STEM Stats v1 - Day 1

Statistics is a way of reasoning, along with a collection of tools and methods, designed to help us understand the world.

Statistics is about **variation.**

Data are values along with their context. Context is provided by the "W"s

Why do we care about the data? (What are you trying to figure out?)

Who are described by the data? (Which individuals or things?)

What variables do the data contain? (What questions were asked?)

When, Where, How } was the data collected?

For histograms:
# of bars = $\sqrt{\#\text{ observations}}$
bar width = $\frac{\max - \min}{\#\text{ of bars}}$

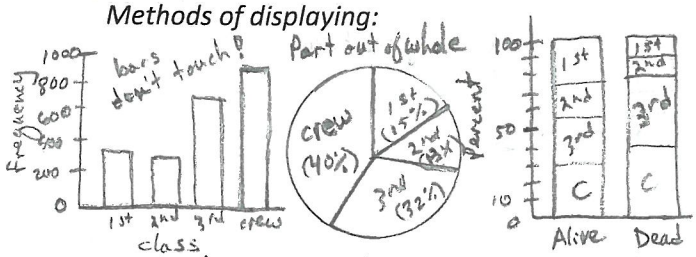
(Necessary)

Data is either:

Categorical

counts or **%s** of individuals into groups or categories.

Methods of displaying:



bar pie segmented bar

Numerical Summaries:

Contingency tables - categorize the individuals on all variables at once to reveal possible patterns.

Class

Survival	First	Second	Third	Crew	Total
Alive	203	118	178	212	711
Dead	122	167	528	673	1490
Total	325	285	706	885	2201

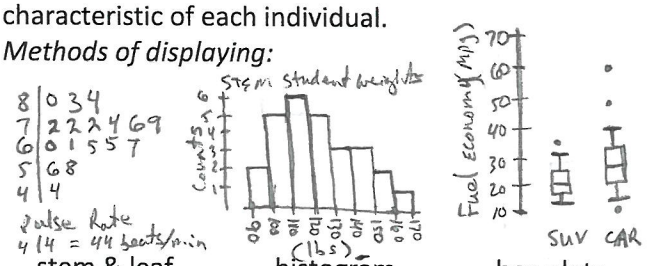
		First	Second	Third	Crew	Total
Alive	Count	203	118	178	212	711
	% of Row	28.6%	16.6%		29.8%	100%
	% of Col.	62.5%	41.4%		24.0%	32.3%
	% of Table	9.2%	5.4%		9.6%	32.3%
Dead	Count	122	167	528	673	1490
	% of Row		11.2%		45.2%	
	% of Col.		58.6%		76.0%	
	% of Table		7.6%		30.6%	
Total	Count	325	285	706	885	2201
	% of Row	14.8%		32.1%		100%
	% of Col.	100%		100%		
	% of Table	14.8%		32.1%		

If rows (columns) of contingency table have equal distributions, then variables are **independent** o.w. they are **associated**. (see segmented bar)

Quantitative

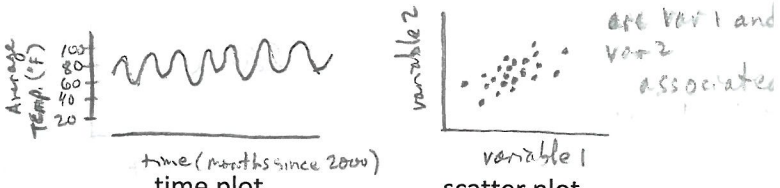
numerical values with units that measure some characteristic of each individual.

Methods of displaying:



stem & leaf histogram box plots

describe distribution: shape, center, spread, outliers



time plot scatter plot

Numerical Summaries:

5 number summary → Max, Q3, Median, Q1, Min

If shape is **skewed:**

spread = IQR (Q3 - Q1), center = Median

outlier if > Q3 + (1.5)(IQR)

or < Q1 - (1.5)(IQR)

If shape is **unimodal and symmetric:**

center = **mean**, spread = **standard deviation**

enables comparison of apples to oranges.

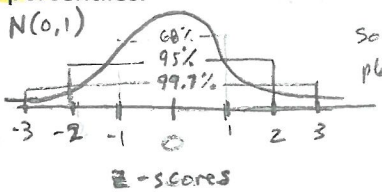
Matthew is 67" tall and Eileen is 65"

$\bar{x}_b = 68"$ $s_b = 3"$ $\bar{x}_g = 66.7"$ $s_g = 3.2"$

z-score = $\frac{67 - 68}{3}$ = $-.\overline{3}$

z-score = $\frac{65 - 66.7}{3.2}$ = $-.5$

Using the standard deviation as a ruler we create z-scores that tell us how unusual a value is, in SDs from its mean. Then we use models (mathematical curves) to change z-scores to **percentiles.**



So 2 SDs above the mean places you in the ___ percentile.