KEY CONCEPT: POPULATION VS.SAMPLE

Population: An entire group of people or objects of interest
Sample: A subset of the population. Hopefully representative of the population.

Samples are used because we can't get the entire population, or it is too expensive to get.

Examples:

- Population: All the data analysts in the world. Sample: all data analysts in Ohio
- Population: All Major League Baseball players. Sample: Cleveland Guardians players

  - Think of some examples yourself and tell us about them

ARE THESE POPULATIONS OR SAMPLES?

1. A teacher gives an exam to their pupils. The teacher wants to summarize the results of the exam.

   Population. The teacher is only interested in these specific pupils' scores and nobody else.

2. A researcher has recruited males aged 45 to 65 old for an exercise study to investigate risk markers for heart disease (e.g., cholesterol).

   Sample. A researcher investigating health related issues will not simply be concerned with only the participants of their study; they will want to show how their sample results can be generalized to the whole population (in this case, all males aged 45 to 65 years old).

3. A census is taken every 10 years in the US to gather information on demographics, etc.

   A census is close to a population but misses some people because it is difficult to find them.

KEY CONCEPT: SAMPLE BIAS

Sample Bias is a systematic misrepresentation of the population data and is to be avoided.

Examples:

- Using temperatures in June to represent the entire year.
- 1948 Presidential election prediction based on a phone survey when many people did not have a phone
- Statistical theory tells us survey answers are plus/minus 3% if there are 1100 respondents
  - Then why are election surveys so bad at predicting the winner? Answer: the answers to the surveys are from a biased set of respondents.
- Think of some examples and share them.

EXTREME BIAS – SURVIVORSHIP BIAS

A famous case of "survivorship bias" happened during WWII. American bombers were suffering significant losses during missions over Germany.

The Air Force was deciding where to put more protective armor on the planes. They studied the damaged planes and found that the fuselage (body) had the most bullet holes. So, they decided that was the area that needed to be reinforced.

But a statistician, Abraham Wald, reasoned differently. He thought the sample of planes with fuselage holes were the ones that returned. The ones that didn't return likely had bullet holes in a different place – the engine and wings.

The armor was installed over the engines. That increased the % of bombers that successfully returned.

DESCRIPTIVE STATISTICS

Techniques for summarizing a set of data

in ways that people can easily interpret.

TYPES OF DATA

- **Numerical** variables
  - Example 1: The number of miles on the odometer of a car.
  - Example 2: The age of an insured customer.
- **Categorical** variables are non-numeric.
  - Example 1: The make of a car (Chevy, Tesla, Ford, etc.)
  - Example 2: The age in groupings: (0-19, 20-39, 40-59, 60-79, 80+)
  - Example 3: Survey Scores (5/4/3/2/1). These are often treated as numeric.

MEAN, MEDIAN, AND MODE IN DEPTH

Numeric values use **Mean** and **Median** as measures of **"central tendency".**

For this data, the median is the best measure of the "central tendency" because there are a few atypically large project times that greatly influence the mean.

IT Project Time	
	50
	80
	90
	100
	130
	160
	180
	190
	400
	700
Average	208
Median	145

However, if the goal is to measure the **improvement** in project time and the improvement is due to the large values being reduced, the median may not pick up the improvement. This is particularly true for small to medium sized data sets.

IT Project Time	
	50
	80
	90
	100
	130
	160
	180
	190
	400
	700
Average	208
Median	145

IT Project Time	
	50
	80
	90
	100
	130
	160
	180
	190
	200
	350
Average	153
Median	145

Cut these 2 times in half.

Improvement
-55
0

ANALYST BOOT CAMP

MEAN, MEDIAN, AND MODE IN DEPTH

Categorical variables use the **mode** to describe "**central tendency**".

The mode of the car companies is Audi. The mode of the survey is 2.

Brand
Audi
Audi
Audi
Audi
Audi
Audi
BMW
BMW
Porsche
VW
VW
VW
Mercedes
Mercedes

GuyGo Insurance Co. Survey	
1=Satisfied 2=Neutral 3=Not satisfied	
Indicate your score for:	**Score**
The company	2
Your Business Leader	2
Your Manager	1
Our mascot	3

Strictly speaking, survey answers are categorical, but people often treat them as numbers and compute average scores.

MEAN, MEDIAN, AND MODE IN DEPTH – KHAN

Khan Academy

Statistics Intro: Mean, median, & mode

To access Khan Academy Lessons:
Control-Click the link to access the material everywhere in the presentation.

STANDARD DEVIATION IN DEPTH

The average is a measure of the “central tendency” of a set of data.

The **standard deviation** is a measure of how **compact** or **spread-out** the data is.

There are 2 Excel calculations, depending on if the data is a population or sample

(Population) Standard Deviation = **stdev.p**

Sample Standard Deviation = **stdev.s**

For theoretical reasons**, stdev.s** is generally a little larger than **stdev.p.**

STDEV ( ) uses **stdev.s( )** to be conservative**.**

STANDARD DEVIATION IN DEPTH – KHAN

Khan Academy

Standard Deviation

Sample Standard Deviation

DISPLAYING DATA

Khan Academy

[Displaying & Describing]

ANALYST
BOOT CAMP

OUTLIERS

Outliers are data values that are significantly different from the other values

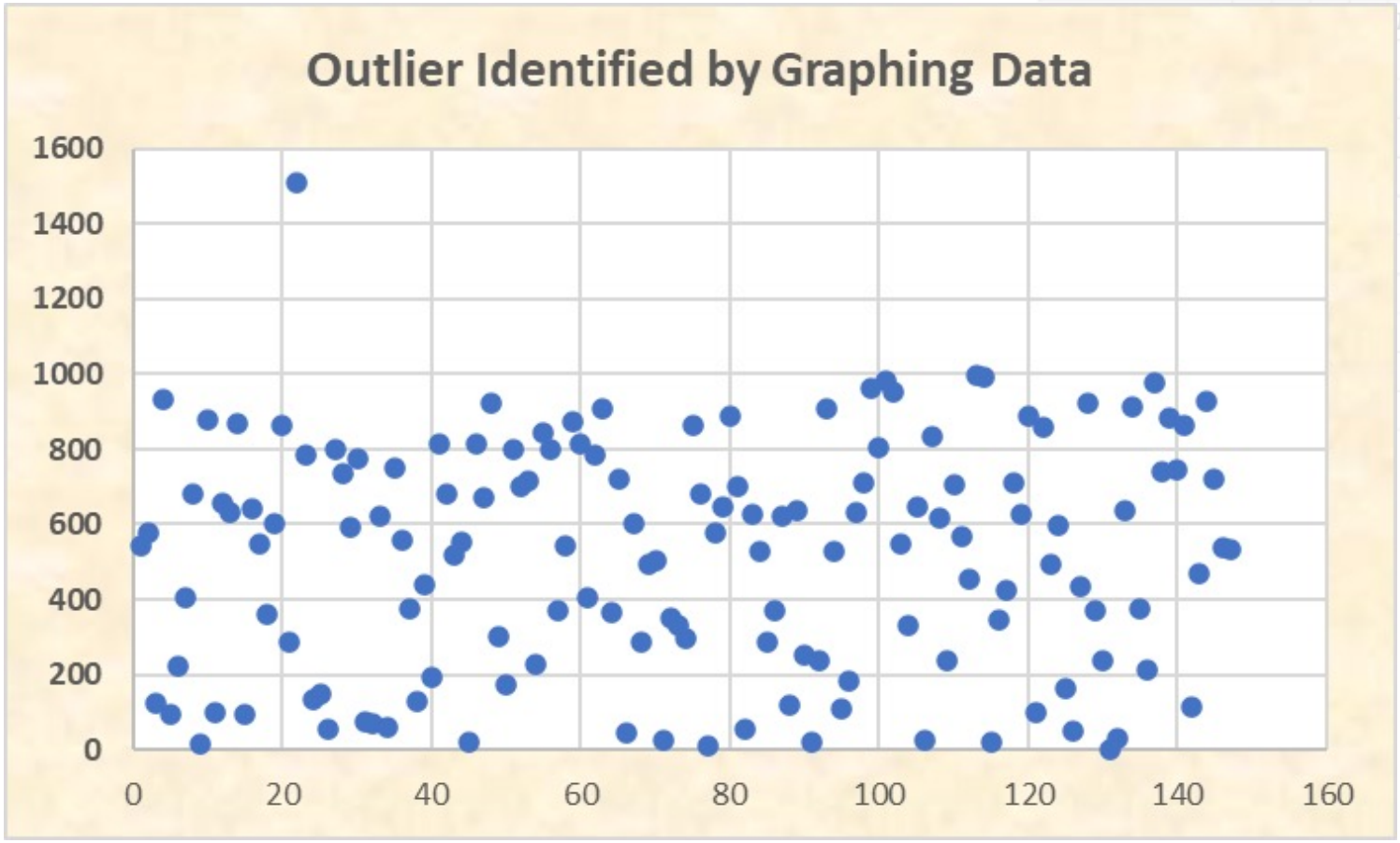
Example: 0, 0, 1, 1, 2,100, 3, 2

- Outliers can dramatically affect the calculations of $\bar{x}$ and s.
  with the outlier, $\bar{x} = (0 + 0 + 1 + 1 + 2 + 100 + 3 + 2) / 8 = 13.6$
  without the outlier, $\bar{x} = (0 + 0 + 1 + 1 + 2 + 3 + 2) / 7 = 1.3$

- You can only remove outliers if you have a valid reason. Example from Claims:
  - 100 could represent the real # of cars involved in a pile up (keep it)
  - 100 could be a "fat fingered" data error (change or remove it).

- PGR Pricing Indications: we remove Large Losses initially, then redistribute them.
- In practice, it is difficult to identify outliers in big datasets. Graphing might help..

OUTLIERS



ANALYST
BOOT CAMP

OUTLIERS– KHAN

Khan Academy

Outliers