2.1 Populations

Contents

> "Statistics is the branch of the scientific method which deals with the data obtained by counting or measuring the properties of populations of natural phenomena. In this definition 'natural phenomena' includes all the happenings of the external world, whether human or not."
>
> – Professor Maurice Kendall

2.1.1 Populations

Here we aim to describe a population using attributes. We do this in a completely non-stochastic manner that relies on descriptive statistics, but that is nevertheless computational and mathematical.

- A **population** is a finite (though possibly huge) set $\mathcal{P}$ of elements.
  - Elements of a population are called **units** $u \in \mathcal{P}$
  - **Variates** are functions $x(u)$, $y(u)$, etc. on the individual units $u \in \mathcal{P}$. For simplicity we will more often use the notation x_u, y_u, etc. when referring to the realized values of these variates for the unit $u = 1, \ldots, N$.
- We will define and explore interesting **population attributes**, denoted generally as $a(\mathcal{P})$. In particular, we will:
  - consider how to calculate them
  - evaluate some of their (non-sampling) properties (e.g. interpretation of the characteristic being captured, sensitivity to outlying points, etc.)

2.1.2 Examples

- The following examples and corresponding datasets are intended to firmly ground the notion of a population (as opposed to a sample).
  - They have been chosen to represent a variety of applications, and of course be rich enough to illustrate concepts relevant to this course.
  - The datasets are populations in the sense that they are finite, complete, and contain everything you might want to learn about them.

Agricultural Census (USA)

- US Census of Agriculture: decline in farms from 2007 to 2012
  - This is an older dataset taken from the book *Sampling: Design and Analysis* by Sharon Lohr (Lohr 2009). The dataset represents a population of $N = 3,078$ units, where each unit is a county (or county equivalent) as defined by the US Census of Agriculture.
  - The following is a list of variates included in the dataset:

Variate	Value
`county`	County name
`state`	State abbreviation
`acres92`	Number of acres devoted to farms in 1992
`acres87`	Number of acres devoted to farms in 1987
`acres82`	Number of acres devoted to farms in 1982
`farms92`	Number of farms in 1992
`farms87`	Number of farms in 1987
`farms82`	Number of farms in 1982
`largef92`	Number of farms, with 100 acres or more, in 1992
`largef87`	Number of farms, with 100 acres or more, in 1987
`largef82`	Number of farms, with 100 acres or more, in 1982
`smallf92`	Number of farms, with 9 acres or less, in 1992
`smallf87`	Number of farms, with 9 acres or less, in 1987
`smallf82`	Number of farms, with 9 acres or less, in 1982
`region`	S=South, W=west, NC=north central, and NE=northeast

Facebook Posts

Moro, Rita, and Vala (2016) report on a study conducted by a cosmetics company who was interested in evaluating the effectiveness of various posts on their Facebook page. This dataset includes information about $N = 500$ posts. Quoting their paper:

> [...] we needed to collect a representative data set of published posts. All the posts published between the 1st of January and the 31th of December of 2014 in the Facebook's page of a worldwide

renowned cosmetic brand were included. As a result, the data set contained a total of 790 posts published. It should be noted that Facebook is the most used social network with an average of 1.28 billion monthly active users in 2014, followed by Youtube with 1 billion and Google+ with 540 million (Insights, 2014)." - (Moro, Rita, and Vala 2016)

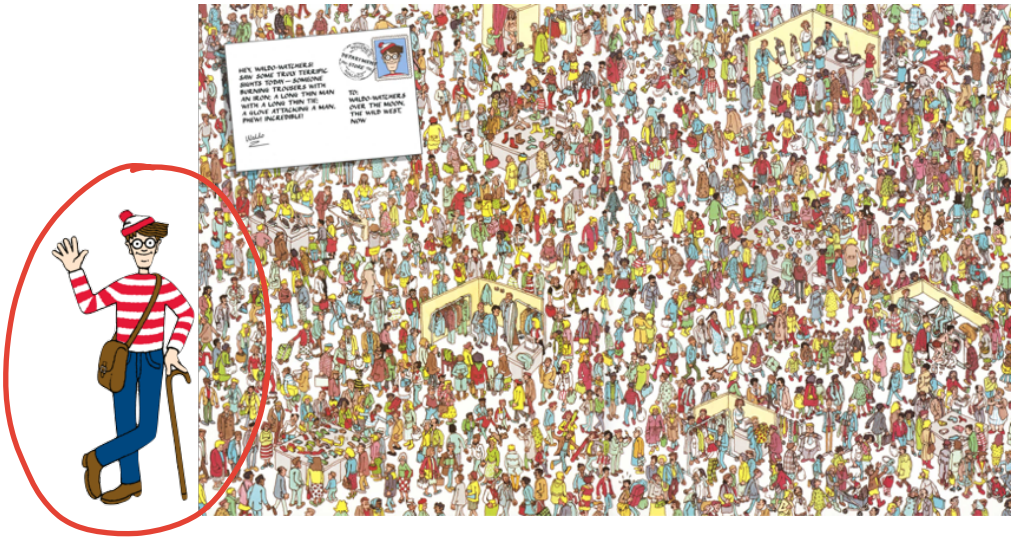
The data were downloaded from the University of California (Irvine) "Machine Learning Repository". Due to privacy issues, the dataset available in the repository contains only 500 of the 790 posts and a subset of the variates analyzed in Moro, Rita, and Vala (2016). The dataset we have access to includes the following 13 variates for each of the 500 posts:

- Facebook Data

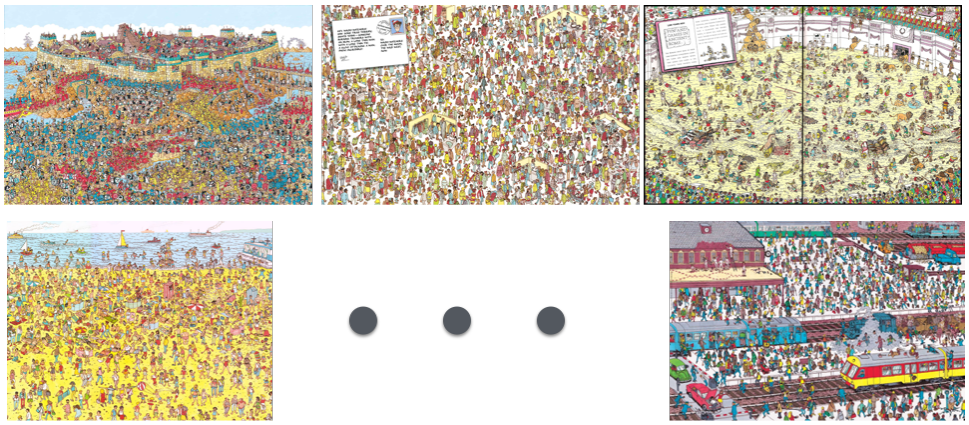
Variate	Value
`share`	the total (lifetime) number of times the post was shared
`like`	the total (lifetime) number of times the post was "liked"
`comment`	the total (lifetime) number of comments attached to the post
`All.interactions`	the sum of `share`, `like`, and `comment`
`Page.likes`	the number of "likes" for the facebook page at the original time of the posting
`Impressions`	the total (lifetime) number of times the post has been displayed, whether the post is clicked or not. The same post may be seen by a Facebook user several times (e.g. via a page update in their News Feed once, whenever a friend shares it, etc.).
`Impressions.when.page.like`	the total (lifetime) number of times the post has been displayed to someone who has "liked" the page
`Post.Hour`	the hour of the day at the original time of the posting (0-23)
`Post.Weekday`	the day of the week at the original time of the posting (1-7, where 1 represents Sunday)
`Post.Month`	the month of the year at the original time of the posting (1-12)
`Category`	the category of the post (as determined by two separate human reviewers according to the campaign associated with the post), one of `Action` (special offers and contests), `Product` (direct advertisement, explicit brand content), or `Inspiration` (non-explicit brand related content)
`Type`	the type of content of the post: either `Link`, `Photo`, `Status`, or `Video`
`Paid`	1 if the company paid Facebook to advertise the post, 0 otherwise

- **Note**: Attributes of interest might include average `Impressions` depending on `Paid` or not. Also, for variates like `Impressions`, `Page.likes`, etc. transforming the data (by square root or logarithms, for example) may yield more interesting values.

Where's Waldo?



- The character Waldo always wore the same shirt, hat, and pants and would appear somewhere in a picture spread across two pages of a book. The objective is to find Waldo in the picture.
- Where's Waldo Population
  - A small population ($N = 68$) is defined by the entire collection of "Where's Waldo?" visual search puzzles taken from an internationally popular children's book series which appeared from 1987 to 2009.



- Where's Waldo Data
  - The population is the set of all two page spreads.
  - An individual unit is any one of the two page spreads.
  - The variates are:

Variate	Value
`Book`	Book number (1 - 7) in which the picture appears
`Page`	Page number of book
`X`	Waldo's Horizontal location measured (in inches?)
`Y`	Waldo's Vertical location measured (in inches?)

- **Note** Possible attributes of interest include density of `X` values; the density of `Y` values; the density of the (`X`, `Y`) pairs; the nature of the relationship between `X` and `Y`; the relationship between any of the aforementioned and page number.

- **Note** The measurements of `X` and `Y` are in error for at least one point. (Check sources to find it.)

The Titanic

From the `help(Titanic)` description in `R`:

> "The sinking of the Titanic is a famous event, and new books are still being published about it. Many well-known facts—from the proportions of first-class passengers to the 'women and children first' policy, and the fact that that policy was not entirely successful in saving the women and children in the third class—are reflected in the survival rates for various classes of passenger.
>
> These data were originally collected by the British Board of Trade in their investigation of the sinking. Note that there is not complete agreement among primary sources as to the exact numbers on board, rescued, or lost."

- The population is the set of all people on board the Titanic's maiden (and only) voyage in 1912. The variates are:

Variate	Value
`Class`	`1st`, `2nd`, `3rd`, or `Crew`
`Sex`	`Male`, `Female`
`Age`	`Child`, `Adult`
`Survived`	`No`, `Yes`

Great White Shark Encounters

Data on known great white shark encounters with humans has been gleaned by Professor P-J Bergeron of the

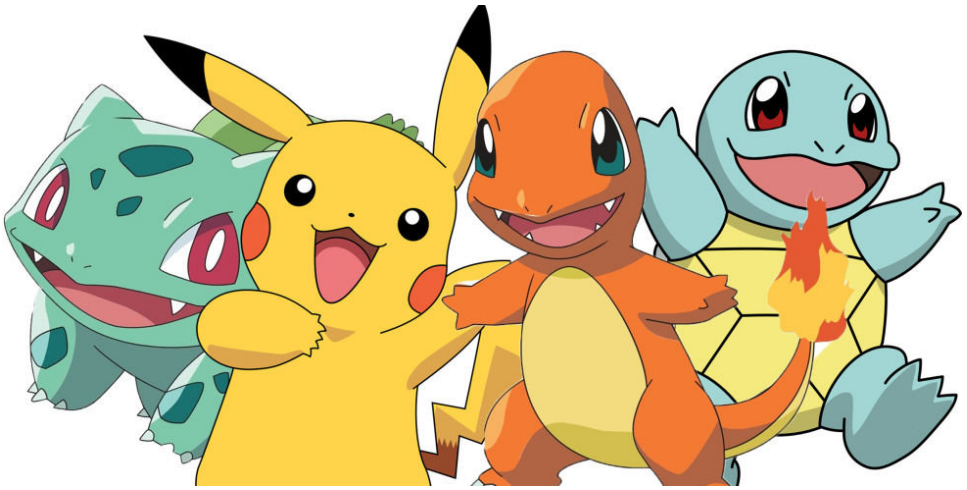
University of Ottawa from a variety of tables which appeared on the now defunct site http://sharkattackinfo.com/shark_attack_news_sas.html.

The population of interest to us is the set of $N = 65$ great white shark encounters in which a person was recorded to have been bitten by a shark.

The variates in the Great White Shark Data are:

Variate	Value
`Year`	the year in which the encounter occurred
`Sex`	sex of the victim (`M` = male, `F` = female)
`Age`	age of the victim in years
`Time`	time of the encounter (`AM` or `PM`)
`Australia`	1 if encounter was in Australian waters, 0 if not
`USA`	1 if encounter was in USA waters, 0 if not
`Surfing`	1 if the victim was surfing at the time of the encounter, 0 otherwise (N.B. other unrecorded activities might be "free diving", "fishing", "pearl diving", etc.)
`Scuba`	1 if the victim was scuba diving at the time of the encounter, 0 othewise (N.B. other unrecorded activities might be "free diving", "fishing", "pearl diving", etc.)
`Fatality`	1 if the victim died after being attacked (though not necessarily directly because of the attack), 0 if they survived
`Injury`	1 if the victim was injured by the encounter, 0 if not
`Length`	the recorded length in inches of the shark thought to have encountered the victim

Pokémon



The Pokémon media franchise is composed of card games, console video games, computer games, mobile games, television shows and feature length films all based on *Pokémon* (a portmanteau of *pocket monsters*). As a Pokémon trainer, your goal is to collect, train, and battle your Pokémon.

The Pokémon population we will deal with contains $N = 801$ Pokémon each with the following variates:

Variate	Value
`pokedex_number`	1 to 801
`name`	Character string containing their English name

Variate	Value
`generation`	1 to 8
`base_happiness`	A numeric happiness score
`hp`	The number of hit points the Pokémon has
`attack`	A measure of the Pokémon's attack capabilities
`defense`	A measure of the Pokémon's defensive capabilities
`speed`	A measure of the Pokémon's speed
`height`	The Pokémon's height in metres
`weight`	The Pokémon's weight in kilograms
`type`	At most two of {grass, fire, water, bug, normal, poison, electric, ground, fairy, fighting, psychic, rock, ghost, ice, dragon, dark, steel, flying}
`is_legendary`	1 if the Pokémon is classified as *legendary*, 0 otherwise