Computing and reporting descriptive statistics

Understanding variable types and data types

Jenine Harris
Brown School

Data types for categorical variables

- Arguably, the two data types most commonly used in social science research are **factor** and **numeric.**
- The **factor** data type is used for information that is measured or coded in categories, or **categorical variables**.
- There are two types of categorical variables, **nominal** and **ordinal.**
    - Nominal categorical variables are variables measured in categories that do not have any logical order, like marital status, religion, sex, and race.
    - Ordinal categorical variables have categories with a logical order.
        - One way to remember this is to notice that ordinal and order both start with *o-r-d*, so ordinal variables have an order.
        - For example, income can be measured in categories such as:
            - < 10k
            - 10k - 24,999
            - 25k - 99,999
            - 100k+
        - These categories have a logical order from the lowest income group to the highest income group.

Ordinal variables with Likert scales

- One common ordinal way of measuring things is the use of a **Likert scale**.
- Likert scales have categories go in a logical order from *least to most* or *most to least*.
- For example, measuring agreement with the statement that *R is awesome* could use a Likert scale with the following options:
    - Strongly agree
    - Somewhat agree
    - Neutral
    - Somewhat disagree
    - Strongly disagree
- Often people refer to the number of categories as the *points* in a Likert scale.
- The agreement scale from strongly agree to strongly disagree is a 5-point Likert scale because there are five options along the scale.

Data types for continuous variables

- The **numeric** data type in R is used for continuous variables.
- **Continuous** variables can take *any* value along some continuum, hence **continuous**.
- Just like *o-r-d* is in order and **ordinal**, *c-o-n* is in continuum and **continuous**, which can be a good way to remember this variable type.
- Examples of continuous variables include age, height, weight, distance, blood pressure, temperature, etc.
- The **numeric** data type is also often used for variables that do not technically qualify as continuous but are measured along some sort of a continuum.
    - Age in years would be one example since it falls along a continuum from 0 years old to over 100.
        - However, by specifying age *in years* the values for this variable cannot be any value between 0 and over 100, but instead are limited to the whole numbers in this range.
    - Likewise, the number of cars in a parking lot, or the number of coins in a piggy bank, would be numeric but not truly continuous.
- If the variable is a whole number, the **integer** data type could also be used, but it has more limitations for analysis than the **numeric** data type in R.

Check your understanding

Identify the most appropriate data type for the following variables:

- number of healthy days per month
- marital status
- religious affiliation
- smoking status
- number of alcohol beverages per week

Answer

Identify the most appropriate data type for the following variables:

- number of healthy days per month **numeric** or **integer**
- marital status **factor**
- religious affiliation **factor**
- smoking status **factor**
- number of alcohol beverages per week **numeric** or **integer**