

# Data

Professor Jarad Niemi

STAT 226 - Iowa State University

August 22, 2018

## Important terminology/concepts:

- Data
  - Individuals and variables
  - Categorical vs numerical variables
  - Nominal vs ordinal variables
  - Random variables vs observations
- Descriptive vs inferential statistics
  - Population vs sample
  - Parameters vs statistics
- Time series - out of place

# Individuals and Variables

## Definition

**Individuals** are subjects/objects of the population of interest; can be people but also business firms, common stocks or any other object that we want to study.

## Definition

A **variable** is any characteristic of an individual that we are interested in. A variable typically will take on different values for different individuals.

## 2. Dataset basics - Data types

Aa Aa 

Students in a business statistics class developed a pricing model for diamond stones.

The top and bottom portions of the data set that the students collected are reproduced in the following table; dots indicate that the intervening rows in the data set are not displayed. [Source: S. Singfat Chu, "Pricing the C's of diamond stones," *Journal of Statistics Education* 9(2) (2001).]

Diamond ID	Price (Singapore dollars)	Weight (Carats)	Color	Clarity	Certification Body
1	8,873	1.01	H	VS2	1
2	3,635	0.52	E	VS1	1
3	11,696	1	F	VVS1	3
4	8,095	1	I	VS1	3
5	3,501	0.5	F	VVS2	1
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
304	4,401	0.63	G	VVS2	1
305	2,942	0.46	E	VVS2	1
306	3,706	0.55	F	VVS2	2
307	1,555	0.31	E	VS1	1
308	1,098	0.33	I	VS2	1

Note that color purity is a desirable characteristic of a diamond. A grade of D indicates top color purity, a diamond graded E has less color purity than a diamond graded D, a diamond graded F has less color purity than a diamond graded E, and so on. Clarity is also a desirable characteristic. The top clarity rating is IF (internally flawless); other clarity ratings, in descending order, are VVS1, VVS2, VS1, and VS2. (VVS is the notation for "very, very slightly imperfect," and VS is shorthand for "very slightly imperfect.") Certification Body has three different values, which are coded as 1 = Gemological Institute of America, 2 = International Gemological Institute, and 3 = HRD Antwerp.

Category	Region	Subcategory	Revenue	Profit	Cost
Books		<a href="#">Art &amp; Architecture</a>	\$480,173	\$110,012	
		<a href="#">Business</a>	\$400,871	\$89,274	
		<a href="#">Literature</a>	\$296,229	\$57,986	
		<a href="#">Books - Miscellaneous</a>	\$315,929	\$53,007	
		<a href="#">Science &amp; Technology</a>	\$811,787	\$184,275	
		<a href="#">Sports &amp; Health</a>	\$335,106	\$74,724	
Electronics		<a href="#">Audio Equipment</a>	\$3,782,832	\$633,169	
		<a href="#">Cameras</a>	\$5,061,148	\$900,830	
		<a href="#">Computers</a>	\$1,928,998	\$338,585	
		<a href="#">Electronics - Miscellaneous</a>	\$4,671,957	\$810,424	
		<a href="#">TV's</a>	\$3,837,906	\$679,393	
		<a href="#">Video Equipment</a>	\$5,108,464	\$927,202	
Movies		<a href="#">Action</a>	\$617,565	\$37,746	
		<a href="#">Comedy</a>	\$669,642	\$33,243	
		<a href="#">Drama</a>	\$698,840	\$42,376	

Keyword Set:

buy shoes in Boulder Colorado

Rank:	Site Name	Google Business Photos?	POI Photos	Other Images	Google Reviews	Star Rating	DA	PA	Linked Domains	URL Match?	Domain Age (Years)
1	Nordstrom Rack Twenty Ninth Street	NO	0	0	6	3.9	86	39	28402	NO	19.3
2	Boulder Running Company	NO	0	5	175	4.7	44	53	311	NO	13.8
3	Rocky Mountain Kids	NO	0	0	23	4.5	22	34	31	NO	14.7
4	Perry's Shoe Shop Inc	NO	0	2	16	3.3	25	32	24	YES	8.7
5	Pedestrian Shops	YES	16	2	8	3.5	40	47	237	YES	16.7
6	Boulder Army Store	NO	0	0	13	3.7	26	36	41	NO	9.9
7	Two Sole Sisters	NO	0	0	22	4.5	28	39	38	NO	6.2

# Categorical Variables

## Definition

A **categorical variable** is a variable that can take on one of a limited, and usually fixed number of possible values, assigning each individual to a particular group based on some qualitative property. An **ordinal variable** is a categorical variable for which the values can be ordered. A **nominal variable** is a categorical variable that has no ordering.

- Nominal: order not meaningful
  - gender, religion, race
  - type of stock
  - pattern of a carpet
- Ordinal: order may be meaningful
  - grades: A, A-, B+, B, B-, ...
  - educational degrees
  - Likert scales: disagree, neutral, agree

# Numerical variables

## Definition

A **numerical, or quantitative, variable** take numerical values for which arithmetic operations such as adding and averaging make sense.

Examples:

- height/weight of a person
- temperature
- time it takes to run a mile
- currency exchange rates
- number of webpage hits in an hour

For numerical variables, we also consider whether the variable is a **count** and whether or not that count has a technical upper limit.



## 2. Dataset basics - Data types

Aa Aa 

Students in a business statistics class developed a pricing model for diamond stones.

The top and bottom portions of the data set that the students collected are reproduced in the following table; dots indicate that the intervening rows in the data set are not displayed. [Source: S. Singfat Chu, "Pricing the C's of diamond stones," *Journal of Statistics Education* 9(2) (2001).]

Diamond ID	Price (Singapore dollars)	Weight (Carats)	Color	Clarity	Certification Body
1	8,873	1.01	H	VS2	1
2	3,635	0.52	E	VS1	1
3	11,696	1	F	VVS1	3
4	8,095	1	I	VS1	3
5	3,501	0.5	F	VVS2	1
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
304	4,401	0.63	G	VVS2	1
305	2,942	0.46	E	VVS2	1
306	3,706	0.55	F	VVS2	2
307	1,555	0.31	E	VS1	1
308	1,098	0.33	I	VS2	1

Note that color purity is a desirable characteristic of a diamond. A grade of D indicates top color purity, a diamond graded E has less color purity than a diamond graded D, a diamond graded F has less color purity than a diamond graded E, and so on. Clarity is also a desirable characteristic. The top clarity rating is IF (internally flawless); other clarity ratings, in descending order, are VVS1, VVS2, VS1, and VS2. (VVS is the notation for "very, very slightly imperfect," and VS is shorthand for "very slightly imperfect.") Certification Body has three different values, which are coded as 1 = Gemological Institute of America, 2 = International Gemological Institute, and 3 = HRD Antwerp.

Category	Region	Subcategory	Revenue	Profit	Cost
Books		<a href="#">Art &amp; Architecture</a>	\$480,173	\$110,012	
		<a href="#">Business</a>	\$400,871	\$89,274	
		<a href="#">Literature</a>	\$296,229	\$57,986	
		<a href="#">Books - Miscellaneous</a>	\$315,929	\$53,007	
		<a href="#">Science &amp; Technology</a>	\$811,787	\$184,275	
		<a href="#">Sports &amp; Health</a>	\$335,106	\$74,724	
Electronics		<a href="#">Audio Equipment</a>	\$3,782,832	\$633,169	
		<a href="#">Cameras</a>	\$5,061,148	\$900,830	
		<a href="#">Computers</a>	\$1,928,998	\$338,585	
		<a href="#">Electronics - Miscellaneous</a>	\$4,671,957	\$810,424	
		<a href="#">TV's</a>	\$3,837,906	\$679,393	
		<a href="#">Video Equipment</a>	\$5,108,464	\$927,202	
Movies		<a href="#">Action</a>	\$617,565	\$37,746	
		<a href="#">Comedy</a>	\$669,642	\$33,243	
		<a href="#">Drama</a>	\$698,840	\$42,376	

Keyword Set:

buy shoes in Boulder Colorado

Rank:	Site Name	Google Business Photos?	POI Photos	Other Images	Google Reviews	Star Rating	DA	PA	Linked Domains	URL Match?	Domain Age (Years)
1	Nordstrom Rack Twenty Ninth Street	NO	0	0	6	3.9	86	39	28402	NO	19.3
2	Boulder Running Company	NO	0	5	175	4.7	44	53	311	NO	13.8
3	Rocky Mountain Kids	NO	0	0	23	4.5	22	34	31	NO	14.7
4	Perry's Shoe Shop Inc	NO	0	2	16	3.3	25	32	24	YES	8.7
5	Pedestrian Shops	YES	16	2	8	3.5	40	47	237	YES	16.7
6	Boulder Army Store	NO	0	0	13	3.7	26	36	41	NO	9.9
7	Two Sole Sisters	NO	0	0	22	4.5	28	39	38	NO	6.2

# Random variables

## Definition

An **observation** in a data set refers to the observed value of a variable on a specific individual.

## Definition

A **random variable** is the as yet unknown outcome of some observation. We typically denote random variables with capital Roman letters at the end of the alphabet, e.g.  $X$ ,  $Y$ , or  $Z$ .

For example,

- $X$ : monthly unemployment rate
- $Y$ : grade on your next Stat 226 exam, and
- $Z$ : education of customer.

are all examples of random variables.

# Observations

Once we “see” an observation, i.e. the outcome of  $X, Y$  and  $Z$  is determined and no longer unknown, we switch to a lower case letter  $x, y$  or  $z$ . For example, the corresponding **observations** could be:

- $x = 3.9\%$  (for July 2018),
- $y = 95$  points, and
- $z = \text{College graduate}$

**TL;DR** Know the difference between a random variable and an observation (data point) and how to distinguish between them in terms of notation!

- upper case letter  $\implies$  not yet observed
- lower case letter  $\implies$  observed

# Population

## Definition

The **population** is the entire group of individuals that we want to say something about.

## Examples:

- all currently enrolled ISU students
- all Starbucks customers nationwide
- all customers banking with Wells Fargo

The population is entirely defined by the target group of interest and the purpose of the study!

# Sample

## Definition

The subset of the population that you have collected data is called the **sample**.

Examples (of extremely non-representative) samples:

- students in STAT 226, Section A, Fall 2018 (who came to class)
- Starbucks customers visiting 2302 Lincoln Way, Ames from 11-11:30am today
- Wells Fargo customers visiting 3910 Lincoln Way, Ames, IA 50014 today

<https://www.abc15.com/lifestyle/what-too-much-alcohol-can-do-to-your-health>:

## What too much alcohol can do to your health

For example, a [2002 study](#) of almost 25,000 Finnish men and women over five-year intervals found that moderate alcohol consumption, combined with a physically active lifestyle, no smoking and healthy food choices, "maximizes the chances of having a normal weight."

A [2017 study](#) of nearly 2 million Brits with no cardiovascular risk found that there was still a modest benefit in moderate drinking, especially for women over 55 who drank five drinks a week. Why that age? Alcohol can alter cholesterol and clotting in the blood in positive ways, experts say, and that's about the age when heart problems begin to occur.

[Another 2018 study](#) found that consistently drinking a moderate amount of alcohol, within recommended guidelines, had a protective effect on the heart over time. Unstable drinking habits were associated with a higher risk of heart disease, which the authors reflected might indicate broader lifestyle changes, such as poor health or stress. Former drinkers were also at greater risk.



# Descriptive versus Inferential Statistics

## Definition

**Descriptive statistics** is the collection, presentation and description of data in form of **graphs**, **tables**, and **numerical summaries** that provide meaningful information about the sample.

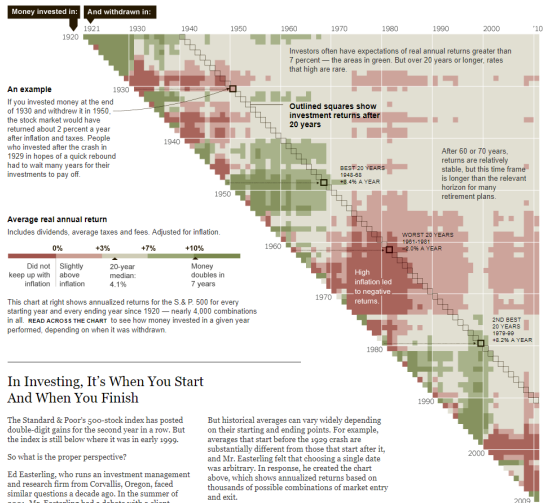
## Goals:

- look for patterns
- summarize and present data

Descriptive statistics focuses on obtaining a better understanding about the **distribution**, **variability**, and **central tendency** that a variable of interest exhibits.

Geomorphological Structure Type	Area (km <sup>2</sup> )	Area (acres)	% of Total Reef Area
<b>Total Coral Reef and Hardbottom</b>	<b>74.8</b>	<b>18473.1</b>	<b>68.8</b>
Pavement	48.5	11981.7	44.6
Aggregate Reef	17.1	4221.7	15.7
Spur and Groove	5.5	1353.4	5.0
Rubble	1.6	384.9	1.4
Aggregated Patch Reef	0.9	217.0	0.8
Rock/Boulder	0.5	115.2	0.4
Individual Patch Reef	0.5	113.2	0.4
Scattered Coral/Rock	0.3	86.0	0.3
<b>Total Unconsolidated Sediment</b>	<b>33.9</b>	<b>8376.5</b>	<b>31.2</b>
Sand	33.4	8251.9	30.7
Mud	0.5	124.6	0.5
<b>Total Reef Area</b>	<b>108.7</b>	<b>26872.1</b>	<b>100.0</b>

Table B. Thematic content summary of geomorphological structure



# Inferential Statistics

## Definition

**Inferential statistics** deals with drawing conclusions and making generalizations based on data for a larger group of subjects (a population).

## Goals:

- making statements about the population
- making data-based decisions

## **Your Brain Tries to Change Focus Four Times per Second, Study Finds**

Depressed patients see quality of life improve with nerve stimulation

Study focuses on people not treated effectively with antidepressants

## **A Low-Carb Diet Could Cut 4 Years Off Your Life, So Just Eat the Damn Pasta**

Keto dieters, be warned.

# Statistic

## Definition

A (summary or sample) **statistic** is any function of the data.

Examples:

- Mean, median, mode
- Tables
- Charts, figures

# Parameter

## Definition

A (population) **parameter** is a characteristic of the population.

Examples:

- Mean summary salary of ISU students
- Median expenditure of Starbucks customers
- Standard deviation of savings account dollars of Wells Fargo customers

Numerical statistics are often used to estimate population parameters.

## Iowa Governor - Reynolds vs. Hubbell



RCP Senate Map | Senate Polls | RCP House Map | Generic Vote | RCP Governor Map | Governor Polls | All 2018 Polls

Candidates		Iowa Snapshot
 Kim Reynolds (R)* <a href="#">Bio   Campaign Site</a>	 Fred Hubbell (D) <a href="#">Bio   Campaign Site</a>	<p>RCP Ranking: <b>Leans GOP</b></p> <p>-----PAST KEY RACES-----</p> <p> <b>2016:</b> President   Senate   IA-1, IA-3  <b>2014:</b> Governor   IA-1   IA-2   IA-3   IA-4  <b>2012:</b> President   IA-1   IA-2   IA-3   IA-4  <b>2010:</b> Governor   Senate   IA-1   IA-2   IA-3  <b>2008:</b> President  <b>2006:</b> Governor   IA-1   IA-3  <b>2004:</b> President   Senate   IA-3           </p>

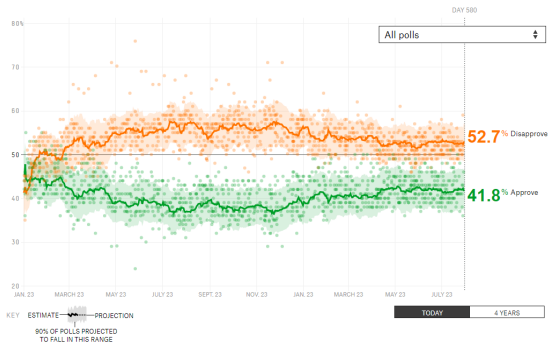
Polling Data						
Poll	Date	Sample	MoE	Reynolds (R)	Hubbell (D)	Spread
Des Moines Register	1/28 - 1/31	555 LV	4.2	42	37	Reynolds +5

The proportion of voters who will vote for Reynolds (parameter) is estimated to be 42% (statistic) with a 95% confidence interval of  $42\% \pm 4.2\% = (37.8\%, 46\%)$  (statistic).



## How **unpopular** is Donald Trump?

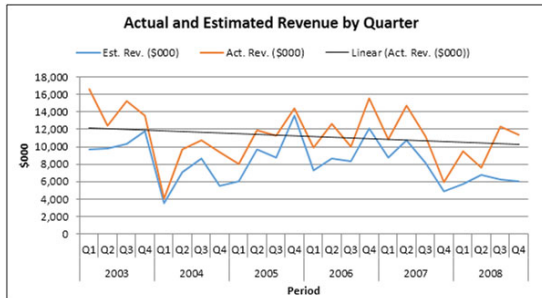
An updating calculation of the president's approval rating, accounting for each poll's quality, recency, sample size and partisan lean. [How this works »](#)



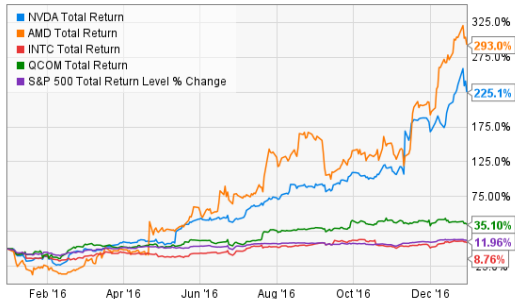
DATES	POLLSTER	GRADE	SAMPLE	WEIGHT	APPROVE		DISAPPROVE		ADJUSTED
* AUG. 19-21	Rasmussen Reports/Pulse Opinion Research	C+	1,500 L V	all 0.95	46%	52%	41%	52%	
* AUG. 19-20	YouGov	B	1,000 R	all 0.30	43%	51%	44%	53%	
* AUG. 17-20	American Research Group	C-	1,100 R	all 1.15	36%	59%	38%	56%	

# Time series

Sometimes, variables are **collected over time**. Typically plot these data as a **time series** where time is on the x-axis.



## 2016 Total Returns: NVIDIA vs. Select Competitors/Peers &amp; the S&amp;P 500



The Motley Fool

Dec 30 2016, 11:32PM EST. Powered by **YCHARTS**