

Data Science Module on the Types of Variables, Summary Statistics & Graphical Displays

Prerequisites/Assumptions. Algebra.

Duration. The duration of this module depends heavily on how much previous statistics students have had. Students with AP Statistics can skip this. Students who have never seen mean, median, etc. might require more time.

Tips for teaching.

Additional Reading

1. ???
2. ???

Lecture – See ToV SummStat Graphics.pptx

Learning Outcome – Types of Variables, Summary Statistics, Graphical Displays

Goal #1: Identify Types of Variables

Key Ideas

The following table breaks down variable types for storage/computing and analysis

Analysis	Storage/Computing	Examples
Categorical – Dichotomous (2 categories)		Gender (if M/F) Aged 65 or more? (yes/no)
Categorical – Polychotomous (>2 categories)		Political Party Religious Affiliation
Ordinal (labels, but with a specific ordering)		Class year (Fr, So, Ju, Se) Grade (A+, A, A-, etc.)
Numeric (sometimes separated into discrete (finite) and continuous (measurement))	Continuous – specify precision Date/Time	Age GPA
Text Analysis	Text / String	

Video Resource

- [Types of Variables](#) – OpenIntro

Goal #2: Calculate and interpret summary statistics. Understand properties, including resistance to outliers.

Key Ideas

Types of summary statistics:

Numeric data

Center – mean, median, (mode)

Dispersion – variance/standard deviation, range, interquartile range

Additional – quartiles, percentiles

Categorical data

Frequency (or count), percent (or relative frequency)

Ordinal data

Frequency, percent, cumulative frequency, cumulative percent

For two categorical/ordinal variables:

Contingency Tables: Numerical counts of two categorical variables, including marginal distributions, joint distributions, and conditional probability.

For two numeric variables: correlation, line of best fit (regression), and R^2

Video Resources

- [Mean, median, and mode](#) – OpenIntro
- [The Effects of Outliers](#) – StatisticsLectures.Com
- [Line fitting, residuals, and correlation](#) - OpenIntro

Goal #3: Create and interpret graphical displays to summarize NUMERIC data.

Key Ideas

Dotplot: Used for summarizing a numerical variable. Can be split with a number of categorical variables.

Histograms: Used for summarizing a numerical variable. This is the most common form.

Boxplot is graphical representation of the five-number summary (min, Q_1 , med, Q_3 , max)

It can be split with a number of categorical variables.

Outlier limits: $Q_1 - 1.5 \text{ IQR}$ and $Q_3 + 1.5 \text{ IQR}$

Extreme Outlier limits: $Q_1 - 3 \text{ IQR}$ and $Q_3 + 3 \text{ IQR}$

Whisker ends at most extreme non-outlier, outliers denoted with different symbol

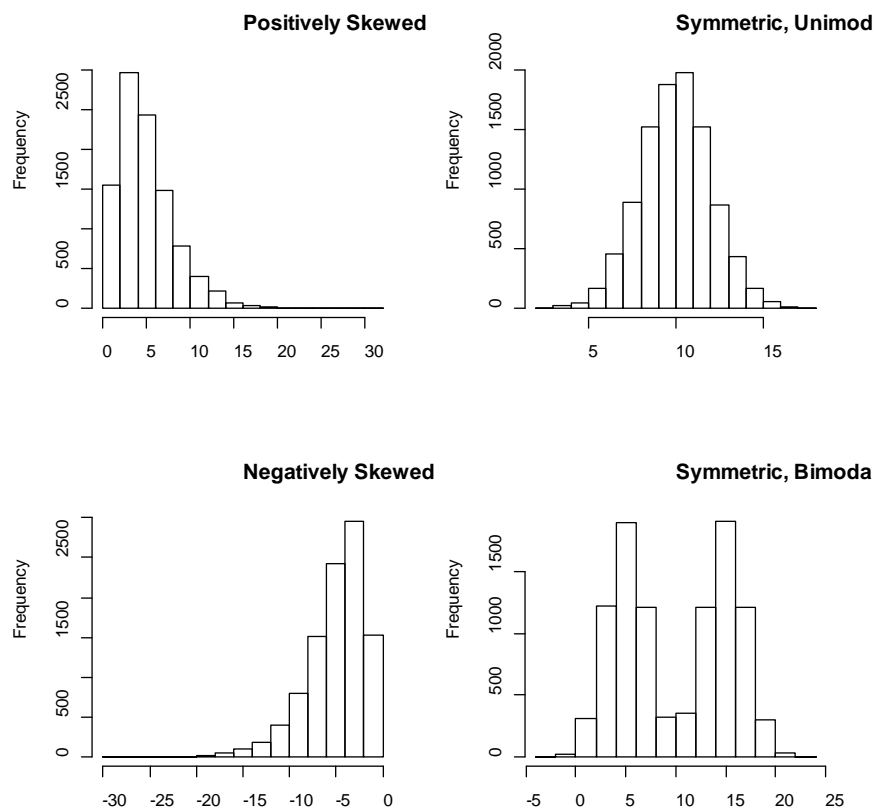
Efficient visualization when splitting by categorical variable(s).

Scatterplots: Used for comparing two numeric variables.

Describing a distribution: Shape (skewed, symmetric), Center, Dispersion, Other

What else is interesting? What is the takehome message? (SOCS+IT)

mean > median => positive skew, mean < median => negative skew



Video Resources

- [Examining Numeric Data](#) – OpenIntro
- [Distributions of Numeric Variables](#) – OpenIntro
- [Reading BoxPlots](#) – Khan Academy

Goal #4: Create and interpret graphical displays to summarize CATEGORICAL data.

Key Ideas

Pie Charts: Used for summarizing a single categorical variable

Bar Charts: Used for summarizing any number of categorical variables

Video Resources

- [Exploring Categorical Data](#) - OpenIntro
- [Relationships Between Categorical Variables](#) – OpenIntro

Additional Resources:

These videos were assembled and rated on their Statistical Content, Production Quality, and Entertainment Value by one or two undergraduate students (last column). Each is also categorized as Procedural, Conceptual, or Informative (second to last column). Some extend beyond the scope for this course.

Types of Variables:

<u>Medium</u>	<u>Statistical Topic</u>	<u>Application Area</u>	<u>Source</u>	<u>Title</u>	<u>PCI</u>	<u>Rating (S/P/E)</u>
Video	Types of variables		konan351		C	5/2/1
Video	Types of variables		Statslectures		C	5/5/2
Video	Types of variables		CreativeHeuristics		C	5/4/3
Video	Types of variables		statslectures		C	5/5/2
Video	Types of variables		Yourteacher.com		C	4/2/1
Video	Types of variables		wjschne		C	5/4/2
Video	Types of variables		sowk300	Independent Dependent Variable	C	4/2/1

Summary Statistics:

<u>Medium</u>	<u>Statistical Topic</u>	<u>Application Area</u>	<u>Source</u>	<u>Title</u>	<u>PCI</u>	<u>Rating (S/P/E)</u>
Video	Categorical Data		ProfessorRosever		P	5/2/2
Video	Categorical Data		ProfessorRosever		P	5/3/2
Video	Categorical Data		ProfessorRosever		P	5/3/2
Video	Categorical Data	Electronics	RRCwiseguys		P	5/4/2
Video	Categorical Data		ProfessorRosever		C	4/2/2
Video	Categorical Data		ProfessorRosever		C	5/2/2
Video	Categorical Data		TwoMinuteTeacher		C	5/2/2
Video	Categorical Data		ProfessorRosever		C	5/3/3
Video	Categorical Data	Health	Against All Odds	Vehicle emissions standards and medical studies of cholesterol provide real-life examples of how to use Normal Calculations An illustrative example shows how a city government used statistical methods to correct inequity between men's and women's salaries.		//
Video	Center	Income	Against All Odds		PCI	//
Video	Center		David Mortimer		PC	5/4/2
Video	Center	Popularity	Udacity	Introduction/Overview to Statistics	PC	5/5/4
Video	Center		ProfessorRosever		P	5/1/1
Video	Center		yourteachermathhelp		P	5/2/1
Video	Center		ProfessorRosever		P	5/1/2
Video	Center		ProfessorRosever		P	5/1/2
Video	Center		ProfessorRosever		P	5/2/2
Video	Center		ProfessorRosever		P	5/2/2
Video	Center		statslectures		P	5/5/2
Video	Center		ProfessorRosever	Logic questions based on the mean 4.6 Statistical Mythbusters (Most Better than Average)	C	4/2/2
Video	Center		Udacity		C	5/5/4

Video	Center		Khan Academy			5/5/3
Video	Center/ Dispersion		profrobbob		C	3/4/3
	Chebyshev's					
Video	Therem		ProfessorRosever	(Not required for our class)	P	5/2/2
Video	Dispersion		ProfessorRosever		P	4/3/2
Video	Dispersion		ProfessorRosever		P	5/2/2
Video	Dispersion		ProfessorRosever		P	5/2/2
Video	Dispersion		ProfessorRosever		P	5/2/2
Video	Dispersion		ProfessorRosever		P	5/3/2
Video	Dispersion		ProfessorRosever		P	5/3/2
Video	Dispersion		ProfessorRosever		P	5/3/2
Video	Dispersion		ProfessorRosever		P	5/3/2
Video	Dispersion		ProfessorRosever		P	5/4/2
				Reading Bar Graphs to Compare Standard Deviations		
Video	Dispersion		ProfessorRosever		C	4/3/2
Video	IQR		dougsimmsonline		P	5/2/1
Video	IQR		TheMathClips		P	5/4/2
Video	Mean	Age	Udacity		PC	5/5/4
Video	Mean		Khan Academy		P	5/5/3
Video	Mean vs. Median		maysterchief		C	4/3/2
Video	Mean vs. Median		kbower50		C	5/5/2
	Mean, Median,					
Video	StDev		radicaltutor		C	5/4/3
Video	Mean, St Dev		Profrobbob		C	5/4/3
Video	Mean, StDev		ExamSolutions		C	3/3/2
Video	Mean, StDev		kbower50		C	5/4/2
Video	Median, IQR		ExamSolutions		P	5/3/2
Video	Percentiles/Quartiles	Age	Udacity	Unit 4 Outliers and Normal Distribution	PC	5/4/4
Video	Percentiles/Quartiles		ProfessorRosever		P	5/3/2
Video	Range		Khan Academy			5/5/3

Video	Sample vs. Pop Mean	Height	Khan Academy		P	5/5/3
Video	SE Mean		Khan Academy	Standard Error of the Mean (move to SD/CLT)	C	5/5/3
Video	StDev		ExamSolutions		PC	5/5/3
Video	StDev		treelmuw		P	4/3/2
Video	StDev		Khan Academy		P	5/4/3
Video	StDev		EnderlePhD		C	4/3/2
Video	StDev		gm2231		C	5/3/2
Video	StDev		TAMUC DrDawg		C	5/3/3
Video	StDev, IQR		Educator		P	4/4/3
Video	Variance	Age	Udacity		PC	5/4/4
Video	Variance		Khan Academy	Alternate Variance Formulas	C	5/3/3
Video	Variance		Khan Academy	Sample Variance	C	5/4/3
Video	Variance, StDev		statisticsfun		P	5/5/2

Graphical Displays:

<u>Medium</u>	<u>Statistical Topic</u>	<u>Application Area</u>	<u>Source</u>	<u>Title</u>	<u>PCI</u>	<u>Rating (S/P/E)</u>
Video	Bargraphs		jlinchstats		PC	5/3/2
Video	Bargraphs		Prof. Rosever		P	4/2/2
Video	Bargraphs		Khan Academy	Bargraphs	P	5/4/3
Video	Bargraphs		Prof. Rosever	Computations for Bar Charts	P	3.5/2/2
Video	Bargraphs		ProfessorRosever	Reading Bar Graphs to Compare Means	C	4/3/2
Video	Bargraphs		ProfessorRosever	Reading Bar Graphs to Compare Means	C	4/3/2
Video	Bargraphs	Valuing Houses	Udacity		C	5/4.5/4
Video	Bargraphs, Histograms		Alice Keeler		PC	5/2/2
Video	Bargraphs, Piecharts		StatsLectures		P	5/4/2
Video	Boxplots	Public Health	ProfessorMcComb		PC	5/3/2
Video	Boxplots		Educator		PC	5/4/3
Video	Boxplots		Brainingcamp		PC	5/4/2.5
Video	Boxplots		jimbabweiberg	Interpretation	P	5/2/2
Video	Boxplots		ExamSolutions	Box and Whisker Plots : Skewness and Outliers. The Five Number Summary, Interquartile	P	5/4/2
Video	Boxplots		StatsLectures	Range(IQR), and Boxplots	P	5/4/2
Video	Boxplots		Khan Academy	Boxplots	P	5/5/3
Video	Boxplots		Khan Academy	Reading Box and Whisker Plots	P	5/5/3
Video	Boxplots		Arnoldsk		P	5/2/1.5
Video	Boxplots		Prof. Rosever		P	5/3.5/2.5
Video	Boxplots		Rcvmathprof	Outliers present	P	5/4.5/2.5
Video	Boxplots		TenMarksInstructor		C	5/4/2
Video	Center		ProfessorRosever	Reading Bar Graphs to Compare Means	C	4/3/2
Video	Dotplots		Adam Gonzalez		P	5/3/2

Video	Dotplots		Educator		P	5/4/2
	Dotplots, Stem-and- Leaf		ssurinal		P	5/2/2
Video	Histograms		Profbbob		PC	5/4/3
Video	Histograms		Statslectures		P	5/4/2
Video	Histograms		Prof. Rosever		P	4/2.5/1.5
Video	Histograms Histograms, Stem-and- Leaf		EscellsFun		C	5/3.5/2.5
Video	Leaf		Statslectures			5/4/2
Video	Pictograms		Khan Academy	Reading Pictographs	P	5/5/3
Video	Pie Charts		Jinchstats		P	5/2/2
Video	Piecharts		Prof. Rosever	Interpreting Pie Charts	P	4/2/2
Video	Piecharts		Khan Academy	Reading Pie Graphs and Circle Graphs	P	5/5/3
		Election, Age Distribution Valuing Houses				
Video	Piecharts		Udacity		C	5/5/4
Video	Scatterplots Stem-and- Leaf		Udacity		PC	5/4.5/4
Video	Leaf		profrobbob		PC	4/4/4
Video	Stem-and- Leaf		Prof. Rosever	Interpreting Stem-and-Leaf Displays	P	4/2/2
Video	Stem-and- Leaf		Prof. Rosever	Back-to-back Stem-and-Leaf	P	4/2/2
Video	Stem-and- Leaf		VCEfurthermaths	Back-to-back Stem-and-Leaf	P	5/3/2
Video	Stem-and- Leaf	NFL	Educator		P	5/4/2
Video	Stem-and- Leaf	Basketball	Khan Academy	Stem-and-Leaf Plots	P	5/5/3

Video	Stem-and-Leaf	MrRaup		P	2.5/2.5/2
Video	Stem-and-Leaf	Yourteachers.com		P	4.5/3/1
Video	Stem-and-Leaf	EHow		P	4/3/2.5
Video		Perdiscotv	Presenting categorical and numerical data	PC	5/3/3
			Hans Rosling discuss the difficulty in measuring progress in Human Rights in the form of comparable numerical statistics. He also shows the surprisingly weak correlation between existing estimates for democracy and socio-economic progress.		
Video	Human Rights Global	Gapminder		I	-/3.5/4
Video	Development	Gapminder	Hans Rosling answers Reddit users questions	I	-/3.5/4
Video	Health	Hans Rosling	Breast Cancer Statistics	I	-/4.5/4.5
Video	Health	Hans Rosling	Swine Flu Alert	I	-/4/4
Video	Politics	Nate Silver	Nate Silver answers to controversial questions about race in politics	I	-/4/4
Video	Health	Hans Rosling	Child Mortality	I	-/5/5
Video	Health	Hans Rosling	Religion/Babies	I	-/5/5
			Hans Rosling uses images and animated Gapminder charts that present global economic growth since 1858 to predict when China and India will catch up with the United States in terms of income per person.		
Video	Global Development	Gapminder		I	-/5/5
			Hans Rosling uses Trendalyzer software to show that we can make the impossible possible in terms of national health and wealth		
Video	Global Development	Gapminder		I	-/5/5
			At the US State Department Hans, Rosling shows overall global trends in health and income over the last 200 years, the development of the HIV/AIDS-epidemic and how China is catching up on the richest countries.		
Video	Global Development	Gapminder		I	-/5/5

Video	Global Development	Gapminder	Hans Rosling about low and middle income countries that, with economic and health progress, are catching up with high income countries – countries we used to call the western world.	I	-/5/5
Video	Health	TED		I	-/5/5
Video	Health	TED	Hans Rosling unveils new data visuals that untangle the complex risk factors of HIV and argues that preventing transmissions -- not drug treatments -- is the key to ending the epidemic.	I	-/5/5
Video	Global Development, History, Health,	TED	Hans Rosling offers students ideas as to how to choose appropriate descriptive methods in various data analysis situations, and how to use it to make predictions and discuss correlations by telling the story of the world in 200 countries, 200 years, using 120,000 numbers with augmented reality animation.	I	
Video		jlinchstats	How To Describe Graphs	C	5/4/2
Video			The Beauty of Data Visualization	C	-/4/4
Video	Global Health	Hans Rosling	“The Best Stats You Have Ever Seen”	C	-/5/5
Video	GPA	Udacity		C	4.5/4/3.5
Video	Meteorology, Health	Against All Odds			

Correlation/Regression:

<u>Medium</u>	<u>Statistical Topic</u>	<u>Source</u>	<u>Title</u>	<u>PCI</u>	<u>Rating (S/P/E)</u>
Video	Calculating a Regression Line	StatisticsFun		P	5/5/3
Video	Calculating a Regression Line	Khan Academy	Squared Error of Regression Line - Overview		
Video	Calculating a Regression Line	Khan Academy	Squared error proof 1/4	P	55/44/33
Video	Calculating a Regression Line	Khan Academy	Squared error proof 2/4	P	55/54/33
Video	Calculating a Regression Line	Khan Academy	Squared error proof 3/4	P	55/54/33
Video	Calculating a Regression Line	Khan Academy	Squared error proof 4/4	P	55/54/33
Video	R-squared	Khan Academy	Calculating R-squared	P	55/54/33
Video	R-squared	Khan Academy	R-squared or Coefficient of Determination	P	55/54/33
Video	Regression Line Example	Khan Academy		P	55/54/33
Video	Regression Line Example 2	Khan Academy		P	55/54/33
Video	Correlation	OnlineStats	Demo - guessing correlations		
Video	Regression	OnlineStats	Overview of all Regression Concepts		
Video	Intro to Linear Regression	OnlineStats			
Video	Inference for Regression	OnlineStats			