Populations

Populations

> "Statistics is the branch of 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

Populations

The approach here is to remain completely non-stochastic, more traditionally a descriptive statistics approach, but nevertheless computational and mathematical.

- A **population** is a finite (though possibly huge) set $\mathcal{P}$ of elements.
  - Elements of a population are called **units** $u \in \mathcal{P}$, and
  - **variates** are functions $x(u)$, $y(u)$, etc., on individual units $u \in \mathcal{P}$, or more simply as 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 $a(\mathcal{P})$, we consider
  - how they can be calculated, e.g. mean, variance, median
  - and some of their (non-sampling) characteristics (e.g. interpretation of feature being captured, sensitivity to outlying points, etc)

Real datasets

- The datasets given are used to firmly ground the idea of a population.
  - They have been choosen to show variety of applications, and of course be rich enough to illustrate the concepts.
  - The Datasets are obvious as populations in the sense of being finite, complete, and containing all that you want to learn about.

Agricultural census (USA)

- US Census of Agriculture: decline in farms from 2007 to 2012
  - This is an older dataset are taken from the book Sampling: design and analysis by Sharon Lohr (Lohr 2009). The dataset has N=3,078 where each unit is a county (or county equivalent) as defined by the US Census of Agriculture.

Agricultural census (USA) Data

Variates included are

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

not numerical

Variate	Value
`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

This is a subset of data, N=500, collected by (Moro, Rita, and Vala 2016). This study was conducted for a cosmetics company who had a Facebook page and wanted to see the effectiveness of their various postings on that page. 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". The data set found on that site contains only 500 of the 790 posts and a subset of the variates analysed in (Moro, Rita, and Vala 2016). The data uploaded to the course website is a further reduction of the 19 variates available to only 13.

Facebook Data

Variate	Value
`share`	the total (lifetime) number of times the post was shared
`like`	the total (lifetime) number of times the post "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) beginning with Sunday
`Post.Month`	the month of the year at the original time of the posting (1-12)

Variate	Value
`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, one of `Link`, `Photo`, `Status`, or `Video`
`Paid`	1 if the company paid Facebook for advertising, 0 otherwise

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

Where's Waldo

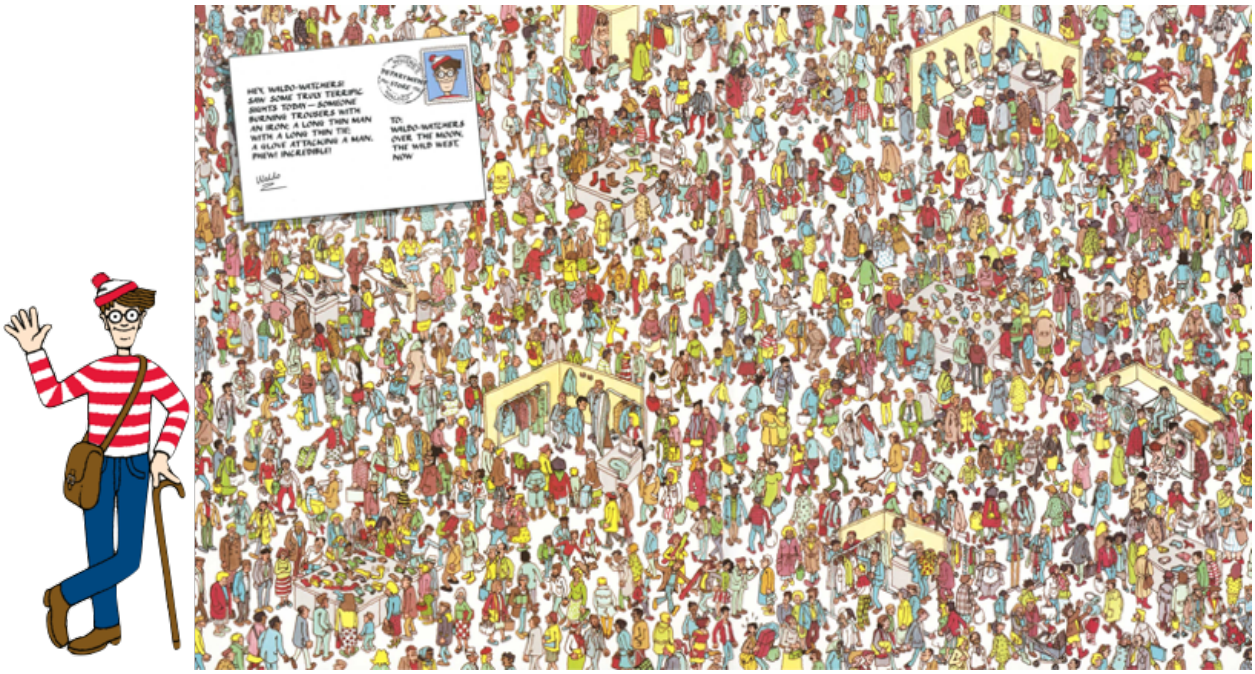

Figure 1: Waldo appears somewhere in the store.

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 $N = 68$ population is defined to 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.

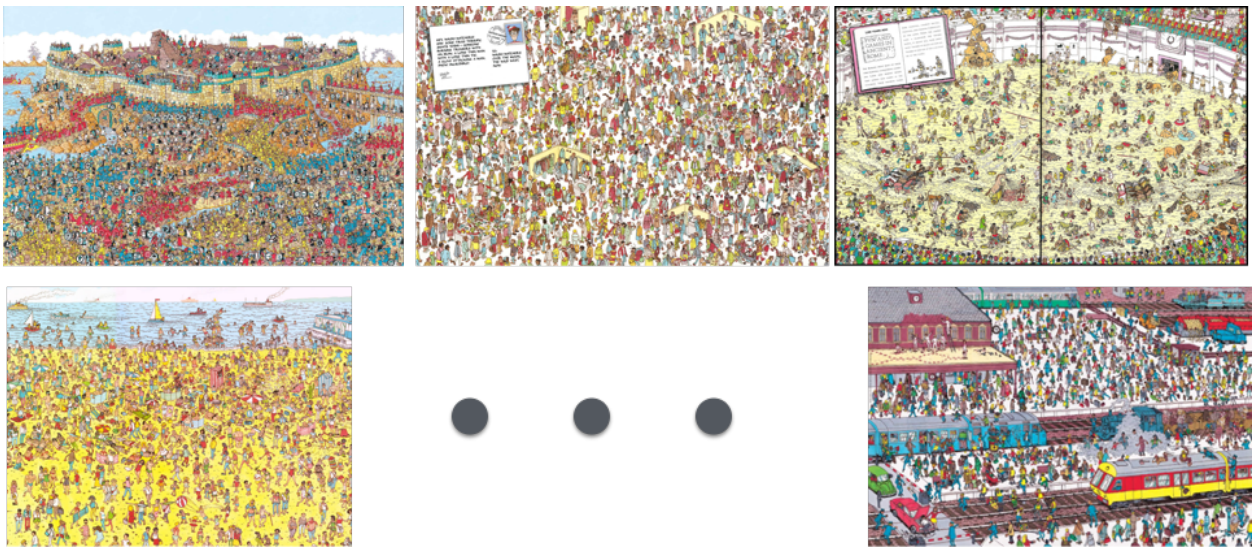

Figure 2:

- An individual unit is any one of the two page spreads.
- The Variates are

Variate	Value
`Book`	Book number (1 - 7) in which 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, of `Y` values, of the pair (`X`, `Y`), of a straight line relationship (or smooth) of `Y` on `X`. Relation between either of `X` or `Y` and even or odd page number?
- **Note** The measurements of `X` and of `Y` is in error, for at least one point. (Check sources to find it.)

The Titanic

learn things as population, not as sample
Not to project into future
cannot simulate

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 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

Variate	Value
`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 could have been "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

Gadsby

- This is a novel, available online, containing approximately 50,000 words in 43 chapters.
  - The population could be the collection of *unique* words in the novel itself ($N < 50,000$), or perhaps the collection of word positions in the book.
  - Variates would include the number of letters of a word, the number of each vowels or consonants in each word, the number of times that word appears, the chapters of the novel in which that word appears, the word at that position, etc.
  - Population attributes are up to your imagination.
- **Note** Perhaps the **most amazing feature of this data** is that nowhere in the text of the novel does the letter `e` appear! This can be for the students to discover, or to motivate counting the frequency of other vowels in the tex

Baseball Data

- Baseball Data 1
  - Major League Baseball Data from the 1986 and 1987 seasons.

```
library(ISLR)
data(Hitters)
head(Hitters)
```

```
##                   AtBat Hits HmRun Runs RBI Walks Years CAtBat CHits
## -Andy Allanson      293   66     1   30  29    14     1    293    66
## -Alan Ashby         315   81     7   24  38    39    14   3449   835
## -Alvin Davis        479  130    18   66  72    76     3   1624   457
## -Andre Dawson       496  141    20   65  78    37    11   5628  1575
## -Andres Galarraga   321   87    10   39  42    30     2    396   101
## -Alfredo Griffin    594  169     4   74  51    35    11   4408  1133
##                   CHmRun CRuns CRBI CWalks League Division PutOuts Assists
## -Andy Allanson         1    30   29     14      A        E     446      33
## -Alan Ashby           69   321  414    375      N        W     632      43
## -Alvin Davis          63   224  266    263      A        W     880      82
## -Andre Dawson        225   828  838    354      N        E     200      11
## -Andres Galarraga     12    48   46     33      N        E     805      40
```

Figure 3: Friendly great white shark

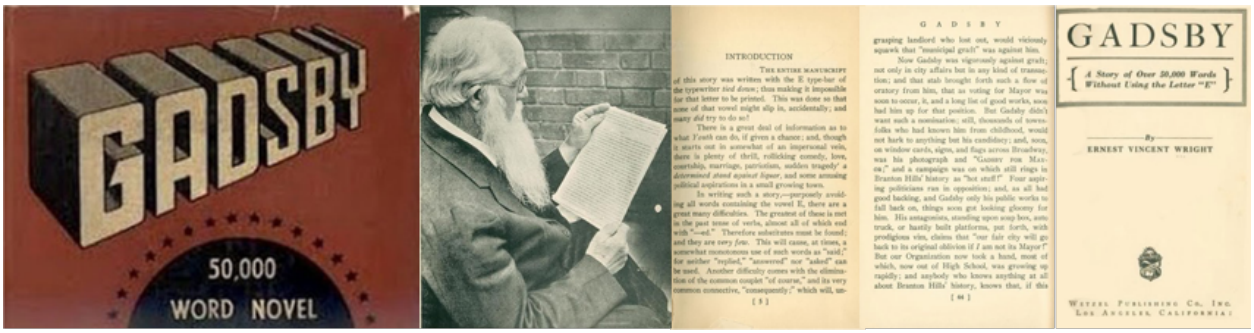

Figure 4: Gadsby: A novel of more than 50,000 words.

```
## -Alfredo Griffin      19   501  336    194      A        W     282     421
##                   Errors Salary NewLeague
## -Andy Allanson        20     NA          A
## -Alan Ashby           10  475.0          N
## -Alvin Davis          14  480.0          A
## -Andre Dawson          3  500.0          N
## -Andres Galarraga      4   91.5          N
## -Alfredo Griffin      25  750.0          A
```

if use to predict => sample
if learn about properties => population

- Baseball Data 2
  - This data frame contains batting statistics for a subset of players collected from http://www.baseball-databank.org/. There are a total of 21,699 records, covering 1,228 players from 1871 to 2007. Only players with more 15 seasons of play are included.

```
library(plyr)
data(baseball)
```

Fire Emblem Heroes

Fire Emblem Heroes is a free-to-play tactical role-playing game developed by Intelligent Systems and Nintendo for iOS and Android devices.

- The Population is 168 characters with varying type and movement.
  - Some variates are below but the game mechanics allow for the development of new variates.

Figure 5:

```
feh = read.csv("feh.csv")
head(feh)
```

```
##       Name        Type      Move HP ATK SPD DEF RES Total
## 1     Abel Blue Lance   Cavalry 39  33  32  25  25   154
## 2 Alfonse  Red Sword Infantry 43  35  25  32  22   157
## 3      Alm  Red Sword Infantry 45  33  30  28  22   158
## 4  Amelia  Green Axe  Armored 47  34  34  35  23   173
## 5     Anna  Green Axe Infantry 41  29  38  22  28   158
## 6  Arthur  Green Axe Infantry 43  32  29  30  24   158
```