Exam:
Midterm 1 (Takehome)

MATH 150

Due: Feb 12, 2024

Sepehr Akbari

Problem 7

This question refers to the gss data set, available on Moodle. **Use R for all calculations. Include all code used to generate answers.**

(a) Which of the following variables are quantitative and which are categorical? What is the level of measurement of each?

- year (explain in just a sentence or two)

  Categorical — Ordinal: In this case, 'year' is not measuring age, but it refers to 'something' in a calendar, therefore performing as a label, which could be used to categorize data (for example, divorce rate in 2007), hence its a categorical data. Moreover, although 'year' would still make sense unordered, it does have a natural order, no matter what calendar system one uses, therefore it is of an ordinal level.

- marital (no explanation needed)

  Categorical — Nominal: Each observation in the marital status variable contains distinct categories marked by labels, such as 'married' and 'divorced', making it a categorical data type. Although this column could be sorted in some order, but does not have any natural order or ranking which at least most data scientist can agree on, and therefore is at a nominal level.

- age (no explanation needed)

  Quantitative — Ratio: Age, represented as an integer, is considered quantitative data, as it can be mathematically manipulated. Moreover, the age variable has a meaningful zero point (which signifies birth) and ratios (20 years old is twice older as 10 years old). Therefore is at a ratio level.

(b) Find the mean number of hours of television watched per day. Also determine the variance of this variable. Recall that you can force R to ignore missing values by adding the argument na.rm = TRUE to most functions.

Using R:

```
library(readxl)
gss <- read_excel("Documents/LFC/MATH 150/DSs/gss.xlsx")
View(gss)

mean(gss$tvhours, na.rm = TRUE) # 2.981945
var(gss$tvhours, na.rm = TRUE) # 6.707056
```

.
Therefore the mean hours of television watched is about 2.98 hours, with a variance of 6.7.

It is also important to have in mind that this is not the most accurate calculation, as a large part of the sample did not respond to the question, and therefore were ignored in the calculations.

(c) Find the standard deviation of hours of television watched per day. Briefly interpret your answer.

Using R:

```
sd(gss$tvhours, na.rm = TRUE) # 2.589798
```

.
Therefore the standard deviation is about 2.59 hours. This indicates that on average, individual responses for hours of TV watched tend to deviate by approximately 2.59 hours from the mean of 2.98 hours. This suggests that while the mean provides a central tendency around which the responses cluster, the actual number of hours watched by each individual can vary from this average value by roughly 2.59 hours.

As previosly said this calculation is not very reliable and accurate as it does not contain all the sample's repsonses. That said, using standard deviation to measure spread is a better than using variance as we are dealing with a right-skewed histogram with some outliers.

(d) Determine the five number summary and IQR of the ages of respondents to this survey.

Using R:

```
quantile(gss$age) # 18 33 46 59 89
IQR(gss$age) # 26
```

.
Five-Number summary:

0% (min)	25% (Q1)	50% (M)	75% (Q3)	100% (max)
18	33	46	59	89

IQR: $Q3 - Q1 = 59 - 33 = 26$

(e) What is the 90th percentile of ages in this survey? Briefly but clearly explain what this means.

Using R:

```
quantile(gss$age, 0.9) # 72
```

.
A percentile is a statistical measure used to indicate the value below which a given percentage of observations in a dataset fall. In this case for example the 90th percentile is 72, that means that 90% of the observations in the dataset have an age value equal to or less than 72. This can also indicate that 10% of the dataset have an age greater than 72.