



# STRINGS AND FACTORS

Jeff Goldsmith, PhD  
Department of Biostatistics

# Strings vs Factors

- They both look like character vectors, but:
  - Strings are just strings
  - Factors have an underlying numeric structure with character labels sitting on top
- Factors generally make sense for variables that take on a few meaningful values
  - Sex
  - Race
  - BMI category
- Strings make sense for less structured character values

# Strings vs Factors in R

- Sort of a long story
- Base R, in a variety of ways, has a bias towards factors
  - e.g. character variables are factors when imported using `read.csv`
- This bias stems from historical use
  - R is a statistical language
  - Factors make more sense for classical statistical analysis (e.g. determining sex or race disparities in health outcomes)
- Not so clear there should still be a bias
  - Some folks are upset by base R's preference ...

# Strings vs Factors in R

- Sort of a long story



ors  
ed using read.csv

analysis (e.g.  
comes)

- Not so clear  
– Some

## Package 'hellno'

August 29, 2016

**Type** Package

**Title** Providing 'stringsAsFactors=FALSE' Variants of 'data.frame()' and 'as.data.frame()'

# Common string operations

- There are lots of things you can do with strings
- Some are very common:
  - Concatenating: joining snippets into a long string
  - Shortening, subsetting, or truncating
  - Changing cases
  - Replacing one string segment with another
- The stringr package is the way to go for the majority of your string needs



# Regular expressions

- String operations are “easy” when you know exactly what you’re looking for
- When you know a general pattern but not an exact match, you need to use **regular expressions**
  - Instead of looking for the letter “a” you might look for any string that starts with a lower-case vowel
- Regular expressions take some getting used to

# Factors

- Controlling factors is critical in several situations
  - Defining reference group in models
  - Ordering variables in output (e.g. tables or plots)
  - Introducing new factor levels
- Common factor operations include
  - Converting character variables to factors
  - Releveling by hand
  - Releveling by count
  - Releveling by a second variable
  - Renaming levels
  - Dropping unused levels
- The forcats package is the way to go for the majority of your factor needs
  - (forcats = “for cats”; also an anagram of “factors”)



# Factors

- Controlling factors is critical in several situations
  - Defining reference group in models
  - Ordering variables in output (e.g. tables or plots)
  - Introducing new factor levels

- Common factors
  - Convert
  - Relevel
  - Relevel
  - Relevel
  - Relevel
  - Renaming
  - Dropping



- The forcats package is the way to go for the majority of your factor needs
  - (forcats = “for cats”; also an anagram of “factors”)