What do you mean it's "statistically significant"?

Shivani Patel, Ventura County Star
Briana Rice, Cincinnati Enquirer
Kristi Tanner, Detroit Free Press
Tim Webber, Des Moines Register
Tyler Whetstone, Knoxville News Sentinel

Outline

Statistical terms: the basics

Your data is only as good as its source

Sampling error

Real world application

Statistical terms: the basics

Stuff we should all know

Average: People want to know how healthy/fit/happy/rich people compare to others

- **Mean: It is calculated by dividing the sum of values in a collection by the number of values.**

- Example: 400+500+600=1500. Divide this by 3. Mean=500.
- You should use the mean when you numbers are close together and you want to find the central tendency.

Median: The median separates the higher half of data from the lower half. Great for times when there is an outlier like a person with a high income among a group of low-income individuals

- Median: 300+400+500+600+700. Median is 500
- 300+400+500+600. 400 + 500 = 900; 900 / 2 = 450.
- **Mode: Most recurring value in a data set.**

What is and isn’t a rate?

A rate isn’t just a number. It’s a comparison of two quantities, such as deaths and people. A rate refers to the occurrence of events over a specific interval in time.

If 600 people die, the death rate is not 600.

Example: 100 miles/ 10 hours.

It takes 10 hours to travel 100 miles. Simply: 10 miles per hour.

Terms to avoid:

- Likely: It comes from probability measures. If a community has a 10% infection rate, that does NOT mean they are 10% more likely to get the infection. Keep it simple and just use the percentage.
- Significant: An actual statistical measure, so don't use it even if you think it is socially and journalistically significant. Avoid the word.
- Correlation is also a measure. Do not use correlation to mean two things that are happening at the same time..

Percent change vs. percentage point change

Percent: One part per 100 (50% = 50/100 or ½)

Percentage: Fraction of a whole (To find what percentage 23 of 50 is, divide 23/50 then multiply by 100 to get 46%)

Percent change: Difference of final and initial values, divided by initial value.

Percentage point change: Difference between two percentages

Percent change vs. percentage point change (cont.)

Suppose you have a student loan with an annual interest rate of 4 percent.
One day your lender announces that the interest rate will soon increase to an annual rate of 5 percent.*

Percent change:
(5-4)/4 = 25% increase

Percentage point change:
5-4 = 1 percentage point increase

**Percent change is good for analyzing data over time.

- *context is important* - When writing about school enrollment, for ex., a school with a low population could see a high percentage change even if the change is just a matter of a few students so explain why in your copy.
- keep it simple - Don't list out every number, and instead focus on key finds in your article. Throw the rest into a table or chart.

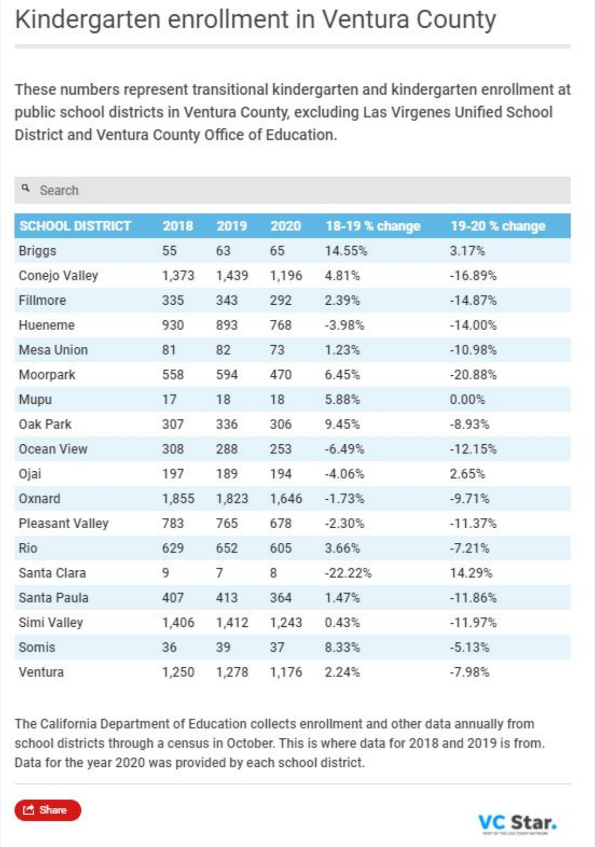
**example courtesy reed college*

(cont.)

Easy websites to use for data visualization: Infogram, Datawrapper, Flourish (Google Journalist Studio)

Check out: https://www.journaliststoolbox.org/2021/06/28/online_journalism/.

"Ojai, Briggs and Santa Clara school districts saw an increase in enrollment. However, because Briggs and Santa Clara are two of the smallest districts in the county, an increase or decrease of a student or two in those districts can translate to a high percentage."



Kindergarten enrollment in Ventura County

These numbers represent transitional kindergarten and kindergarten enrollment at public school districts in Ventura County, excluding Las Virgenes Unified School District and Ventura County Office of Education.

SCHOOL DISTRICT	2018	2019	2020	18-19 % change	19-20 % change
Briggs	55	63	65	14.55%	3.17%
Conejo Valley	1,373	1,439	1,196	4.81%	-16.89%
Fillmore	335	343	292	2.39%	-14.87%
Hueneme	930	893	768	-3.98%	-14.00%
Mesa Union	81	82	73	1.23%	-10.98%
Moorpark	558	594	470	6.45%	-20.88%
Mupu	17	18	18	5.88%	0.00%
Oak Park	307	336	306	9.45%	-8.93%
Ocean View	308	288	253	-6.49%	-12.15%
Ojai	197	189	194	-4.06%	2.65%
Oxnard	1,855	1,823	1,646	-1.73%	-9.71%
Pleasant Valley	783	765	678	-2.30%	-11.37%
Rio	629	652	605	3.66%	-7.21%
Santa Clara	9	7	8	-22.22%	14.29%
Santa Paula	407	413	364	1.47%	-11.86%
Simi Valley	1,406	1,412	1,243	0.43%	-11.97%
Somis	36	39	37	8.33%	-5.13%
Ventura	1,250	1,278	1,176	2.24%	-7.98%

The California Department of Education collects enrollment and other data annually from school districts through a census in October. This is where data for 2018 and 2019 is from. Data for the year 2020 was provided by each school district.

Your data is only as good as its source

Know what your data can and can’t do

- Ask yourself what the data is and isn’t able to tell you
- Is the data regularly updated? When was it last updated?
- Does your dataset capture all of the relevant data, or might it be missing some?

Understand how your data is defined

- Who put the data together and why?
    - Look at the methodology
    - Don’t take numbers directly from press releases — check them yourself!
- How specific is your data?
- How are terms defined within your data? Don’t assume each source classifies things in the same way

Verify that your data is accurate

- Double-check the math!
- If you’re using multiple sources, make sure they match
- Make sure you’re working with the correct version of the data

Sampling Error

Population vs. Samples

The entire universe of possible data is called the ***population***. *A* subset of the population is called the ***sample***. There are many different types of samples.

A **probability sample** incorporates randomness. Examples include:

➔ **Simple random samples**: Each element has the same chance of selection.

➔ **Stratified samples**: Divide the population into groups of some sort (gender, race, income, many others) and sample from each stratum.

Q. Why is Probability Important?

A. ***Representativeness***: A sample must contain essentially the same variations that exist in the population. This is usually achieved if all members of the population have an equal chance of being selected in the sample.

A ***nonprobability sample*** discounts randomness and is usually not representative of the larger population of interest.

Confidence Intervals

Confidence intervals use the standard error calculation and your choice of confidence level to define how sure we are the population parameter falls within a specific range.

Or X% of the time the confidence interval will contain the true value of what we are trying to estimate.
Confidence interval: Xbar +/- s.e. *t

An example:

➔ **90%** :Xbar +/- s.e. * 1.64
➔ **95%** : Xbar +/- s.e. * 1.96
➔ **99%** : Xbar +/- s.e. * 2.58

Example: Statistical significance and Census data

- ➔ American Community Survey

- ➔ U.S. Census Bureau. (2020, February). *Understanding and Using American Community Survey Data: What Journalists Need to Know.*

- ➔ U.S. Census Bureau. Statistical Testing Tool

Other questions to ask

- What was the sampling strategy?
- Is it a probability sample?
  - Sampling error: What is the margin of error? sample size? confidence level?
  - Weighting: Were the results weighted? How?
- Are there other forms of bias? E.g. question wording, question order and interviewer effects.
- When were the interviews conducted?

American Association for Public Opinion Research (AAPOR). (2021). Education/Resources for Media.

Real world applications

Real world example (budget season)

- Numbers matter | People matter more
- Knox County cut their indigent care budget 33%, $4.5 million - $2.9 million
- This came out through the budgeting process, where you can see year over year change
    - Find it first!
- The numbers are boring. Explain why it matters
    - This program specifically helps the poor and vulnerable and the county cuts it

- Tyler Whetstone

Questions?