Extra-credit #1: Categorical Data

Milica Cudina

2022-10-03

Problem 1. (10 points total)

The data set at hand concerns "pre-existing medical conditions of 92 children involved in a study on the optimal duration of antibiotic use in treatment of tracheitis, which is an upper respiratory infection."

(2 points) First, you should read the data from our csv file "antibiotics.csv" into a data.frame:

```
babies<-read.csv("antibiotics.csv")
```

(2 points) Now, create and print out the table of counts of different conditions present in the above dataset.

```
baby.list<-table(babies)
print(baby.list)
```

```
## condition
##    Cardiovascular  Gastrointestinal Genetic/metabolic Immunocompromised
##                16                 2                 6                 2
##     Neuromuscular       Prematurity       Respiratory             Trauma
##                10                33                13                10
```

Now, you feel that your table is not pretty enough. Try the following command (don't forget to un-comment the line of code once you insert the name of your table):

```
#knitr::kable(name-of-your-table-goes-here)
knitr::kable(baby.list)
```

condition	Freq
Cardiovascular	16
Gastrointestinal	2
Genetic/metabolic	6
Immunocompromised	2
Neuromuscular	10
Prematurity	33
Respiratory	13
Trauma	10

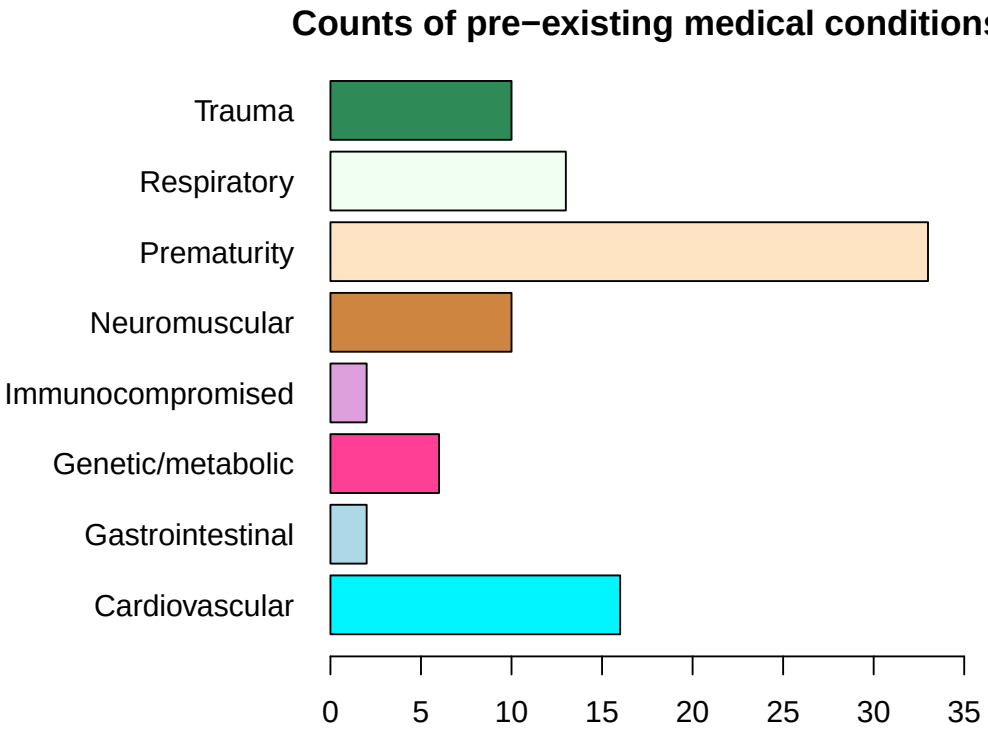
(6 points) Now, you want to plot a bargraph of the data. You do **not** want all of your columns to be gray and you **do** want all the labels of the bars to be visible. To figure out what the available colors are, you can use the command `colors()` in your console. In class, we have used the `barplot` command. To be able to adequately represent the data at hand, you will need to make some modifications to what we did in class. Please, format your barplot so that it satisfies the following properties:

- your chart must have a descriptive *title*,
- the barplot is *horizontal*, i.e., bars stretch from left to right,

- your labels for both axes are readable without tilting your head, i.e., they are readable from left to right,
- your labels for the *conditions* are must be readable in their entirety, i.e., none of the text gets cut off due to inadequate margin sizes.

This barplot customization tutorial is sure to be useful.

```
par(mar=c(4,11,2,4))
barplot(baby.list, col=c("turquoise1","lightblue","violetred1",
                         "plum","peru","bisque", "honeydew", "seagreen"),
        las=1, horiz=TRUE,
        main="Counts of pre-existing medical conditions",
        xlim=c(0,35))
```

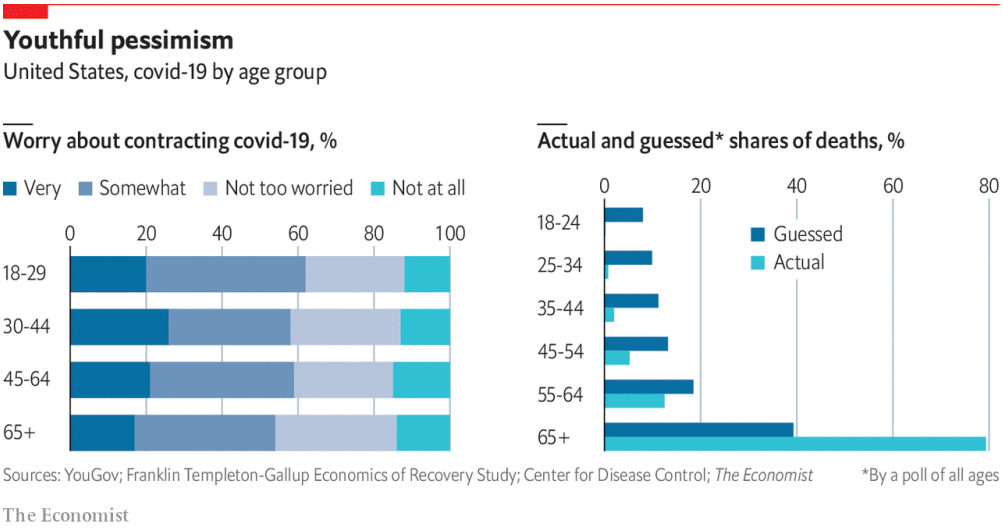


Problem 2. (5 points)

Remember that The University of Texas has a subscription to *The Economist* and that you can enjoy their charts and articles whenever you want.

Consider the following charts which appeared in *The Economist* in late 2020:

```
knitr::include_graphics("bar-graph.png")
```



According to the above charts, which of the following statements is **not** correct? **Why? Justify** your answer by going through all of the statements below and providing supporting evidence of their veracity (based on the provided charts, of course).

a. More than three in five 18- to 29-year-old Americans are "very" or "somewhat" worried about contracting COVID-19.
b. Young people do not fall ill with the virus as often as older people.
c. The findings **suggest** that many people underestimate the age of a typical COVID-19 victim.
d. Young people are **underestimating** the number of COVID-19 victims in their age group.
e. People aged 65 and older account for more than a half of the COVID-19 victims.

Solution: a. is **true** because we can see in the chart on the left-hand side that the combined *very* and *somewhat* measure over the 60% mark (which is exactly what "three in five" would be).

b. could be understood as **true** because the combined light-blue bars in the chart on the right-hand side for all the people under the age of 64 do not exceed the light-blue bar for people 65 and over - and one could say that the number of ill individuals is correlated with the number of deaths. On the other hand, if you want to judge just from the graphs, you could say **false** or **I don't know.** All the answers are accepted, as long as you demonstrate your thought process.

c. is **true** because the dark-blue bars in the chart on the right-hand side are longer than the light-blue bars in the same chart for all but the oldest age groups. Also, the word "suggest" saves the day for this statement (and many statements you would find in the press).

d is, by all the evidence and some surmising, **false**. Looking at the chart on the right-hand side, we see that for all but the oldest age brackets the dark blue *guessed* bar is longer than the light blue *actual* bar representing total covid-19 deaths. While the footnote says the bars represent "all the people's responses", there is no reason for us to think that the thought process of any particular age group is different from the total. I will accept **I don't know** as the answer as well as long as you explained your thought process.

e. is **true** because the light-blue bar in the chart on the right-hand side corresponding to the oldest age group exceeds the 50% mark.