col="darkgray"))`.

Cada histograma corresponde a la región indicada sobre el mismo. Podemos observar, por ejemplo, que en la OCDE hay mayoría de paises que presentan un porcentaje de habitantes en ciudades de, al menos, el 50 %. En Europa Oriental casi la totalidad de los países presentan un porcentaje de habitantes en ciudades entre el 50 y el 75 % y en África no hay países que tengan una proporción de habitantes en ciudades superior al

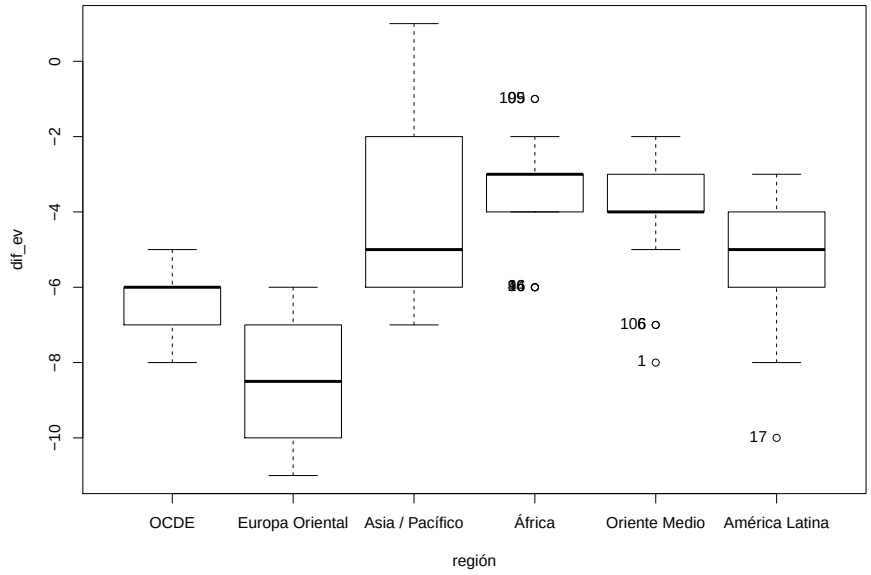

Figura 4.8: Gráfica de cajas

50 %. En Oriente Medio hay una mayor proporción de países que tienen un porcentaje alto (superior al 50 %) de habitantes que viven en ciudades.

h) Construir una variable de nombre *dif_ev* como la diferencia entre las variables *espvidam* y *espvidaf*. Representar e interpretar gráficos de cajas para la nueva variable, según las distintas regiones.

Una vez construída la nueva variable como *espvidam-espvidaf* seleccionamos, en el menú, `Gráficas→Diagrama de caja` la variable *dif_ev* y en la ventana emergente del botón `Gráfica por grupos` seleccionamos como variable de grupo la variable región. Pulsamos el botón `Aceptar` obteniendo el gráfico de la figura 4.8.

En el gráfico de la región Asia/Pacífico, la línea que corresponde a la mediana se encuentra más próxima al lado inferior de la caja que al superior y el bigote inferior es más corto que el superior. Esto nos indica que el conjunto de datos, en esta región, presenta asimetría positiva, esto es, que la distribución es sesgada a la derecha.

En el gráfico de la región América Latina, la línea de la media está centrada en la caja y el bigote superior es ligeramente más corto que el inferior. Nos indica esto que la distribución presenta una ligera asimetría negativa.

Por el contrario, en el gráfico de Europa Oriental la línea de la mediana está centrada en la caja y los bigotes tienen la misma longitud. En consecuencia, podemos decir que la distribución es simétrica.

Para el resto de regiones podemos hacer interpretaciones similares. Indicar también que los casos atípicos son los casos 16, 44, 96, 95, 109, 1, 6, 106 y 17. Aunque en el gráfico se solapan y no es posible distinguirlos, podemos encontrarlos en la ventana de salida de R-Commander.

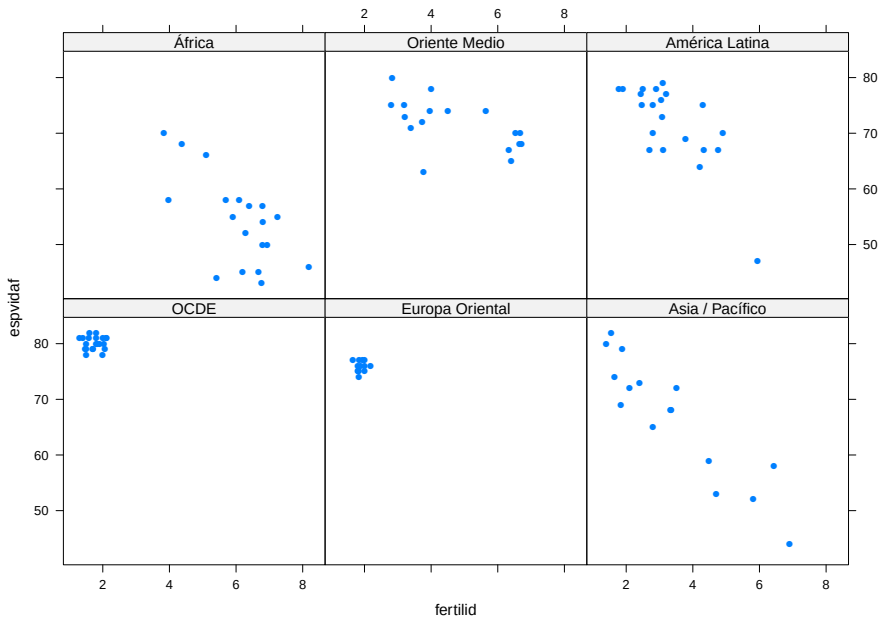

Figura 4.9: Gráfico XY

i) Construir un gráfico XY para estudiar la esperanza de vida femenina en función de la tasa de fertilidad, eligiendo como condición el valor de la variable *región*. Interpretar los resultados obtenidos.

El gráfico que se obtiene está representado en la figura 4.9. En las regiones OCDE y Europa Oriental, tanto la esperanza de vida femenina como la fertilidad presentan muy poca variabilidad por lo que la nube de puntos aparece 'colapsada' y no se aprecia que exista relación entre ambas variables.

En Oriente Medio y América Latina parece apreciarse una débil relación inversa entre las variables: una mayor tasa de fertilidad va asociada a una menor esperanza de vida, aunque hay que insistir en que la relacion entre las variables es muy débil. En América Latina, además, se aprecia un punto muy alejado del resto por lo que habría que estudiar este caso por separado porque parece se trata de una observación atípica.

En África y Asia/Pacífico también se aprecia, de manera clara, que existe relación inversa entre las variables y que la dependencia entre ambas es muy fuerte: en estas regiones, claramente, una mayor tasa de fertilidad va asociada a una menor esperanza de vida. Hay que indicar también que este hecho no implica causalidad, esto es, que la mayor tasa de fertilidad cause por si misma una menor esperanza de vida. Esta relación puede deberse a otros motivos como, por ejemplo, a las condiciones socieconómicas de los países agrupados en estas regiones.