

1.1 + 1.2 Analyzing Categorical + Quantitative Data

Statistics: the science ^{of} data

data analysis: organizing, displaying, summarizing, and asking questions about data

individuals: objects described by a set of data.
people = subjects
other things = units

Variable: characteristic of an individual. There can be many for each individual.

Categorical variables: place an individual into one of several groups or categories

quantitative variables: takes numerical values for which it makes sense to find an average.

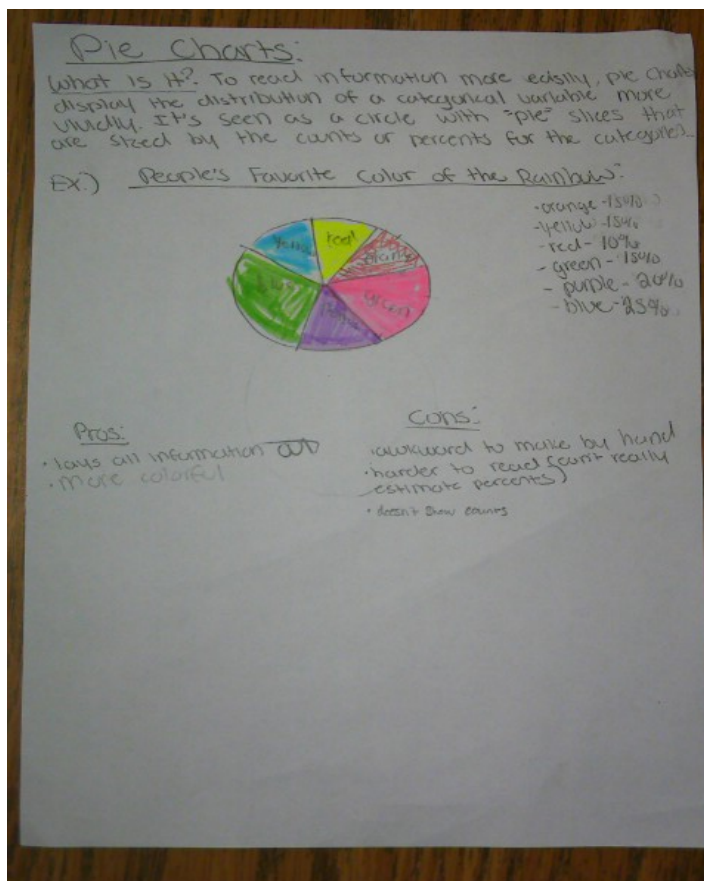
the distribution of a variable tells us what values the variable takes and how often it takes them.


Frequency table : table that displays all values of a variable and the count (frequency) of each variable.

Ex: Birthday month of 5th hour AP Stats

month	Tally	count	relative frequencies
Jan.		0	0%
Feb.		3	10%
March		3	10%
April	-	1	3%
May		4	16%
June		3	10%
July		0	0%
August		0	0%
Sept.		3	10%
Oct.		4	16%
Nov.		4	13%
Dec.		3	10%

Relative Frequency Table : table that use percents (relative frequencies) instead of counts.

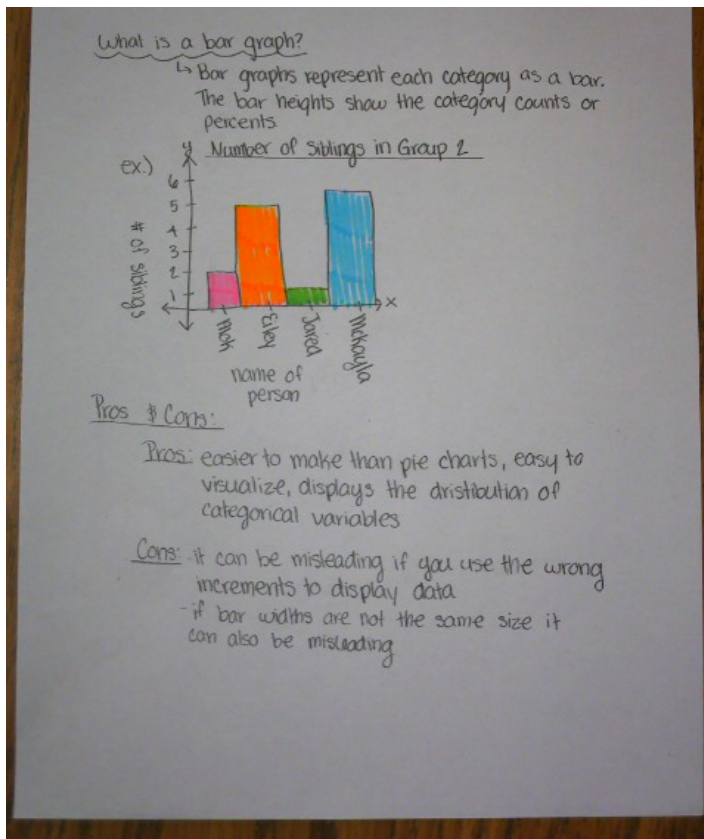


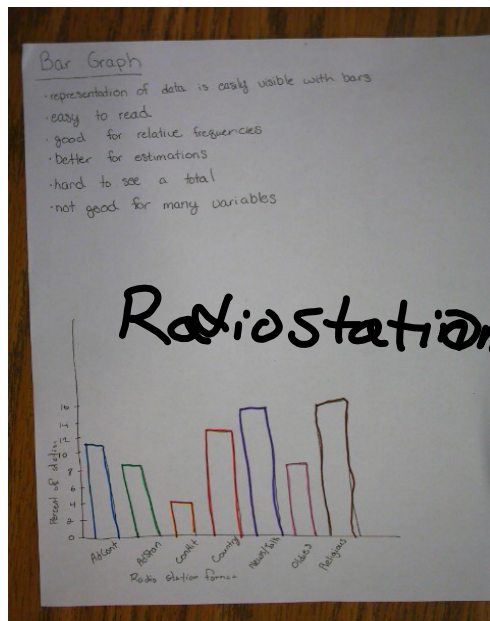


Pie Charts

Pros	Cons
<ul style="list-style-type: none">- Visualize Percentages- Easy to compare multiple graphs- Easy to read- Highly convincing	<ul style="list-style-type: none">- Difficult to measure properly- Difficult to find exact percentages- Difficult to show large sets of data- It can be misleading

Definition: Pie charts show distribution of a numerical value as a "pie" whose slices





Favorite Sports

example:

(Favorite Sport)	(gender)		Total
	male	female	
football	96	78	174
basketball	426	286	712
soccer	696	720	1416
volleyball	663	758	1421
tennis	486	597	1083
<u>total</u>	<u>2367</u>	<u>2459</u>	<u>4826</u>

row variable

column variable

pros: allows you to compare 2 different variables

cons: does not make good visual for data
not easy to find percentage value of data

marginal distribution →

total of a row variable over total of all individuals

Conditional distribution → total of 1 column variable over total individuals in that column

Two-Way Table:

↳ It describes two categorical variables. The distribution of a categorical variable will tell how often it occurs. The totals of the variables is also shown.

Example:

Car Usage between SUV's & Sports cars, between gender

	Sport Utility Vehicle (SUV)	Sports Car	Totals
male	21	39	60
female	135	45	180
totals	156	84	240

Pro's:

- Easy to manipulate
- Easily comparable
- More accurate in calculations

Con's:

- How long it takes to calculate

This shows a marginal distribution; it is the distributional value of a variable among all individuals, which are described by the table.

This also shows conditional distribution; the values of a variable among individuals who have a specific value of another variable.

Example:

- ↳ SUV Users: $\frac{156}{240} = 65\%$
- ↳ Sports Car Users: $\frac{84}{240} = 35\%$

Example:

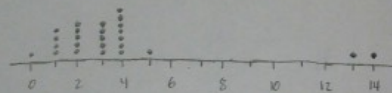
- ↳ For Females
- ↳ SUV Users: $\frac{135}{180} = 75\%$
- ↳ Sports Car users: $\frac{45}{180} = 25\%$

1. What is it?

A number line, starting with the minimum value and ending with the maximum value, where every piece of data is represented as a dot above its respective number value

2.

U.S. Women's Soccer Goals



3. Pros:

- Easily read/interpreted
- Easy to construct
- Works well with many types of data

Cons:

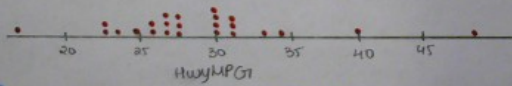
- Outliers can make the spread very large
- Would not work large with large samples of data

DOTPLOT

What is it?
 Simple graph that shows each data value as a dot above its location on a number line.

Example

Model	Miles	Model	Miles	Model	Miles
Acura RL	24	Dodge Avenger	30	Nexus	30
Audi A8	28	Ford Fusion	25	benz E350	30
Bentley Mullanne	18	Hyundai Elantra	40	Mitsubishi Galant	30
BMW 550i	23	Jaguar XF	23	Nissan Maxima	26
Buick LaCrosse	27	Kia Optima	34	Saab 9-5 sedan	28
Cadillac CTS	27	Lexus ES 350	28	Subaru Legacy	31
Chevrolet Malibu	33	Lincoln MKZ	27	Toyota Prius	48
Chrysler 200	30	Mazda 6	31	Volkswagen Passat	31
				Volvo S80	26



Pros

- Shows each individual data point
- Simple to make with small set of data

Cons

- Can be time consuming w/ lots of data points to make
- Have to count to get the total
- Hard to interpret

Stemplot: shows shape of distribution while including the actual number values in the graph

Stem: All leading numbers (can be 0)

Leaves: the last number

Pros: easy to make and use, easy to understand, useful for small data sizes, can use either quantitative or categorical variables

Cons: gets messy with large data sizes,

Ex: Pairs of shoes teenage girls have:

50 26 26 31 57 19 24 22 23 38
13 50 15 34 23 30 49 13 15 51

Pairs of shoes	
1	3359
2	233466
3	0148
4	9
5	0017
Stems	Leaves

Stemplot

Pairs of shoes	
1	333
2	59
2	2334
2	66
3	014
3	8
4	
4	9
5	001
5	7
Stems	Leaves

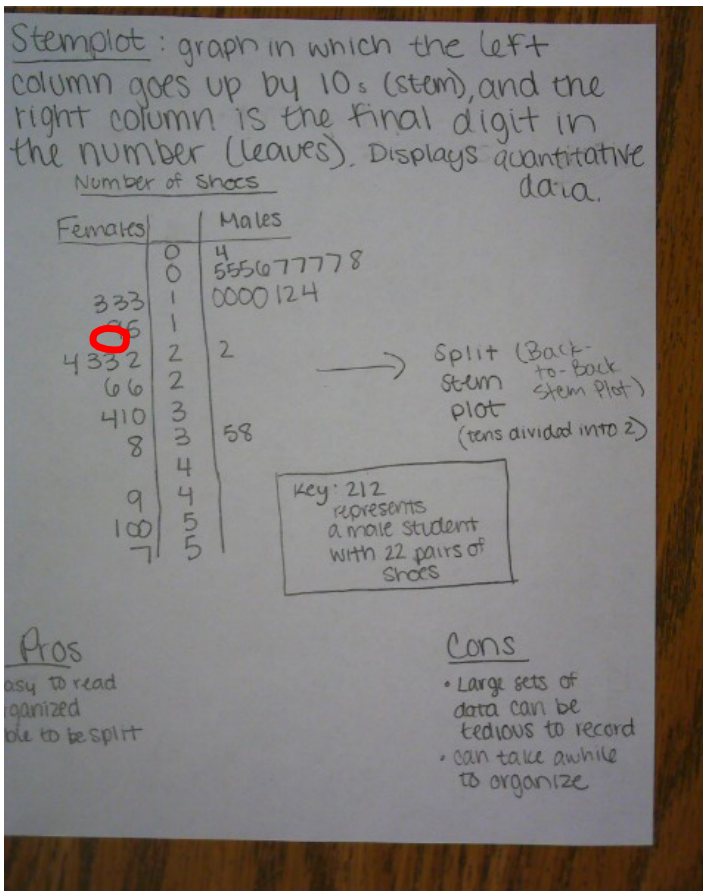
"Splitting Stems" plot

pairs of shoes teenage boys have:

14 7 6 5 12 38 8 7 10 10
10 11 4 5 22 7 5 10 35 7

Females		Males
	0	4
	0	555677779
333	1	0000124
95	1	
4332	2	2
66	2	
420	3	
8	3	58
9	4	
100	4	
7	5	

back-to-back stemplot



Therefore, 5 will only fit into one group, not two.

Step 2: find out the count frequency or percent (percent only if you're making a relative frequency table)

Frequency table		relative frequency table	
class	count	class	percent
0 to 5	20	0 to 5	41
5 to 10	13	5 to 10	27
10 to 15	9	10 to 15	18
15 to 20	5	15 to 20	10
20 to 25	2	20 to 25	4
total	49	total	100

Step 3: label and scale axis and draw Histogram

each bar represents a class
 & the bar height is the class frequency or relative frequency
 & draw the bars w/ no horizontal space between them unless a class is empty

Pros:

- great for larger class ranges
- use w/ quantitative data.

Cons:

- not as efficient w/ smaller class ranges
- can't use categorical variables.

what is it?

A graph of the distribution of a quantitative variable made by plotting the counts or percents of values in equal-width classes

Example: Hours spent on phone by age group

Age Group	Hours Spent (approx.)
0-10	0.5
10-20	5.5
20-30	4.5
30-40	3.5
40-50	2.5
50-60	1.5
60-70	1.0
70-80	0.5
80-90	0.2
90-100	0.1

Pros	cons
<ul style="list-style-type: none"> - easy to compare categories - can do on calculator - relate to dot plots for easy comparison - can use percents and counts 	<ul style="list-style-type: none"> - not as easily as a bar graph, but also confused often - not very specific (not specific data) - cannot use categorical data

* Bar graph is categorical and Histogram is quantitative data & connected *
 * if (for example) there is a number like 5, which falls in both 0-5 and 5-10, then it is really 0 to <5, then 5 to <10
 * quantitative values *

Shape → symmetry, skewness, mode
 unimodal → one peak
 bimodal → two peaks
 trimodal → three peaks
 multimodal → more than 2 peaks

Outliers → a value that falls outside the overall pattern

Center → the midpoint/median of all the values

Spread → how the data varies

states with foreign born residents

think of ski hill which way would you ski = which way it's skewed

- skewed right, unimodal, and long tail right
- no outliers
- midpoint from 10-15%
- data is 0% to 30%

Amount of shoes

0	14
0	555077
1	0000124
1	19
2	2
2	2
3	58

KEY: 212 represents a male student with 22 pairs of shoes

Shape
Outliers
Center
Spread

stemplots:

0	5	8		
1	2	6	8	
2	2	5	5	9
3	3	4	7	
4	0	2		

Amount of Candy eaten by second graders
Key: 1 = 1 child

* Shape: symmetrical, unimodal at the 20's

* Outliers: none because there is no large jump in between numbers ex. 100

* Center: the center is 25

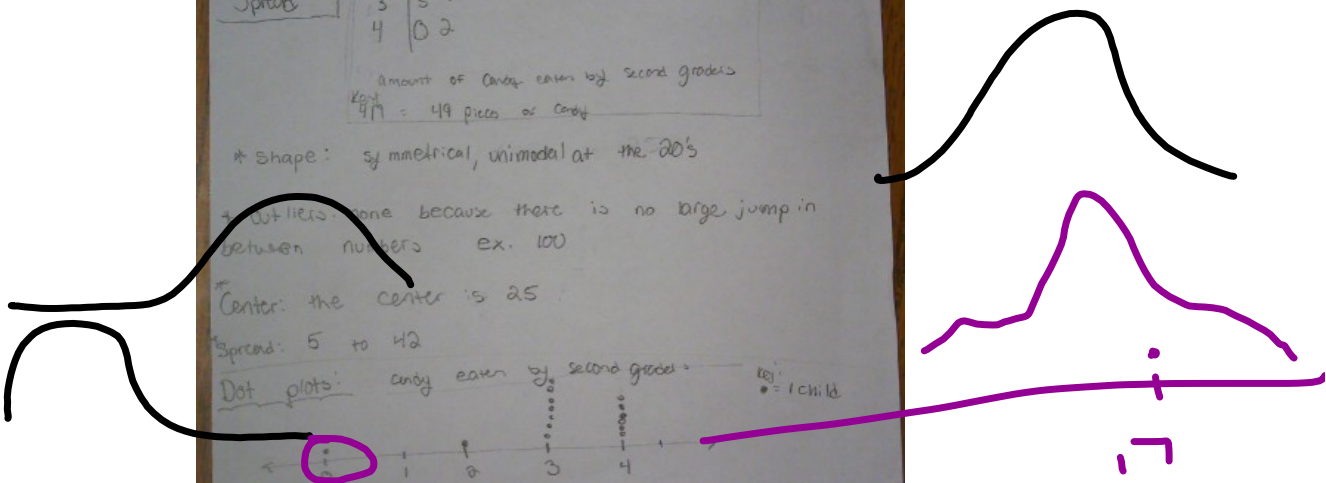
* Spread: 5 to 42

Dot plots: candy eaten by second graders

* Shape: skewed to the left, mode is 3, only one mode

* outliers: 0

* Center: the center is 3



Histograms

* Shape: skewed to right
unimodal at 2

* outliers: none

* Center: center at 2

* Spread: 5 to 25

Amount of candy eaten by 1st graders