Data Basics

Ryan Miller

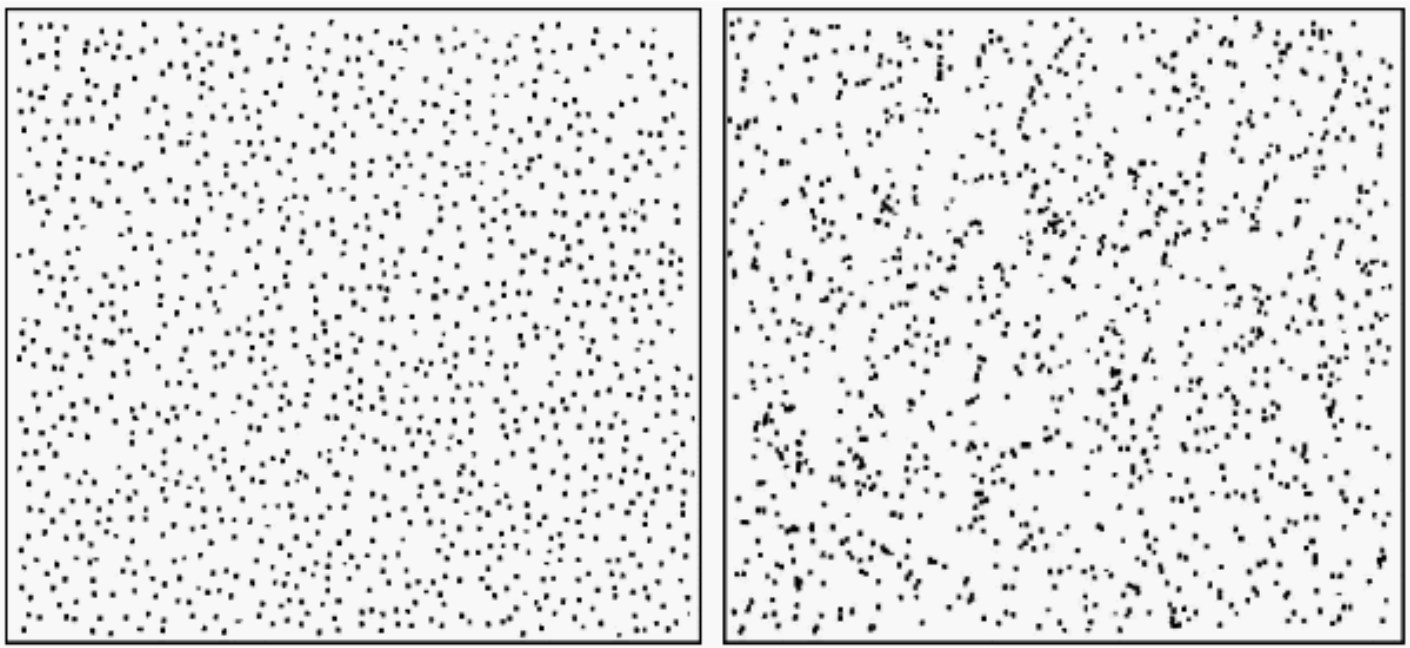
Grinnell College
Statistics

Overview

1. What is "statistics" as a discipline?
2. Data formats
   - Cases vs. variables
   - Tabular or "tidy" data
   - Longitudinal, unstructured, and other types of data
3. Types of variables
   - Categorical vs. quantitative
   - Grey areas

Statistics

One panel displays *randomly positioned* dots, the other shows dots whose positions reflect *meaningful patterns* (ie: biological/physical truths, etc.)



Which panel is which?

Statistics

A fundamental goal of "statistics" is to identify and understand meaningful patterns that exist in the *presence of uncertainty*. To do this we need:

1. Ways of expressing or describing patterns (descriptive statistics, visualizations, and models)
2. Contextual understanding (how were the data collected, what types of patterns are practically meaningful)
3. Methods to judge the role uncertainty in what we observed (is the pattern real, or could it be explained by chance?)

Data Basics

- We'll use *data* to find patterns
  - **Data** is defined as "a collection of discrete or continuous values that convey information"
  - This is a very broad definition, we'll largely restrict our attention to "tidy data" or "tabular data"
- A "tidy" data set is organized such that each row represents an *observation/case* and each column represents a *variable*
  - An **observation** or **case** is defined to be a single unit of analysis (ie: person, subject, etc.)
  - A **variable** is any characteristic or attribute that is recorded for each case

Tidy Data

Below is an example of "tidy data" obtained from the Washington Post's database of police shootings:

name	date	age	race	armed	city	state	body_camera
Tim Elliot	1/2/2015	53	A	armed	Shelton	WA	FALSE
Lewis Lee Lembke	1/2/2015	47	W	armed	Aloha	OR	FALSE
John Paul Quintero	1/3/2015	23	H	unarmed	Wichita	KS	FALSE
Matthew Hoffman	1/4/2015	32	W	armed	San Francisco	CA	FALSE
Michael Rodriguez	1/4/2015	39	H	armed	Evans	CO	FALSE
Kenneth Joe Brown	1/4/2015	18	W	armed	Guthrie	OK	FALSE
Kenneth Arnold Buck	1/5/2015	22	H	armed	Chandler	AZ	FALSE
Brock Nichols	1/6/2015	35	W	armed	Assaria	KS	FALSE
Autumn Steele	1/6/2015	34	W	unarmed	Burlington	IA	TRUE
Leslie Sapp III	1/6/2015	47	B	armed	Knoxville	PA	FALSE
Patrick Wetter	1/6/2015	25	W	armed	Stockton	CA	FALSE
Ron Sneed	1/7/2015	31	B	armed	Freeport	TX	FALSE

In this data set, the cases are individual instances of police-involved shootings, and the variables describe characteristics of these incidents.

Types of Variables

There are different types of variables, and the statistical methods we use will differ by variable type.

- **Categorical Variables** divide the cases into *groups*
  - **Binary** - two mutually exclusive categories
  - **Nominal** - many groups with no natural ordering
  - **Ordinal** - groups with a natural order
- **Quantitative Variables** record a *numeric* value for each case
  - **Discrete** - countable (ie: integers)
  - **Continuous** - uncountable (ie: real numbers)

For which of these types could you calculate an average? Are there any where you could calculate a median but *not* an average?

Grey Areas

There are plenty of grey areas regarding how we decide to handle different variables:

- When analyzing data on the students in our class, we might treat "graduation year" as categorical despite it being a discrete quantitative variable
- Similarly, we might record encode Likert response (strongly disagree, disagree, . . .) as numeric values so that we can calculate averages, correlations, etc.

Going forward you should always be aware of variable types as they will guide how a variable can be summarized, graphed, and analyzed, but you should also recognize the room for flexibility.