Chapter 0 and Chapter 1: An Introduction to Statistics

1 Welcome

Welcome to MATH 2332: Probability and Data Analysis. Probability and Data Analysis is an outstanding course as part of a technical education. Everyone (sciences and humanities) can benefit from a better understanding of statistics and data. Everyone. Statistics is important in every day life and decision making. Daily news presents statistics about issues both important and mundane. What does the margin of error in a political poll mean? Which of those side effects from a new medication should I worry about? What past performance data should I consider when drafting my fantasy sports team? What past performance data should I consider when making any decision about the future? Why are canned soups and ice cream frequently "buy one get one free" at the grocery store but gasoline never is? Adding statistical skills into your studies of mathematics creates breadth in a working world where data is easy to collect but challenging to organize and understand.

2 What is Statistics?

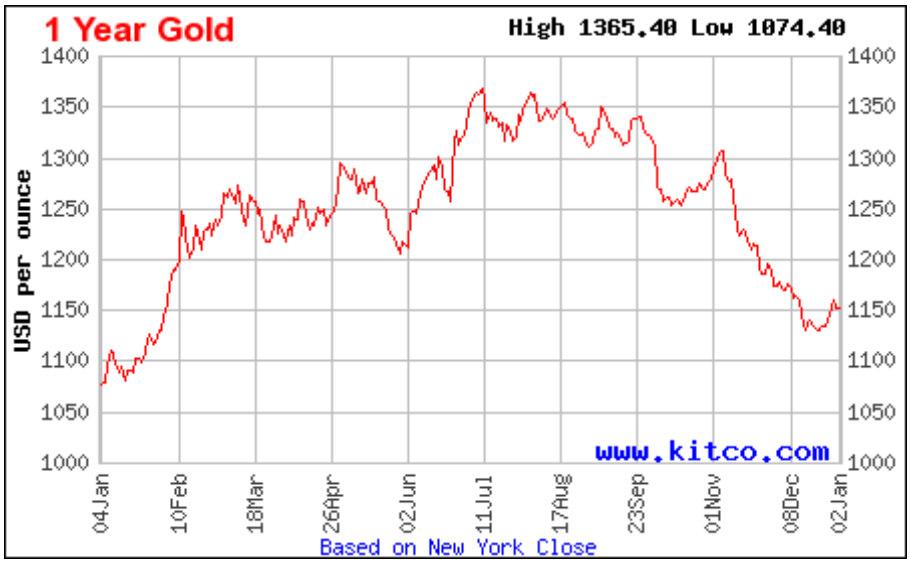
Statistics is the science of data! Collecting, classifying, organizing, analyzing, interpreting, making decisions from, etc. Statistics is a part of your everyday life even if you haven't always noticed.

Example 1 *The average score on Test 1 in STAT 8000 at the University of Nowhere is 75.*

Example 2 *The tallest ice cream cone was over 9 feet tall and scooped in Italy.(http://www.icecream.com/icecreaminfo).*

Example 3 *California produces the most ice cream in America (http://www.icecream.com/icecreaminfo).*

Example 4 *How has the price of gold fluctuated in 2016? A* ***Time Series*** *studies changes in variables over time.*



http://www.kitco.com/charts/popup/au0365nyb.html

Example 5 *Top 5 interstates for most fatal accidents per mile in 2013 (http://commuting.blog.ajc.com/2015/1 deadliest-interstate-is-in-georgia-study-says/)*

I-285 in Georgia
I-710 in California
I-240 in Oklahoma
I-495 in Delaware
I-240 in Tennessee

Example 6 *Georgia is the seventh-worst state in the country for fatal car accidents in total (1,085 incidents in 2013). Texas ranked no. 1, with 3,044 deaths from car accidents in 2013.*

The first example provided here is a work of fiction. I made up the data and it looks nice. I refer to such examples as toy data. Such data sets are good to play with and can make a point. The other examples are real. As a science, Statistics is very important because of its applications in the real world. I grab lots of example data from Wikipedia. It is a nice source for pop culture stuff. Most all sports data comes from http://www.sports-reference.com/. It contains lots of data about baseball, basketball, football, hockey and the Olympics. Using real data can help answer the timeless question in every statistics/mathematics course. What is this good for?

3 Raw Data is Ugly

Graphical representations of data always look pretty in newspapers, magazines and books. What you haven't seen is the blood, sweat and tears that it sometimes takes to get those results.

3.1 State of Birth for US Players in NHL

Example 7 *Consider the state of birth for US players in the NHL.*

NHL Players Born in Georgia, United States

Change birth place: United States | Georgia | Find Players

3 Players Glossary • CSV • PRE

					Scoring Stats						Goalie Stats							
Rk	Player	From	To	Pos	GP	G	A	PTS	+/-	PIM	GP	W	L	T/O	SV%	GAA	Birth Date	Birth City
1	Eric Chouinard	2001	2006	LW	90	11	11	22	-8	16							1980-07-08	Atlanta
2	Mark Mowers	1999	2008	C	278	18	44	62	0	70							1974-02-16	Decatur
3	Jean-Marc Pelletier	1999	2004	G	7	0	0	0	0	0	7	1	4	0	.857	3.90	1978-03-04	Atlanta

NHL Players Born in Hawaii, United States

Change birth place: United States | Hawaii | Find Players

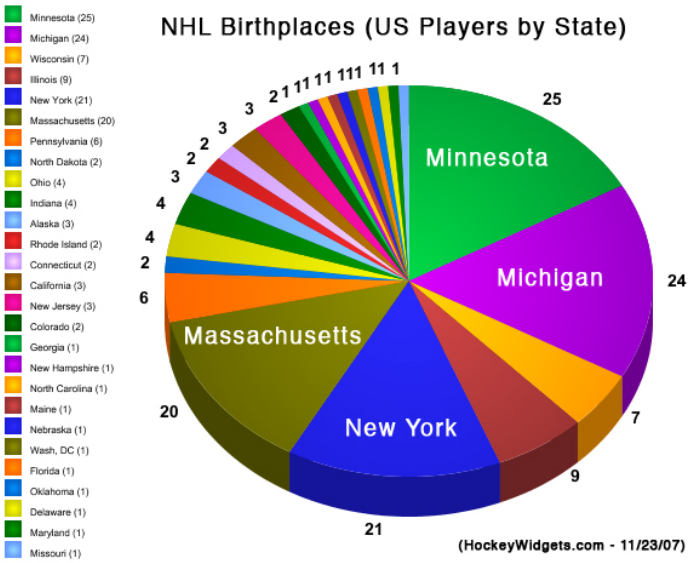
No NHL players born in Hawaii, United States.

NHL Players Born in New York, United States

Change birth place: United States | New York | Find Players

92 Players Glossary • CSV • PRE

					Scoring Stats						Goalie Stats							
Rk	Player	From	To	Pos	GP	G	A	PTS	+/-	PIM	GP	W	L	T/O	SV%	GAA	Birth Date	Birth City
1	Tom Askey	1998	1998	G	7	0	0	0	0	0	7	0	1	2	.894	2.64	1974-10-04	Kenmore
2	Zach Bogosian	2009	2012	D	264	29	60	89	-37	208							1990-07-15	Massena
3	Jason Bonsignore	1995	1999	C	79	3	13	16	-22	34							1976-04-15	Rochester
4	Francis Bouillon	2000	2012	D	676	29	105	134	-26	481							1975-10-17	New York
5	Jesse Boulerice	2002	2009	RW	172	8	2	10	-17	333							1978-08-10	Plattsburgh
6	Rich Brennan	1997	2003	D	50	2	6	8	-8	33							1972-11-26	Schenectady
7	Greg Britz	1984	1987	RW	8	0	0	0	-1	4							1961-01-03	Buffalo
8	Dustin Brown	2004	2012	LW	595	163	196	359	-30	430							1984-11-04	Ithaca
9	Jack Brownschidle	1978	1986	D	494	39	162	201	-49	151							1955-10-02	Buffalo
10	Jeff Brownschidle	1982	1983	D	7	0	1	1	-9	2							1959-03-01	Buffalo



http://www.hockeywidgets.com/newblog/uploaded_images/graph2-775937.jpg

3.2 Health Statistics

Example 8 *Here are the first 10 entries for a data set containing health information. Note that each row is an* ***observation*** *about an* ***individual*** *and each column is a* ***variable****. The full data set has 5209 observations. In this example the observations are made about one person. Many variables such as gender, weight and height are recorded for each observation.*

Obs	Status	DeathCause	AgeCHDdiag	Sex	AgeAtStart	Height	Weight	Diastolic	Systolic	MRW
1	Dead	Other	.	Female	29	62.50	140	78	124	121
2	Dead	Cancer	.	Female	41	59.75	194	92	144	183
3	Alive		.	Female	57	62.25	132	90	170	114
4	Alive		.	Female	39	65.75	158	80	128	123
5	Alive		.	Male	42	66.00	156	76	110	116
6	Alive		.	Female	58	61.75	131	92	176	117
7	Alive		.	Female	36	64.75	136	80	112	110
8	Dead	Other	.	Male	53	65.50	130	80	114	99
9	Alive		.	Male	35	71.00	194	68	132	124
10	Dead	Cerebral Vascular Disease	.	Male	52	62.50	129	78	124	106

Obs	Smoking	AgeAtDeath	Cholesterol	Chol_Status	BP_Status	Weight_Status	Smoking_Status
1	0	55	.		Normal	Overweight	Non-smoker
2	0	57	181	Desirable	High	Overweight	Non-smoker
3	10	.	250	High	High	Overweight	Moderate (6-15)
4	0	.	242	High	Normal	Overweight	Non-smoker
5	20	.	281	High	Optimal	Overweight	Heavy (16-25)
6	0	.	196	Desirable	High	Overweight	Non-smoker
7	15	.	196	Desirable	Normal	Overweight	Moderate (6-15)
8	0	77	276	High	Normal	Normal	Non-smoker
9	0	.	211	Borderline	Normal	Overweight	Non-smoker
10	5	82	284	High	Normal	Normal	Light (1-5)

Let's drill down on the variable, Cholesterol.

Analysis Variable : Cholesterol				
N	Mean	Std Dev	Minimum	Maximum
5057	227.4174412	44.9355238	96.0000000	568.0000000

What does $n = 5057$ imply? Data is sometimes incomplete.
Next let's drill down on the variable, Sex.

Sex	Frequency	Percent
Female	2873	55.15
Male	2336	44.85

Why did we not compute a mean (average) for Sex?

Definition 9 *A data set with a single variable for each observation is called* ***univariate****. The data set containing only the birth state for NHL players is univariate.*

Definition 10 *A data set with two variables for each observation is called* ***bivariate****. A data set with more than two variables for each observation is called* ***multivariate****. The health data set above is multivariate.*

There are two types of variables in Data Science.

Definition 11 ***Quantitative*** *variables are data that consist of numbers.*

Definition 12 ***Categorical*** *(or* ***Qualitative****) variables are data that do not consist of numbers.*

Problem 13 *For the health data set above, classify each variable as qualitative or quantitative.*

Furthermore, there are two types of quantitative variables. Each type creates a very different probability model with its own set of rules and computations (as we will see in future sections).

Definition 14 ***Discrete variables*** *can take one of a finite number of distinct outcomes. Discrete random variables jump from one state to the next with nothing in between.*

Example 15 *The number of unopened cans of Coke in my fridge, the number of players on a basketball team and the number of cards in a deck are all discrete variables.*

Definition 16 ***Continuous variables*** *can take any numeric value within a range of values.*

Example 17 *The number of ounces in a can of Coke, the time it takes (in minutes) to run a mile, the temperature outside are all examples of continuous random variables.*

Exercise 18 *Consider an experiment whose population is the set of all Kennesaw State University students. Which of these variables are discrete and which are continuous?*

1. number of classes the student is enrolled in this semester;
2. number of required books for classes this semester;
3. weight (in pounds) of required books for classes this semester;
4. cost (in dollars) of required books for classes this semester;
5. height (in inches) of student;
6. GPA;
7. number of miles driven to campus.

A **population** is the entire collection of data that describe some phenomenon. The above data contains only 5209 observations and thus is not the population of the world, the US or a state. A **sample** is a subset of a population from which we actually collect data. The above data is a sample.

A number that describes a population is called a **parameter**. A number that describes a sample is called a **statistic**.

Example 19 *In Fall 2013, the average age of KSU undergraduate students was 23 (factbook.kennesaw.edu). Since this describes the population of KSU students, this number is a parameter.*

Example 20 *If we use the students present today as a sample of KSU students and find their average age to be 21 then that is a statistic.*

Definition 21 *Using a sample to make an inference about a population is called* ***inferential statistics****.*

We frequently use samples when the population is too large to gather. Say we wish to determine the average number of hours per week students at KSU prepare for classes. It might be very difficult and time consuming to collect this data from every student. It would be easier to use a sample. One possible sample is to use is this class. Another sample would be to use all the students present in this classroom whose birthday is in May. A third sample would be to go to the library and use every student present. A fourth sample would be to place an ad in the school paper inviting students to submit information. The **sampling design** describes how the sample is selected. Unfortunately not all samples are equally useful when practicing inferential statistics.

Problem 22 *Are any of the above sampling designs good to use to estimate the average number of hours per week students at KSU prepare for classes? Explain.*

Definition 23 *A sample selected by taking individuals of the population that are easy to reach is called a* ***convenience sample****. Asking students in this particular class the average number of hours per week they prepare for classes is a convenience sample.*

Definition 24 *A sample selected by taking groups of collected individuals of the population is called a* ***cluster sample****. Sampling every student in a ENG 1101, HIST 1101, ART 1101, and STAT 1107 is a cluster sample.*

Definition 25 *A* ***systematic sample*** *is one where every k^{th} member of the population is included in the sample. Sampling every 100th student leaving the commons is a systematic sample.*

Definition 26 *A sample consisting of people who choose themselves to respond to an appeal for opinions is a* ***self-selected sample*** *or a* ***voluntary response sample****. Putting an ad in the Sentinel requesting information is a self-selected sample.*

Definition 27 *A* ***stratified*** *sample is one where different types of objects are sample in different quantities to ensure representative proportion. Sampling 1000 students at the Kennesaw campus and 300 students at the Marietta campus is a stratified sample.*

Self-selected samples almost always provide biased results.

Problem 28 *Identify misuses of statistics in the following: A local website posted a poll asking readers if Atlantans should approve a 2% increase in sales tax to purchase new stadiums for millionaire athletes? Ten people responded, 83% said "no'" and 17% said "yes" This website reported that Georgians do not want an increase in sales tax to fund local arts and entertainment.*

What makes for a good sample? A simple random sample.

Definition 29 *A **simple random sample** (SRS) of size n is a sample of n individuals from the population such that every individual has an equal chance of inclusion.*

An SRS is like a lottery selection.

4 Understanding Variation

Knowing an average value (or a predicted average) is a good start to understanding data. But it is not enough. We must also understand how data varies. The same unit of measurement may not always have the same value or meaning.

Example 30 *The average score on Test 1 in STAT 8000 at the University of Nowhere is 75. Of the 100 students in the class, half scored a 50 and the other half scored 100.*

Example 31 *The average score on Test 1 in STAT 8000 at the University of Nowhere is 75. Each of the 100 students in the class scored a 75.*

Example 32 *Consider the weight of individuals from our health data set. Is this the best way to present information about weight?*

Analysis Variable : Weight				
N	Mean	Std Dev	Minimum	Maximum
5203	153.0866808	28.9154261	67.0000000	300.0000000

Perhaps we should break down the observations by gender.

Analysis Variable : Weight						
Sex	N Obs	N	Mean	Std Dev	Minimum	Maximum
Female	2873	2869	141.3886372	26.2880439	67.0000000	300.0000000
Male	2336	2334	167.4661525	25.2907044	99.0000000	276.0000000

This comparison is even more difficult when the contexts are dramatically different.

Example 33 *Which music sales record is more impressive (as of June 15, 2016 according to Wikipedia):*

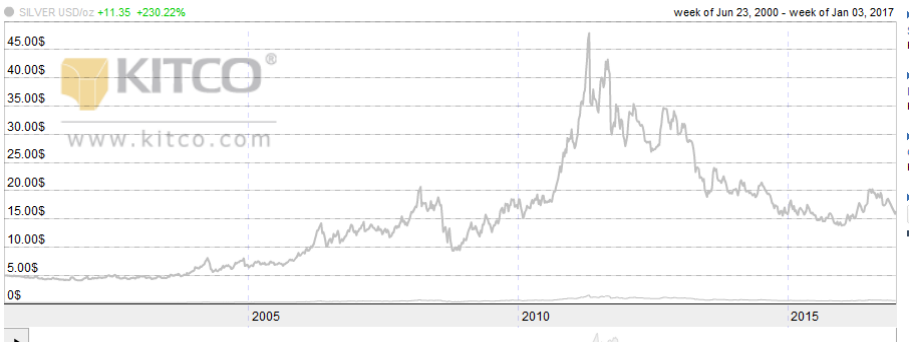
Michael Jackson, *Thriller* (48.1 million albums)
or
Bing Crosby, *White Christmas* (50 million physical singles)
or
Wiz Khalifa featuring Charlie Puth, *See You Again* (20.9 million digital singles)?
We will develop sophisticated techniques to measure variation even in seemingly incomparable situations.[1]

We also need to understand variation in order to determine when a difference exists between two groups or if it is just chance variation in the sampled data. In our health data set, is there a difference in average weight between men and women? Is there a difference in the rate of drinking between men and women? (https://www.theguardian.com/society/2016/oct/24/women-drink-alcohol-men-global-study)

5 Testing Claims

No matter what time of year it is, I see ads on a particular channel claiming that now is the best time ever to buy silver! Is it?

[1] Everyone who told you you cannot compare apples and oranges was lying!

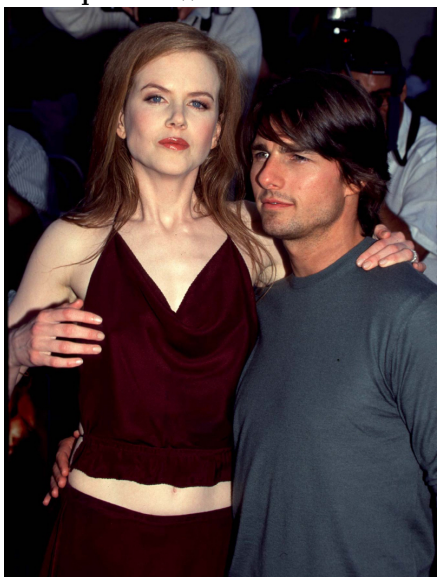


There is clear bias on the part of the advertiser. Bias is extremely damaging to the reliability of statistics.

6 Bias in Samples

The practice of statistics begins with data! All the mathematics is useless unless the data is an accurate representation of reality. Collecting good data is time consuming, messy and occasionally painful. Conducting surveys for data is a horrible technique. It is best avoided when possible. It is best to observe and record data.

Example 34 *We don't ask men how tall they are. We measure their height.*



Example 35 *We don't ask politicians for their income. We ask to see tax returns that document income.*

Example 36 *We don't ask Mean Joe Greene if he prefers Coke or Pepsi. We give him a blind test taste (if we dare).*

Blind taste tests may not be perfect either. Wikipedia has an interesting article on the Pepsi Challenge (https://en.wikipedia.org/wiki/Pepsi_Challenge). One theory is that Pepsi won over Coke because Pepsi is sweeter and with a small cup it made a better impression. It might be inappropriate to imply Pepsi is favored over Coke over longer periods of times. What was overt bias in the taste test is that Pepsi was served chilled and Coke was served at room temperature. Doing so is an example of **self-interest bias** in a study.

Example 37 *If we poll KSU students about the number of hours they study each week or how much alcohol they drink, we might not have confidence in the results. Many undergraduates are under 21 years of age and might not admit to committing a crime. Or wish to look like a hard partier and exaggerate. One might not want to appear to study either too little or too much. These cases are examples of* ***social acceptability bias*** *in a study.*

Example 38 *If we invite students to login to a website to rate their professors, we might only get opinions of the very passionate students (in either direction). This is an example of* ***voluntary*** *(or* ***self-selected****)* ***response bias****.*

Example 39 *"Do you support the development of weapons that could kill millions of innocent people?" Of course you don't if the question is asked in that manner. This is an example of* ***leading question bias****.*

Example 40 *A study asks survivors of motorcycle accidents if they think motorcycles are safe. This is an example of a study that has* ***nonresponse bias****. Those riders who did not survive the accident cannot participate.*

Problem 41 *What are the consequences of sample bias?*

We want to know which baseball team in the 2011 season is the best? Should we just ask the players, managers, team owners and sportswriters which team is best? No! We would be collecting opinions if we did that. The games are played in order to determine which team is best. Even still, playing the games and recording wins may not always provide an easy answer.

Problem 42 *Is the best team in baseball (or football, etc.) the team that wins the World Series (or Superbowl, etc.) each year?*

Problem 43 *Is the best team in baseball (or football, etc.) the team that wins the most games each year?*

Since wild card teams have won the World Series, the answer to both questions cannot be *yes*. From the Chicago Tribune on September 26, 2016 (http://www.chicagotribune.com/sports/spt-cubs-best-record-playoffs-gfx-20160916-htmlstory.html),

> "In a methodical stairstep over two weeks the Cubs locked up the National League Central, home field throughout the NL playoffs and now, the best record in all of Major League Baseball. The bottom line: The team with the best record in baseball usually doesn't win the World Series. Since the beginning of the wild card era only four of the 26 teams with the top record went on to win the World Series: The 1998 and 2009 Yankees, and the 2007 and 2013 Red Sox. That's 7 percent. But don't despair, Cubs fans, the best teams each season have winning records overall in both the division series and the championship series. In the division series, 15 top-record teams won and advanced to the championship series, and then nine went on to play in the World Series."

Perhaps we can never truly know which team is best despite trying to statistically determine it. Statistics frequently tries to answer questions without the ability to ever **know** the right answer.

Problem 44 *How can I determine the effectiveness of my teaching STAT 1107? Should I rely on end-of-semester student evaluations? Check out the web site Rate Your Professor?*

7 Making Sound Data-based Decisions (Predicting the Future)

A key skill in statistics is to predict the unknown (or a future event) by analyzing patterns in data sets. It is a safe bet that retail stores in a mall should hire extra help starting around Thanksgiving. It is also a safe bet to not count on that job lasting past the first week of January.

Predicting the future based on past results works only up to a point. Sales of VHS tapes steadily rose for many, many years. Then they didn't while DVD sales rose. Now, the format of choice is Blu-ray. Unforeseen events or catastrophes wreak havoc with predictions. Wayne Gretksy led his team in scoring in his first 14 seasons as a professional hockey player. It seams natural to predict he would do so again in his 15th season. However a back injury prevented him from playing as many games as he usually did.

8 Chance

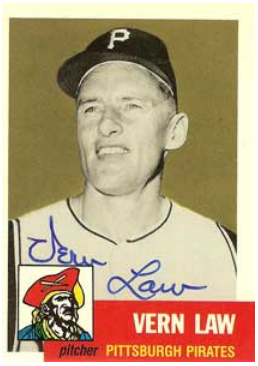
Elvis Presley and David Bowie share the same birthday (January 8). Is this surprising? Is this meaningful? Understanding probability is key to using data to make predictions or estimations.

How many people do you need in a room to have a 50% chance that at least two will share a birthday (disregarding year)?

9 Literacy and Homework

Mark Twain once said that "the man who doesn't read good books has no advantage over the man who can't read them." Reading and writing are critical skills for this class! Yes, this **Statistics** class! One must be careful with words and pay attention to their meaning. We rarely "solve for x" in this course. We analyze scenarios in order to determine what is going on and apply the best technique available. That requires reading skills.

Vern Law, the 1960 Cy Young winner, one said "experience is a hard teacher because she gives the test first, the lesson afterward." This class is much easier since the lessons and homework are provided before the tests. Of course, you must actually do the homework for the lesson to occur before the test!



10 Exercises

1. Identify five misuses of statistics in the following statement.
   A sports reporter on the local country radio station asked his listeners to call in and answer the following question: Do you support the use of random drug tests to catch cheaters who soil the reputation of Baseball by using steroids? The DJ reported that twenty people responded, 87% said yes and 13% said no. The DJ concluded that Americans overwhelmingly support random drug testing.

2. Kokoska 3rd edition: Read Chapter 0
3. Kokoska 3rd edition: Section 1.2: 1.1, 1.2, 1.4, 1.6 a-d, 1.8, 1.9
4. Kokoska 3rd edition: Section 1.3: 1.23, 1.25, 1.27, 1.28, 1.30, 1.31, 1.33, 1.35
5. Kokoska 3rd edition: Section 2.1: 2.1-2.7, 2.12