R Preparation

Quantitative Reasoning

Preparation for class on 2020-08-31

Comparing Lung Cancer Incidence Rates between Countries

The reading assignment from the textbook is about “Understanding and Comparing Distributions”. Our activity during the next class is about a specific application of the material in this chapter: comparing the incidence of lung cancer between different countries.

Globally, lung cancer is one of the most common types of cancer with estimates of about 1.8 million cases or some 12.9% of all new cases of cancer in 2012 alone. The rates of lung cancer incidence may differ among countries because of factors such as levels of air pollution (a topic we briefly explored using the UESI dataset in the previous lesson) and of course, the percentage of the population that smokes. As an introduction to this activity, please watch Hans Rosling’s comment on lung cancer statistics at https://www.youtube.com/watch?v=QBht72_PA-4.

In this activity we will try to figure out if somebody from, say, Viet Nam has a higher probability of getting lung cancer than somebody from the UK. We will compare the following eight countries: Viet Nam, Singapore, the UK, Ethiopia, Austria, China, Georgia and the Philippines. These data come from the UN (Population Division) and the WHO’s GLOBOCAN database that provides recent (2012) estimates of occurrence and mortality for major types of cancers for 184 countries of the world.

Download the dataset we need for the preparatory activity (link on assignment page).

Start a new project called lung_cancer in your QR folder, shift the CSV file to your project folder and import the spreadsheet as a data frame. As usual, copy and paste the `read.csv()` command into a script.

```
lung_cancer <- read.csv("lung_cancer.csv")
```

The dataset contains

- age-specific populations and
- the number of lung cancer cases in different age groups

for the eight countries mentioned above.

Explore the datasets, as usual, with the functions `dim()`, `head()`, `tail()`, `str()`. Confirm with `unique()` that the `Country` column contains exactly the eight countries mentioned above. Use `unique()` to find out which age groups are in the column `AgeClass`.

Overall incidence rate in the UK

By convention, researchers look at the rate of lung cancer incidence as the number of cancer cases occurring in a specified population during a given year. As you may have noticed in the late Prof. Rosling’s video, this is usually expressed as the number of new cancer cases per 100,000 people in a given year,

$$\text{incidence rate} = \frac{\text{number of cases}}{\text{population}} \cdot 100\,000.$$

We can easily determine the incidence rate in a single country based on our existing knowledge of R. For example, we get the incidence rate in the UK with straightforward subsetting operations.

```
uk <- lung_cancer[lung_cancer$Country == "UK", ]
uk  # Let's see what's inside.
```

```
##    Country AgeClass Population Cases
## 21      UK     0-14   11427668     3
## 22      UK    15-39   20700000   207
## 23      UK    40-44    4519231   235
## 24      UK    45-49    4652439   763
## 25      UK    50-54    4201613  1563
## 26      UK    55-59    3637701  2721
## 27      UK    60-64    3685606  4865
## 28      UK    65-69    3219090  6509
## 29      UK    70-74    2516451  6807
## 30      UK      75+    5013201 16709
```

```
uk_total_cases <- sum(uk$Cases)
uk_total_population <- sum(uk$Population)
uk_incidence_rate <- (uk_total_cases / uk_total_population) * 100000
uk_incidence_rate
```

```
## [1] 63.52068
```

During the activity, we are going to use the UK incidence rate as a metrestick to compare incidence rates between countries. It will become clear during the next class what we mean by “metrestick”.

Overall incidence rate by country

We could in principle repeat the same calculation above for the seven remaining countries. So we would copy the code above and change "UK" to the name of another country seven times. This strategy gets the job done, but is not really elegant. Imagine we would have given you the full dataset containing 184 countries! Good computer code should not contain a lot of almost identical lines. A better alternative is the aggregate() function that returns the total number of cases in each country. Let us first see the function in action before we explain how to use it.

```
aggregate(Cases ~ Country, data = lung_cancer, sum)
```

```
##       Country  Cases
## 1     Austria   4576
## 2       China 652842
## 3    Ethiopia   1533
## 4     Georgia   1129
## 5 Philippines  12074
## 6   Singapore   1974
## 7          UK  40382
## 8    Viet Nam  21865
```

What has happened here? You may remember from video tutorial 10 that the tilde operator (~) in R stands for “as a function of”. So the first argument Cases ~ Country instructs R to view the cases as a function of the country in which they occur. In other words, R splits the Cases column by Country. The second argument tells R that it can find the data in lung_cancer. The third argument says that R should sum up all the cases for each country. The return value of aggregate() is a data frame.

Our goal is to calculate the incidence rate in each country. So we also need the population in each country. We can tell aggregate() to add another column to the output with the cbind() function.

```
aggregate(cbind(Cases, Population) ~ Country, data = lung_cancer, sum)
```

```
##       Country  Cases Population
## 1     Austria   4576    8455000
## 2       China 652842 1355386000
## 3    Ethiopia   1533   92191000
## 4     Georgia   1129    4138000
## 5 Philippines  12074   96017000
## 6   Singapore   1974    5300093
## 7          UK  40382   63573000
## 8    Viet Nam  21865   90332264
```

Don't worry too much about the syntax. Just treat it as a recipe that you can adjust when we encounter a similar challenge. As a mnemonic, the "c" in `cbind()` stands for "column". So we bind two columns together: one for the cases, another for the population. Let us store the result in a variable `total` to indicate that the numbers are the sums over all age groups.

```
total <-
  aggregate(cbind(Cases, Population) ~ Country, data = lung_cancer, sum)
```

Now we only need to append one more column that contains the incidence rates. We have known since video tutorial 06 that we can use the $ operator for this purpose.

```
total$Incidence <- (total$Cases / total$Population) * 100000
total
```

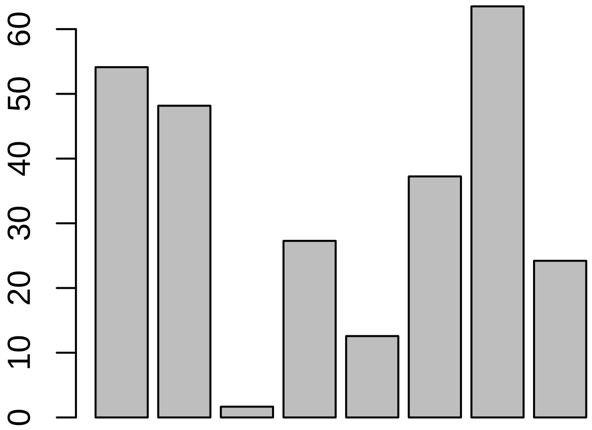
```
##       Country  Cases Population Incidence
## 1     Austria   4576    8455000 54.121821
## 2       China 652842 1355386000 48.166500
## 3    Ethiopia   1533   92191000  1.662852
## 4     Georgia   1129    4138000 27.283712
## 5 Philippines  12074   96017000 12.574857
## 6   Singapore   1974    5300093 37.244629
## 7          UK  40382   63573000 63.520677
## 8    Viet Nam  21865   90332264 24.205084
```

Note that we have replicated the result for the UK from the previous section. But we have also found the incidence rates for all other countries with just a few lines of code.

Bar plot of the incidence rates

In video tutorial 08, we learned how to make a bar chart from a frequency table. In fact, we can also make a bar chart from any numeric vector, for example `total$Incidence`.

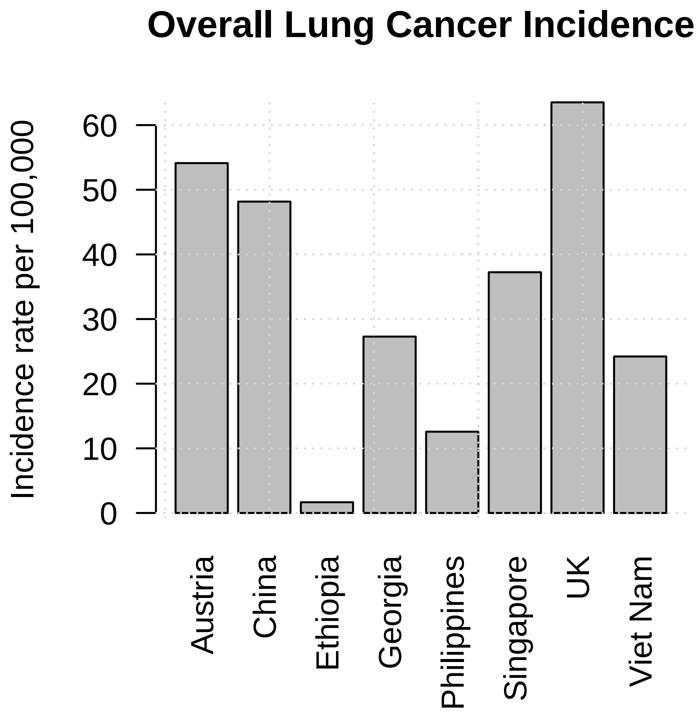
```
barplot(total$Incidence)
```



The plot above does not show which bar corresponds to which country. But by passing a few additional arguments to `barplot()` we can achieve a very presentable result.

```
barplot(total$Incidence,
        names.arg = total$Country,
        main = "Overall Lung Cancer Incidence",
        ylab = "Incidence rate per 100,000",
        las = 2)  # las = 2 rotates the country names by 90 degrees

# Let's add a faint grid on top of the figure. It makes it easier to read the
# numbers from the plot.
grid()
```



Does the overall incidence tell the full story?

Do you find the results shown by the bar chart surprising? What could such a crude, yet quick analysis miss? Think about what differentiates these countries (besides GDP and any other measures of livelihood obviously). How can we go about rectifying this? As you hopefully saw with the Berkeley admissions data, an unsophisticated aggregate approach may lead to a misleading conclusion with ill-advised consequences, especially if you are in a position to prescribe or proscribe policy recommendation to a government or inter-governmental body. So we will take a hopefully more considered and nuanced approach when we revisit this in class.

Practice with `aggregate()`

Consider the spreadsheet `country_info.csv` from an earlier activity (data in same Canvas as this document. Import the data into R.

```
country_info <- read.csv("country_info.csv")
```

As a reminder, here is how the top and bottom of the data frame `country_info` look like.

```
head(country_info)
```

```
##    country continent       pop electr_pct
## 1  Burundi    Africa  11890784       7.59
## 2  Comoros    Africa    869601      77.80
## 3 Djibouti    Africa    988000      51.80
## 4  Eritrea    Africa   3546421      46.70
## 5 Ethiopia    Africa 114963588      42.90
## 6    Kenya    Africa  53771296      56.00
```

```r
tail(country_info)
```

```
##                      country continent       pop electr_pct
## 188               Luxembourg    Europe    625978        100
## 189                   Monaco    Europe     39242        100
## 190              Netherlands    Europe  17134872        100
## 191              Switzerland    Europe   8654622        100
## 192                   Canada  Americas  37742154        100
## 193 United States of America  Americas 331002651        100
```

In our activity, we calculated the population of each continent by subsetting `country_info` five times, namely once for each continent (see the sample solution on Canvas under Files → Week01_Lesson2). This strategy works, but leads to a lot of repeated code.

How can we calculate the population of each continent with `aggregate()`? See the Practice RAT for a solution, but first try it yourself.