Taller 3: Aplicación de wrangling data

J. David Ramos

Objetivo

Con base en los conjuntos de datos `who` y `population` de la librería `DSR` crear data frames en formato tidy que permitan calcular y describir las tasas de incidencia mundial de tuberculosis de los 219 países descritos entre 1995 y 2013, incluida Colombia

Estructura del data frame who

El conjunto de datos `who` está constituido por 7270 observaciones y 60 variables. La estructura del data frame es la siguiente:

```
# Carga del paquete necesarios
library(DSR)
library(tidyverse)

# Encabezado de "who"
slice(who,1:5)
```

```
## # A tibble: 5 x 60
##   country iso2  iso3   year new_sp_m014 new_sp_m1524 new_sp_m2534 new_sp_m3544
##   <chr>   <chr> <chr> <int>       <int>        <int>        <int>        <int>
## 1 Afghan~ AF    AFG    1980          NA           NA           NA           NA
## 2 Afghan~ AF    AFG    1981          NA           NA           NA           NA
## 3 Afghan~ AF    AFG    1982          NA           NA           NA           NA
## 4 Afghan~ AF    AFG    1983          NA           NA           NA           NA
## 5 Afghan~ AF    AFG    1984          NA           NA           NA           NA
## # ... with 52 more variables: new_sp_m4554 <int>, new_sp_m5564 <int>,
## #   new_sp_m65 <int>, new_sp_f014 <int>, new_sp_f1524 <int>,
## #   new_sp_f2534 <int>, new_sp_f3544 <int>, new_sp_f4554 <int>,
## #   new_sp_f5564 <int>, new_sp_f65 <int>, new_sn_m014 <int>,
## #   new_sn_m1524 <int>, new_sn_m2534 <int>, new_sn_m3544 <int>,
## #   new_sn_m4554 <int>, new_sn_m5564 <int>, new_sn_m65 <int>,
## #   new_sn_f014 <int>, new_sn_f1524 <int>, new_sn_f2534 <int>,
## #   new_sn_f3544 <int>, new_sn_f4554 <int>, new_sn_f5564 <int>,
## #   new_sn_f65 <int>, new_ep_m014 <int>, new_ep_m1524 <int>,
## #   new_ep_m2534 <int>, new_ep_m3544 <int>, new_ep_m4554 <int>,
## #   new_ep_m5564 <int>, new_ep_m65 <int>, new_ep_f014 <int>,
## #   new_ep_f1524 <int>, new_ep_f2534 <int>, new_ep_f3544 <int>,
## #   new_ep_f4554 <int>, new_ep_f5564 <int>, new_ep_f65 <int>,
## #   newrel_m014 <int>, newrel_m1524 <int>, newrel_m2534 <int>,
## #   newrel_m3544 <int>, newrel_m4554 <int>, newrel_m5564 <int>,
## #   newrel_m65 <int>, newrel_f014 <int>, newrel_f1524 <int>,
## #   newrel_f2534 <int>, newrel_f3544 <int>, newrel_f4554 <int>,
## #   newrel_f5564 <int>, newrel_f65 <int>
```

Las columnas cinco a sesenta de `who` reportan número de casos nuevos de tuberculosis por año y país. La codificación es la siguiente:

1. Las primeras tres letras de cada columna indican si la columna contiene casos nuevos de tuberculosis (`new`)
2. Las siguientes dos letras describen el tipo de caso que se cuenta:

`rel`: casos de recaída

`ep`: casos de tuberculosis extrapulmonar

`sn`: casos de tuberculosis pulmonar que no pudieron ser diagnosticados por un frotis pulmonar (frotis negativo)

`sp`: casos de tuberculosis pulmonar que podrían diagnosticarse como un frotis pulmonar (frotis positivo)

3. La sexta letra describe el sexo de los casos con tuberculosis. El conjunto de datos agrupa casos por hombres (`m`) y mujeres (`f`).
4. Los números restantes describen el grupo de edad de los casos con tuberculosis. Son siete grupos de edad:

`014`: casos de 0 a 14 años de edad

`1524`: casos que tienen entre 15 y 24 años

`2534`: casos que tienen entre 25 y 34 años

`3544`: casos que tienen entre 35 y 44 años

`4554`: casos de 45 a 54 años de edad

`5564`: casos de 55 a 64 años de edad

`65`: casos que tienen 65 años o más

Cambio de who a estructura tidy

PASO 1: se convierte `who` a formato tidy utilizando `gather()`:

```
who1<- gather(who, "codigo", "casos", 5:60)

head(who1)
```

```
## # A tibble: 6 x 6
##   country     iso2  iso3   year codigo      casos
##   <chr>       <chr> <chr> <int> <chr>       <int>
## 1 Afghanistan AF    AFG    1980 new_sp_m014    NA
## 2 Afghanistan AF    AFG    1981 new_sp_m014    NA
## 3 Afghanistan AF    AFG    1982 new_sp_m014    NA
## 4 Afghanistan AF    AFG    1983 new_sp_m014    NA
## 5 Afghanistan AF    AFG    1984 new_sp_m014    NA
## 6 Afghanistan AF    AFG    1985 new_sp_m014    NA
```

```
tail(who1)
```

```
## # A tibble: 6 x 6
##   country  iso2  iso3   year codigo     casos
##   <chr>    <chr> <chr> <int> <chr>      <int>
## 1 Zimbabwe ZW    ZWE    2008 newrel_f65    NA
## 2 Zimbabwe ZW    ZWE    2009 newrel_f65    NA
## 3 Zimbabwe ZW    ZWE    2010 newrel_f65    NA
## 4 Zimbabwe ZW    ZWE    2011 newrel_f65    NA
```

```
## 5 Zimbabwe ZW    ZWE    2012 newrel_f65    NA
## 6 Zimbabwe ZW    ZWE    2013 newrel_f65   725
```

PASO 2: Se separa la variable `codigo` de `who1` en dos columnas `nuevo_tipo` y `sexo_edad` utilizando la función `separate()` de la librería `tidyr`

```
who2 <- separate(data = who1, col=codigo, into=c("nuevo_tipo", "sexo_edad"),sep=7)

head(who2)
```

```
## # A tibble: 6 x 7
##   country     iso2  iso3   year nuevo_tipo sexo_edad casos
##   <chr>       <chr> <chr> <int> <chr>      <chr>     <int>
## 1 Afghanistan AF    AFG    1980 new_sp_    m014         NA
## 2 Afghanistan AF    AFG    1981 new_sp_    m014         NA
## 3 Afghanistan AF    AFG    1982 new_sp_    m014         NA
## 4 Afghanistan AF    AFG    1983 new_sp_    m014         NA
## 5 Afghanistan AF    AFG    1984 new_sp_    m014         NA
## 6 Afghanistan AF    AFG    1985 new_sp_    m014         NA
```

```
tail(who2)
```

```
## # A tibble: 6 x 7
##   country  iso2  iso3   year nuevo_tipo sexo_edad casos
##   <chr>    <chr> <chr> <int> <chr>      <chr>     <int>
## 1 Zimbabwe ZW    ZWE    2008 newrel_    f65          NA
## 2 Zimbabwe ZW    ZWE    2009 newrel_    f65          NA
## 3 Zimbabwe ZW    ZWE    2010 newrel_    f65          NA
## 4 Zimbabwe ZW    ZWE    2011 newrel_    f65          NA
## 5 Zimbabwe ZW    ZWE    2012 newrel_    f65          NA
## 6 Zimbabwe ZW    ZWE    2013 newrel_    f65         725
```

PASO 3: nuevamente se separa la variable `sexo_edad` de `who2` en dos columnas: `sexo` y `edad`

```
who3 <- separate(data = who2, col=sexo_edad, into=c("sexo", "edad"),sep=1)

head(who3)
```

```
## # A tibble: 6 x 8
##   country     iso2  iso3   year nuevo_tipo sexo  edad  casos
##   <chr>       <chr> <chr> <int> <chr>      <chr> <chr> <int>
## 1 Afghanistan AF    AFG    1980 new_sp_    m     014      NA
## 2 Afghanistan AF    AFG    1981 new_sp_    m     014      NA
## 3 Afghanistan AF    AFG    1982 new_sp_    m     014      NA
## 4 Afghanistan AF    AFG    1983 new_sp_    m     014      NA
## 5 Afghanistan AF    AFG    1984 new_sp_    m     014      NA
## 6 Afghanistan AF    AFG    1985 new_sp_    m     014      NA
```

```
tail(who3)
```

```
## # A tibble: 6 x 8
##   country  iso2  iso3   year nuevo_tipo sexo  edad  casos
##   <chr>    <chr> <chr> <int> <chr>      <chr> <chr> <int>
## 1 Zimbabwe ZW    ZWE    2008 newrel_    f     65       NA
## 2 Zimbabwe ZW    ZWE    2009 newrel_    f     65       NA
## 3 Zimbabwe ZW    ZWE    2010 newrel_    f     65       NA
## 4 Zimbabwe ZW    ZWE    2011 newrel_    f     65       NA
## 5 Zimbabwe ZW    ZWE    2012 newrel_    f     65       NA
```

```
## 6 Zimbabwe ZW    ZWE    2013 newrel_    f      65      725
```

PASO 4: Se eliminan de `who3` las variables innecesarias y se transforman variables adecuadamente

```
who4<-who3 %>% select(-c(2,3,5)) %>%
  mutate(edad=case_when(
    edad=="014"~"0-14",
    edad=="1524"~"15-24",
    edad=="2534"~"25-34",
    edad=="3544"~"35-44",
    edad=="4554"~"45-54",
    edad=="5564"~"55-64",
    edad=="65"~"mas-65"),
    sexo=if_else(sexo=="m","masculino","femenino")) %>%
  mutate_if(.predicate = is.character,.funs = as_factor)

head(who4)
```

```
## # A tibble: 6 x 5
##   country      year sexo      edad  casos
##   <fct>       <int> <fct>     <fct> <int>
## 1 Afghanistan  1980 masculino 0-14     NA
## 2 Afghanistan  1981 masculino 0-14     NA
## 3 Afghanistan  1982 masculino 0-14     NA
## 4 Afghanistan  1983 masculino 0-14     NA
## 5 Afghanistan  1984 masculino 0-14     NA
## 6 Afghanistan  1985 masculino 0-14     NA
```

```
tail(who4)
```

```
## # A tibble: 6 x 5
##   country   year sexo     edad    casos
##   <fct>    <int> <fct>    <fct>   <int>
## 1 Zimbabwe  2008 femenino mas-65     NA
## 2 Zimbabwe  2009 femenino mas-65     NA
## 3 Zimbabwe  2010 femenino mas-65     NA
## 4 Zimbabwe  2011 femenino mas-65     NA
## 5 Zimbabwe  2012 femenino mas-65     NA
## 6 Zimbabwe  2013 femenino mas-65    725
```

PASO 5: se eliminan de `who4`,para cada país, los años anteriores a 1995 (el análisis va de 1995 a 2013) y se imputan los NA's en la variable casos con 0's (con la función `replace_na()` de `dplyr`)

```
who5<-who4 %>%
  filter(year %in% 1995:2013) %>%
  mutate(casos=replace_na(casos,0))

head(who5)
```

```
## # A tibble: 6 x 5
##   country      year sexo      edad  casos
##   <fct>       <int> <fct>     <fct> <dbl>
## 1 Afghanistan  1995 masculino 0-14      0
## 2 Afghanistan  1996 masculino 0-14      0
## 3 Afghanistan  1997 masculino 0-14      0
## 4 Afghanistan  1998 masculino 0-14     30
## 5 Afghanistan  1999 masculino 0-14      8
```

```
## 6 Afghanistan  2000 masculino 0-14     52
```

```
tail(who5)
```

```
## # A tibble: 6 x 5
##   country   year sexo     edad   casos
##   <fct>    <int> <fct>    <fct>  <dbl>
## 1 Zimbabwe  2008 femenino mas-65     0
## 2 Zimbabwe  2009 femenino mas-65     0
## 3 Zimbabwe  2010 femenino mas-65     0
## 4 Zimbabwe  2011 femenino mas-65     0
## 5 Zimbabwe  2012 femenino mas-65     0
## 6 Zimbabwe  2013 femenino mas-65   725
```

Otros data frames a partir de who5

Los data frames `who6` y `population` (incluido en `DSR`), que ya están en formato tidy, constituyen la base para construir otros data frames para análisis posteriores.

Ejemplo: se puede calcular, a partir de `who5`, el número de casos por paises y años, independientes del sexo y el grupo de edad

```
who6<-who5 %>%
  group_by(country,year) %>%
  summarise(casos=sum(casos))

head(who6)
```

```
## # A tibble: 6 x 3
## # Groups:   country [1]
##   country      year casos
##   <fct>       <int> <dbl>
## 1 Afghanistan  1995     0
## 2 Afghanistan  1996     0
## 3 Afghanistan  1997   128
## 4 Afghanistan  1998  1778
## 5 Afghanistan  1999   745
## 6 Afghanistan  2000  2666
```

```
tail(who6)
```

```
## # A tibble: 6 x 3
## # Groups:   country [1]
##   country   year casos
##   <fct>    <int> <dbl>
## 1 Zimbabwe  2008 36060
## 2 Zimbabwe  2009 41768
## 3 Zimbabwe  2010 42872
## 4 Zimbabwe  2011 36960
## 5 Zimbabwe  2012 34391
## 6 Zimbabwe  2013 32899
```

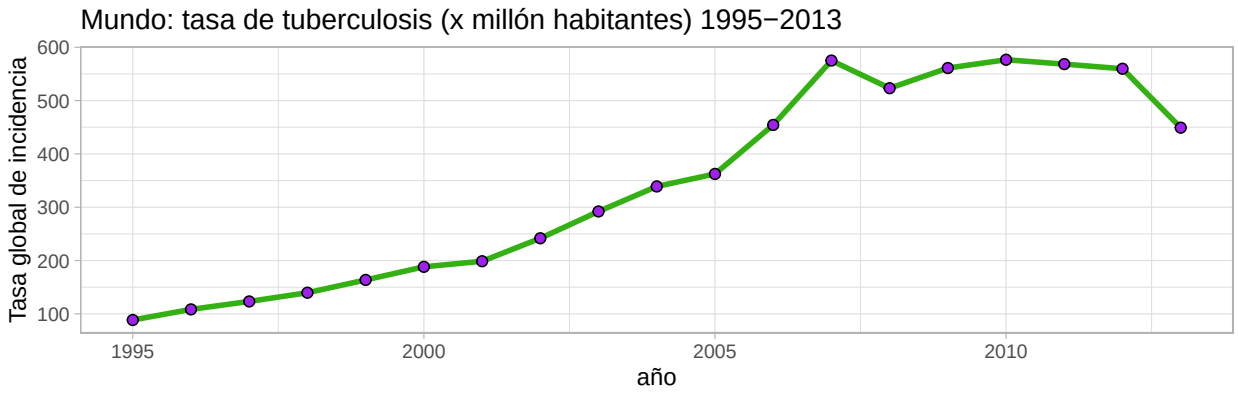
Actividades

Con base en los data frames `who5` y `population` y las distintas funciones utilizadas para wrangling, replique los siguientes resultados donde se explora la distribución de las tasas de incidencia mundiales y a nivel Colombia para 1995-2013:

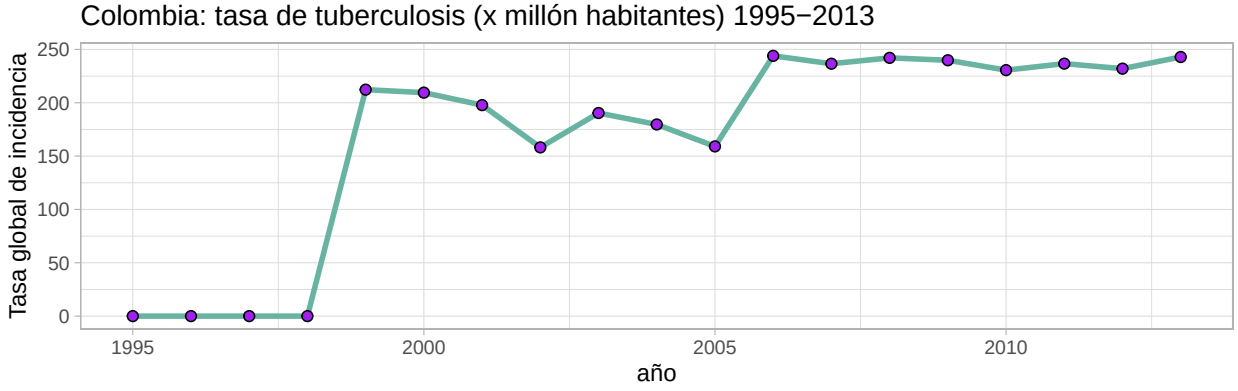
Resultado 1: tabla con incidencias globales de tuberculosis (x millón de habitantes) por país y año (se muestran las diez primeras observaciones)

pais	año	poblacion	casos	tasas
Afghanistan	1995	17586073	0	0.000
Afghanistan	1996	18415307	0	0.000
Afghanistan	1997	19021226	128	6.729
Afghanistan	1998	19496836	1778	91.194
Afghanistan	1999	19987071	745	37.274
Afghanistan	2000	20595360	2666	129.447
Afghanistan	2001	21347782	4639	217.306
Afghanistan	2002	22202806	6509	293.161
Afghanistan	2003	23116142	6528	282.400
Afghanistan	2004	24018682	8245	343.274

Resultado 2: diagrama de puntos con tendencia de las tasas globales de incidencia de tuberculosis (a nivel mundial) entre 1995 y 2013



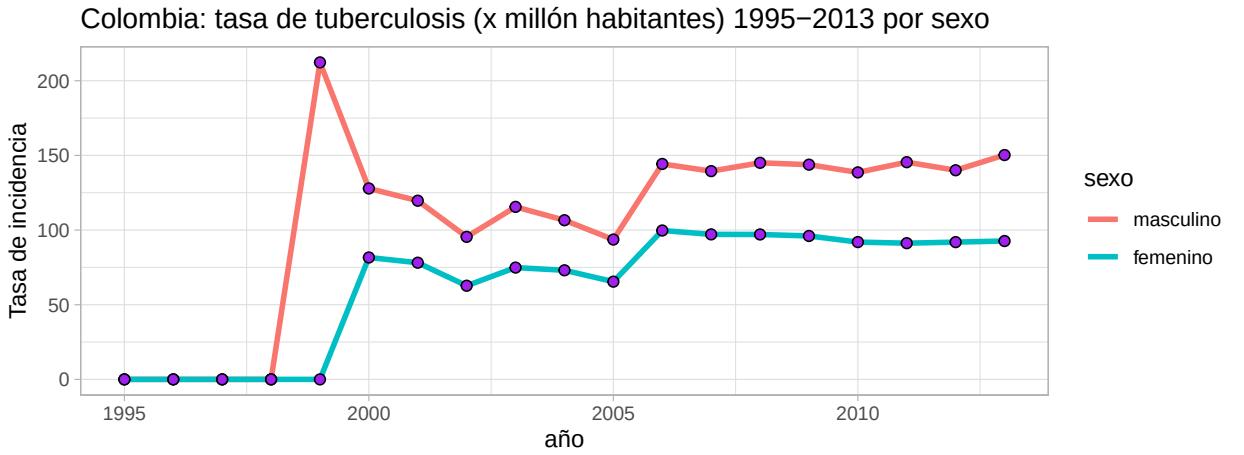
Resultado 3: diagrama de puntos con la evolución de las tasas de incidencias de tuberculosis (x millón de habitantes) para Colombia entre 1995 y 2013



Resultado 4: tabla con incidencias globales de tuberculosis (x millón de habitantes) discriminadas por país, año y sexo (se muestran las diez primeras observaciones)

country	año	sexo	tot_casos	poblacion	tasa_sexo
Afghanistan	1995	masculino	0	17586073	0.000
Afghanistan	1995	femenino	0	17586073	0.000
Afghanistan	1996	masculino	0	18415307	0.000
Afghanistan	1996	femenino	0	18415307	0.000
Afghanistan	1997	masculino	26	19021226	1.367
Afghanistan	1997	femenino	102	19021226	5.362
Afghanistan	1998	masculino	571	19496836	29.287
Afghanistan	1998	femenino	1207	19496836	61.907
Afghanistan	1999	masculino	228	19987071	11.407
Afghanistan	1999	femenino	517	19987071	25.867

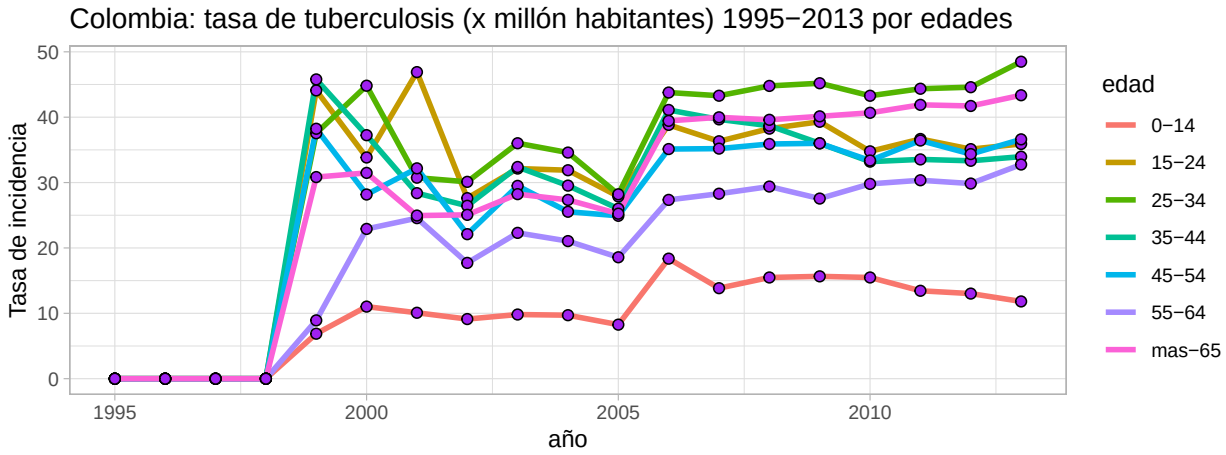
Resultado 5: diagrama de puntos con la evolución de las tasas de incidencias de tuberculosis (x millón de habitantes) para Colombia entre 1995 y 2013 discriminadas por sexo



Resultado 6: tabla con incidencias globales de tuberculosis (x millón de habitantes) discriminadas por país, año y grupos de edades (se muestran las diez primeras observaciones)

country	year	edad	tot_casos	population	tasa_edad
Afghanistan	1995	0-14	0	17586073	0
Afghanistan	1995	15-24	0	17586073	0
Afghanistan	1995	25-34	0	17586073	0
Afghanistan	1995	35-44	0	17586073	0
Afghanistan	1995	45-54	0	17586073	0
Afghanistan	1995	55-64	0	17586073	0
Afghanistan	1995	mas-65	0	17586073	0
Afghanistan	1996	0-14	0	18415307	0
Afghanistan	1996	15-24	0	18415307	0
Afghanistan	1996	25-34	0	18415307	0

Resultado 7: diagrama de puntos con la evolución de las tasas de incidencias de tuberculosis (x millón de habitantes) para Colombia entre 1995 y 2013 discriminadas por grupos de edades



Resultado 8: construya un diagrama de puntos comparando las tasas de incidencias de tuberculosis (x millón de habitantes) para hombres y mujeres a nivel mundial, dentro del periodo 1995-2013

resultado 9: construya una tabla donde se presenten los diez países con las mayores tasas de incidencias de tuberculosis para los hombres, año 2010

resultado 10: construya un diagrama de puntos comparando las tasas de incidencias de tuberculosis (x millón de habitantes) por grupos de edades a nivel mundial, dentro del periodo 1995-2013