**BIOSTATSTICS**

**Why tidy data?** ***The tidy data concept:*** \* provides a standardized layout/organization for data values.

***Standardized aids:*** \* data exploration and analysis \* data sharing \* the develoopment of data cleaning and analysis tools

**Core principles of tidy data** ***The language of datasets:*** \* datasets consist of *values* (usually numbers or strings) \* every *value* belongs to a *variable* and an *observation*

***Structure of a tidy dataset:*** \* *variables* are arranged in *columns* \* *observations* are arranged in *rows* \* each type of observational unit forms a table

**Common causes of messiness** \* column headers are values, not variable names \* multiple variables are stored in one column \* variables are stored in both rows and columns \* multiple types of experimental unit stored in the same table \* on type of experimental unit stored in multiple tables

***Single document type is stored in multiple tables:*** Data values about a single type of observational unit mau be spread out over multiple tables or files. These tables and files are often split up by another variable (e.g., each table represents a singles year, person, or location). As long as the format for individual records is consistent, this is an easy problem to fix by merging tables.

**Tools to tidy and manipulate data** ***The grammer of data cleaning*** Primary *tidyr* package verbs: \* gather(): gathers multiple columns into key-value pair \* spread(): spreads 2 columns (key-value pair) in to multiple columns

**Tools to tidy and manipulate data** ***The grammmer of manipulating data*** Primary *dplyr* package verbs: \* select(): focus on a subset of variables (columns) This function has the inherent ability to arrange columns in any order I please. \* filter(): focus on a subset of rows \* mutate():add new columns \* summarize(): produce summary statistics of variables This function is most powerful when applied to **grouped data**. The main idea behind grouping data is that I want to break up my dataset into groups of rows based on the values of one or more variables. The *group\_by()* function is responsible for doing this. \* arrange(): re-order rows