BMR 617

Types of Variable

Types of variable

- Contrast *categorical* and *quantitative* variables
  - This distinction is extremely important in deciding how to explore, analyze, and make inferences from your data
  - Asking the question “Are the explanatory and response variables categorical or quantitative?” goes a long way to determining the correct analysis to perform
- Further classify categorical variables as “nominal” and “ordinal” and quantitative variables as “interval” and “ratio”

Types of Data

- Remember “NOIR” mnemonic
- Nominal
  - Categorical, no ordering: Gender, Race, Genotype
- Ordinal
  - Categorical with an order: Socio-economic status, Pain scale
- Interval
  - Numerical data on a scale (has units) but no meaningful zero
  - Temperature in Celsius or Fahrenheit
    - Zero is arbitrary. “Doubling the temperature” doesn’t make sense
- Ratio
  - Numerical data with scale and zero.
    - Most measurement data is in this category

Determining the type of variable

- To determine the type of variable we are using, we ask the following questions:
  - Is there an ordering for values of the variable?
  - If there is an ordering, is there a scale?
    - i.e. Does an increase in one unit always mean the same thing?
  - If there is an ordering and a scale, does the value zero have a specific meaning?

- Additionally, we ask if the variable is continuous or discrete.
  - Continuous means there's always a value lying strictly between any two distinct values
    - So it must be able to take on fractional values
  - Discrete means it takes on only specific, disjoint, values.

Nominal variables

- Nominal variables are those whose values have no ordering.
  - Just qualitative categories.
  - Cannot be continuous.
  - Examples:
    - Gender
      - Values are "Male", "Female”
    - Race
      - Values are "Black", "White", "Asian", "Native American", etc…

Ordinal values

- Ordinal variables are variables with qualitative categories which have an ordering, but no scale.
  - Example: Economic status
    - Values are typically stated as "Low", "Medium", or "High", which are computed using a number of factors (income, education level, occupation, wealth).
    - These are ordered because there is a natural ordering low → medium → high.
    - They have no scale because the difference between low and medium is not necessarily the same as the difference between medium and high.

Interval Variables

- Interval variables are variables with ordering and scale, but with no meaningful zero
- Examples:
  - Temperature in celsius or fahrenheit
    - There is a scale, because a difference in one degree means the same thing, no matter what the starting temperature is.
    - However, the choice of a zero value is essentially arbitrary.
  - Time/Date

Operations on interval variables

- Computing differences of values of interval variables makes sense.
  - For example, computing a change in temperature (difference between two temperatures) makes sense, since a change of one unit (one degree) makes sense.
  - Computing ratios of values of interval variables does not make sense, because there is no meaningful zero value.
    - Ratios of values are dimensionless
      - Have no units
      - Should be the same no matter what units we start in.
    - 100°C is not double 50°C
    - These values are equal to 212°F and 132°F respectively.

Ratio Variables

- Ratio variables have both scale and a meaningful zero
  - Most measurements you work with will be ratio variables
  - Length, mass, count data (e.g. number of cells), etc

Operations on Ratio Variables

- It makes sense to compute differences and ratios of ratio variables.
  - A blood pressure of 120 is double the blood pressure of 60.
- Note that the difference of values of an interval variable is always a ratio variable
  - For example, elapsed time (essentially the difference between two dates) is a ratio variable

Examples

- For each of the following, determine the type of the variable (Nominal, Ordinal, Interval, Ratio). Also determine whether it is continuous or discrete.

Variable	Type (N/O/I/R)	Continuous/Discrete
Tumor grade		
Heart rate		
# Heart attacks in a patient's lifetime		
Color		
Weight (mass)		
Disease status		
Pain scale		
Age		
Genotype		
C_T values from RT-qPCR		

Examples

- For each of the following, determine the type of the variable (Nominal, Ordinal, Interval, Ratio). Also determine whether it is continuous or discrete.

Variable	Type (N/O/I/R)	Continuous/Discrete
Tumor grade	Ordinal	Discrete
Heart rate	Ratio	Continuous
# Heart attacks in a patient's lifetime	Ratio	Discrete
Color	Nominal	?
Weight (mass)	Ratio	Continuous
Disease status	Nominal	Discrete
Pain scale	Ordinal (maybe interval)	Discrete
Age	Ratio	Continuous
Genotype	Nominal	Discrete
C_T values from RT-qPCR	Interval	Continuous

Categorical and Quantitative Variables

- Remember, nominal and ordinal variables are *categorical*
  - Take on specific values only
- Interval and ratio variables are *quantitative*
  - Measure some value
- When we come to visualize and analyze data, the distinction between categorical and quantitative is the most important
  - Helps to determine appropriate methods of visualization and analysis

Types of variable in R

- R supports the notion of *types of variable*
- Open Rstudio and type the following in the console (don't worry about what these functions mean, yet):

```
x <- rep(c("a", "b", "c"), each=2)
y <- rnorm(6)
```

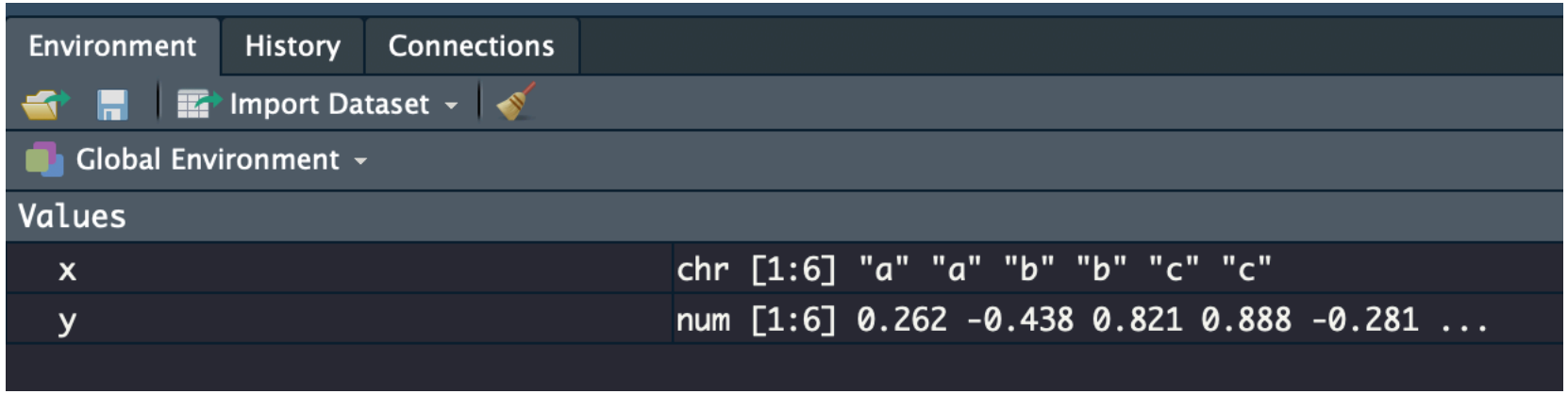
- Look in the "environment" tab. What is the value of `x`? Can you guess what the function `rep` means?
- What is the value of `y`?
  - `rnorm` gives random values from the normal distribution.

Types of variable in R, continued

- In R, we can ask what type a variable is using the `class` function.
- Try the following:

```
class(x)
class(y)
```

- Look at the “Environment” tab. Can you interpret everything that’s displayed there?



Variable types

- Thinking in statistical terms, are x and y *categorical*, or *quantitative*?
- So in R,
  - a *character* variable is …, and
  - a *numeric* variable is …

Variable types

- Thinking in statistical terms, are x and y *categorical*, or *quantitative*?
- So in R,
  - a *character* variable is *categorical*, and
  - a *numeric* variable is *quantitative*

Example

- Imagine a genetic study of obesity, in which we want to determine if the genotype of a particular locus confers obesity
- We could recruit a cohort of patients, measure their BMI, categorize them as obese (yes or no), and determine their genotype at a the locus of interest
  - Let's assume the locus of interest has two possible alleles, C and T

Variable	Categorical (C) or Quantitative (Q)
BMI	
Obese	
Genotype	

Example

- Imagine a genetic study of obesity, in which we want to determine if the genotype of a particular locus confers obesity
- We could recruit a cohort of patients, measure their BMI, categorize them as obese (yes or no), and determine their genotype at a the locus of interest
  - Let's assume the locus of interest has two possible alleles, C and T

Variable	Categorical (C) or Quantitative (Q)
BMI	Q
Obese	C (yes/no)
Genotype	C (CC, CT, TT)

Factors in R

- In R, we could represent our genotype variable with a *character*
- Try:

```
gt <- c("CC", "CC", "CT", "TT", "CT", "CC")
```

- Look in the environment. What is the type of `gt`? Is this what you expect?
- What does the following give?

```
gt[[3]]
```

- What happens if you do

```
gt[[2]] <- "CT"
```

- What about

```
gt[[2]] <- "Meaningless"
```

Factors in R

- When we have a variable that can only take on a fixed set of values, it's useful to force R to only let it have those values
  - This is a common feature of many categorical variables
- In R, a *factor* gives this functionality
- Try the following:

```
gt <- factor(c("CC","CC","CT","TT","CT","CC"))
```

- What does the environment tab display for `gt` now?
- What if you display it in the console (just type `gt` in the console)

Missing values and incorrect values in factors

- Many times when working with data, some values are *missing*
  - Particularly true for clinical and patient/subject-based studies
- R reserves the special value `NA` to represent a missing value
- Now we have gt as a factor, what happens if you do

```
gt[[2]] <- "nonsense"
gt
```

- Using a factor in R allows us to force all values to either be meaningful, or missing.

Summary

- Four types of variable:
  - Nominal
  - Ordinal
  - Interval
  - Ratio
- Nominal and ordinal variables are *categorical*
- Interval and ratio variables are *quantitative*
- The distinction between categorical and quantitative drives the decision as to how to visualize and analyze data

Summary (R)

- In R, we have learned:
  - The `<-` combination of symbols assigns a value to a variable
    - Note you can also use = here: `x=c(2,3,5,7,8)`, but I prefer `<-`
  - You can access individual elements of a variable using `[[ ]]`
    - `x[[3]]` will give the third element of `x`.
    - `x[[2]] <- 5` will change the second element of `x` to `5`.
  - You can find the type of a variable `x` using `class(x)`
  - Use *character* and *factor* types for categorical variables, *numeric* for quantitative variables
    - We'll see other types during the course
  - The special value `NA` represents a missing value