Analyzing Data

A few favorites

percent change

$$\frac{\text{NEW - OLD}}{\text{OLD}}$$

If teacher salaries were $31,500 in 2017 and $32,000 in 2018, we can say:

- Teacher salaries increased by $500.
  - 32000-31500 = 500
- Teacher salaries increased by 1.6 PERCENT.
  - (32000-31500)/31500 = 0.01587

It works the same way with a decrease. If teacher salaries were $31,500 in 2017 and $30,000 in 2018, we can say:

- Teacher salaries decreased by $1,500.
  - 30000-31500 = -1500
- Teacher salaries decreased by 4.8 PERCENT.
  - (30000-31500)/31500 = -0.04762

percent change of a percent

$$\frac{\text{NEW - OLD}}{\text{OLD}}$$

What if we're dealing with changes to something that's ALREADY measured as a percentage?

If 25% of teachers had a masters degree in 2017 and 30% had a masters degree in 2018, we can say:

- The share of teachers with a masters degree increased by 5 PERCENTAGE POINTS.
  - 30-25 = 5
- The share of teachers with a masters degree increased by 20 PERCENT.
  - (30-25)/25 = 0.20

Or, if we're in a decrease situation: If 25% of teachers had a masters degree in 2017 and 18% had a masters degree in 2018, we can say:

- The share of teachers with a masters degree decreased by 7 PERCENTAGE POINTS.
  - 18-25 = -7
- The share of teachers with a masters degree decreased by 28 PERCENT.
  - (18-25)/25 = -0.28

per-capita

How many murders were there in New York City versus Austin, Texas?

To get a reasonable comparison, be sure to account for how many people live in each place!

city	homicide_rate_2017	population_2017	homicides_per_capita	
New York City	290	8,622,698	3.4	per 100,000
Austin, Texas	29	931,830	3.1	per 100,000
Detroit	267	672,795	39.7	per 100,000

** numbers not fact-checked!*

choosing your denominator wisely

How should we measure participation in an election in a particular county?

Some options:

- votes cast / registered voters in the county
- votes cast / eligible voters in the county
- votes cast / U.S. citizens who are at least 18 yrs old
- votes cast / people who live in the county

There’s not necessarily a RIGHT answer. You’re answering a different question with each option.

Types of variables

- Binomial (Yes/No; True/False)
- Categorial (white/Black/Hispanic/Asian)
  - Ordinal (elementary/middle/high school)
- Numeric
  - Continuous (e.g. temperature)
  - Bounded continuous (e.g. age)
  - Discrete (e.g. money in your bank account, employees in your newsroom)

Analyzing one variable

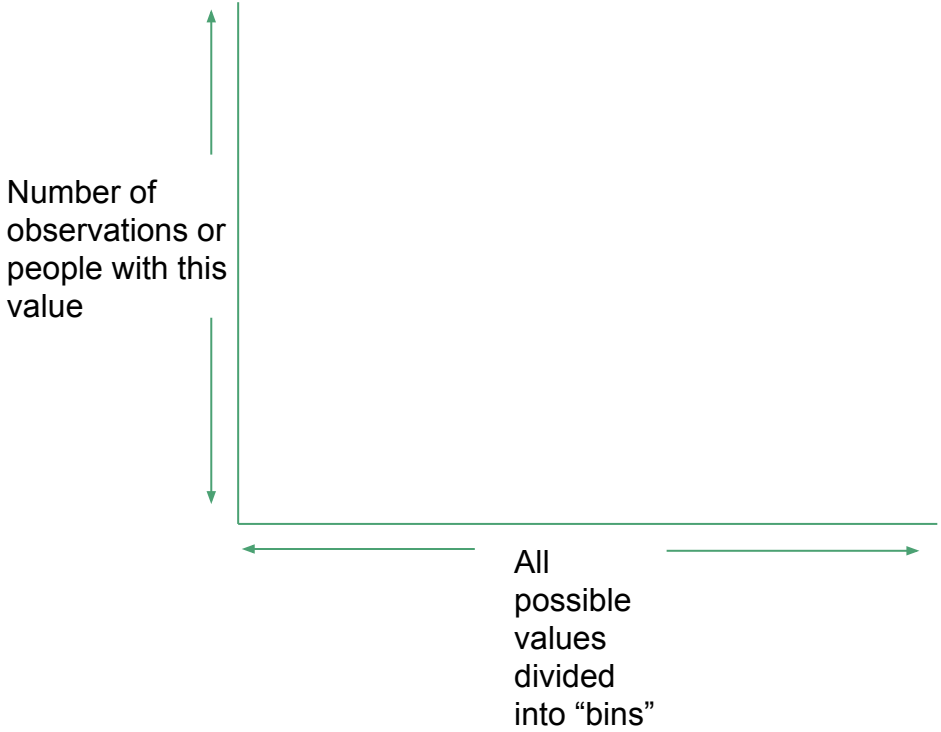
First, a useful tool: distributions
What is a distribution?

It reflects how points are spread out across the range of data.

The range is all the values between the minimum and maximum values in our column.

A good way to see a distribution is to make a histogram.

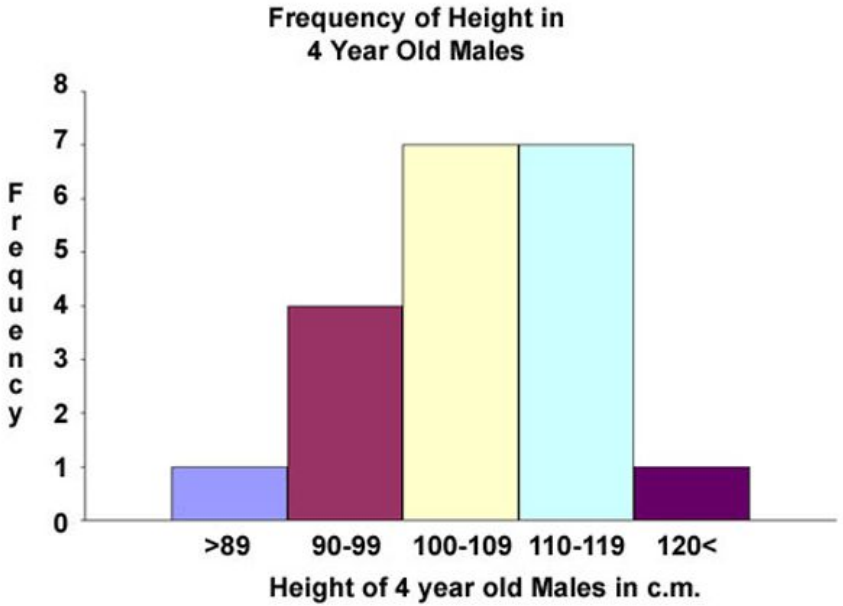
What is a histogram?



A histogram tells us how common each range in the data is. This is called the ‘distribution’ of the data.

What is a histogram?

Here’s an example histogram.



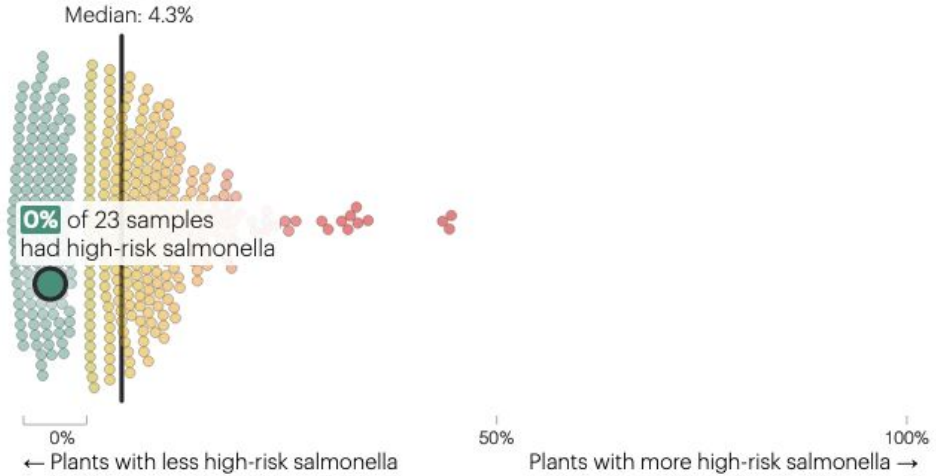
Modesto Food Distributors, Colma, C...

P4985 • Medium plant • *Data from Aug. 2020 to Aug. 2021*

● **Chicken Parts**

How This Plant Compares

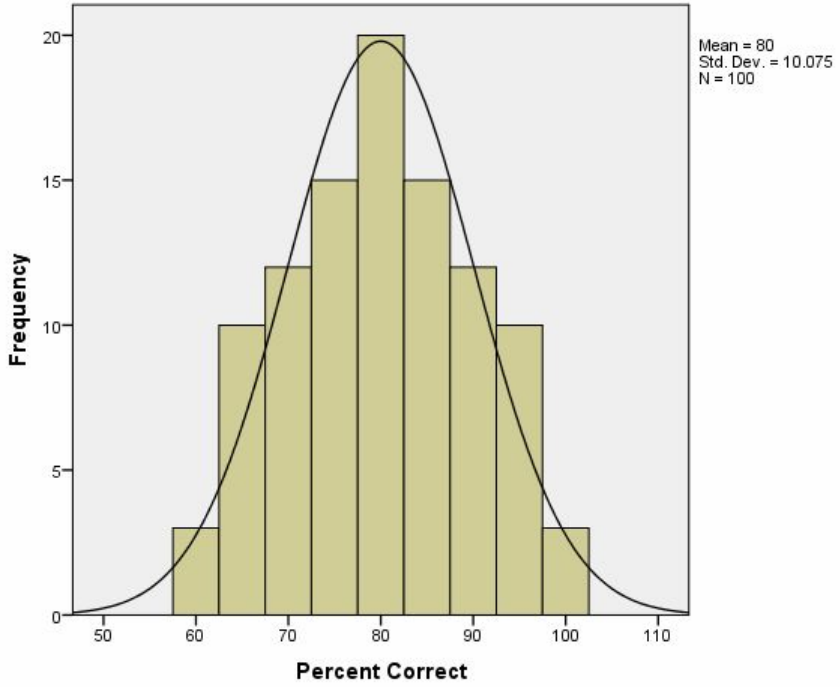
No high-risk salmonella was found on the chicken parts at this Modesto Food Distributors plant.



Here are some common distribution shapes

Normal

Average is an appropriate summary for this column.



Here are some common distribution shapes

Normal with skew

The long right or left tail can move the mean to a value that isn't typical.

You should consider median or mode here.

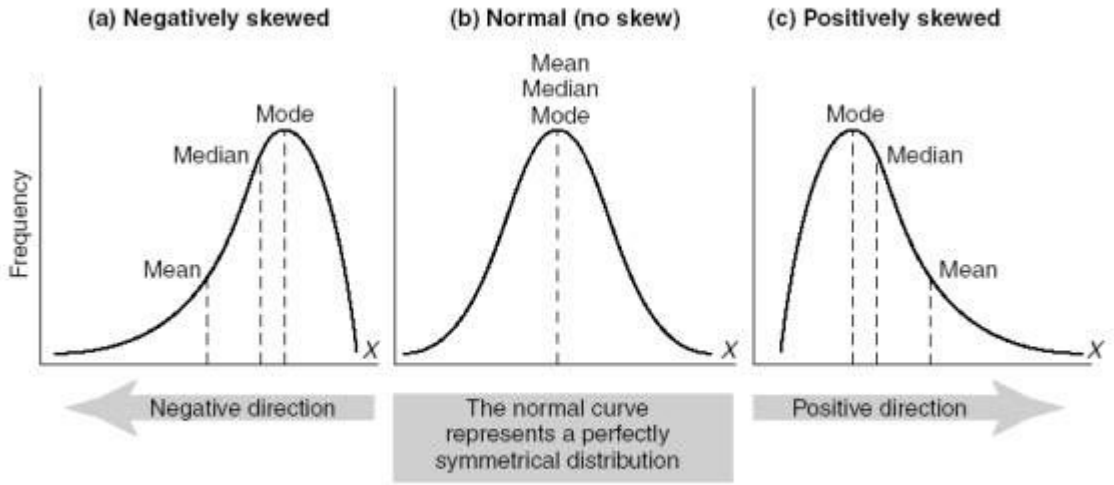
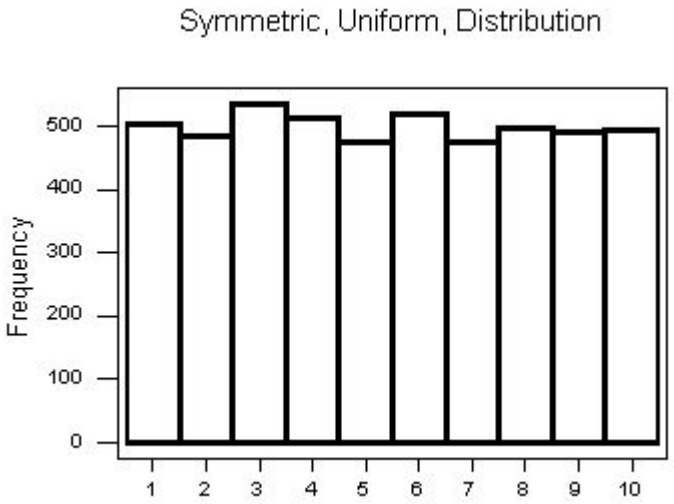


FIGURE 15.6 Examples of normal and skewed distributions

Here are some common distribution shapes

Uniform

The mean here is 5. Does that accurately reflect reality?

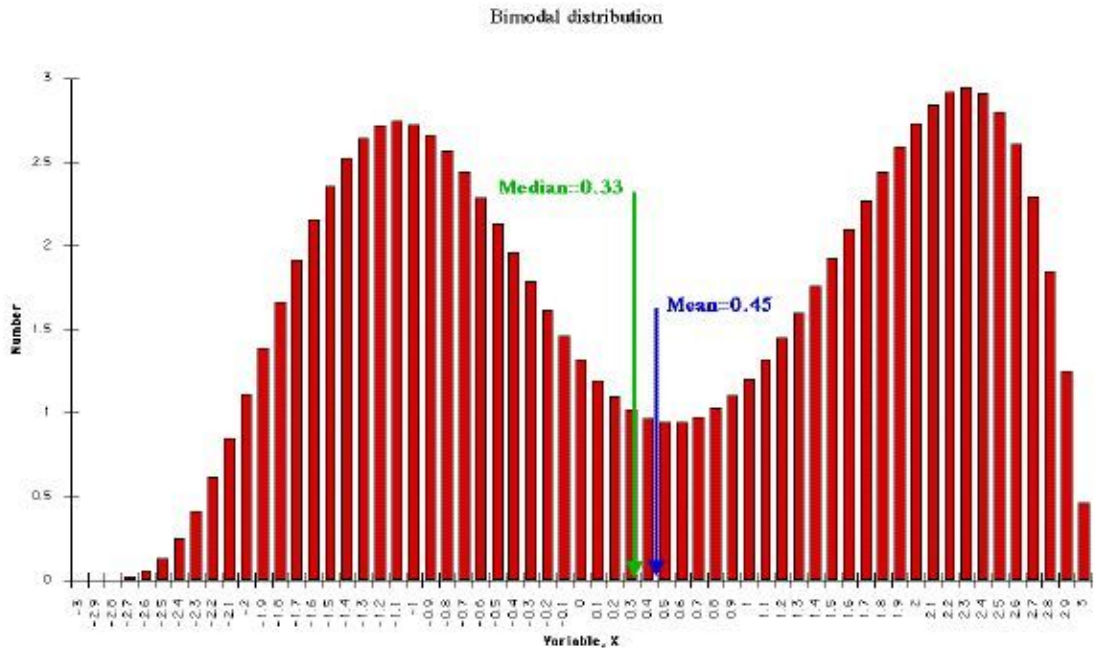


Here are some common distribution shapes

Bimodal

We see this often when we are talking about poverty or race.

Do the mean and median here accurately reflect reality?



Now what?

- Summarizing one variable
  - Average
  - Median
  - Sum
  - Maximum
  - Minimum
- Looking for an outlier in a distribution

How do I know which summary statistic(s) to use?

Well, it depends on the shape of the distribution.

Some vocab:

Mean/Average - This is the sum of all values divided by the number of observations.

Median - If we rank everyone in the data by value, this is the value associated with the person (or people) in the middle.

Mode - This is the value that occurs most frequently in the data.

Analyzing two variables

Two vars, one numeric one categorial

E.g. Was the amount owed different depending on a cases' status? (e.g. had default judgements vs were dismissed vs were pending)

You can answer this by calculating a summary statistic for the continuous variable (amount owed) within each group (case status).

Pivot tables can do this for you!

Two vars, both numeric

Types of questions:

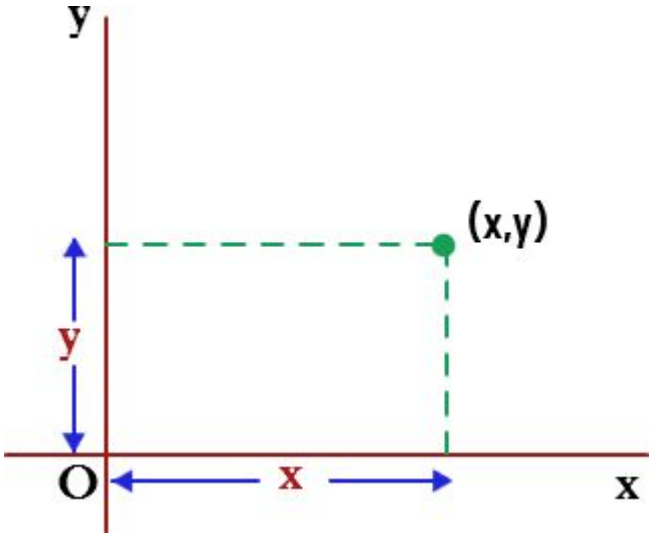
As the rate of poor students at a school increases, what happens to test scores?

Is there a relationship between student teacher-ratio and math test scores?

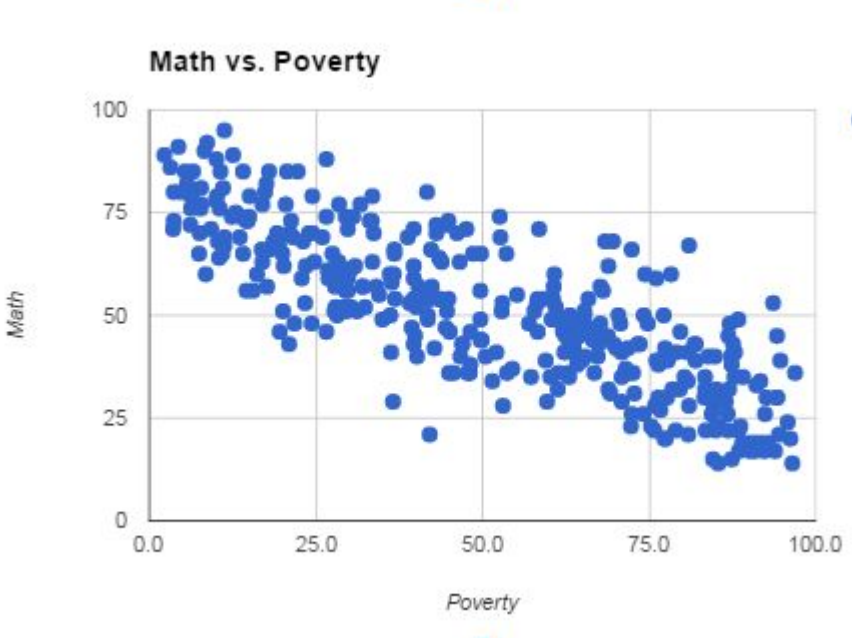
Cartesian coordinate review

In general, the x-axis is the thing we think is a predictor.

The y-axis is the thing we think might be related to the x-axis.



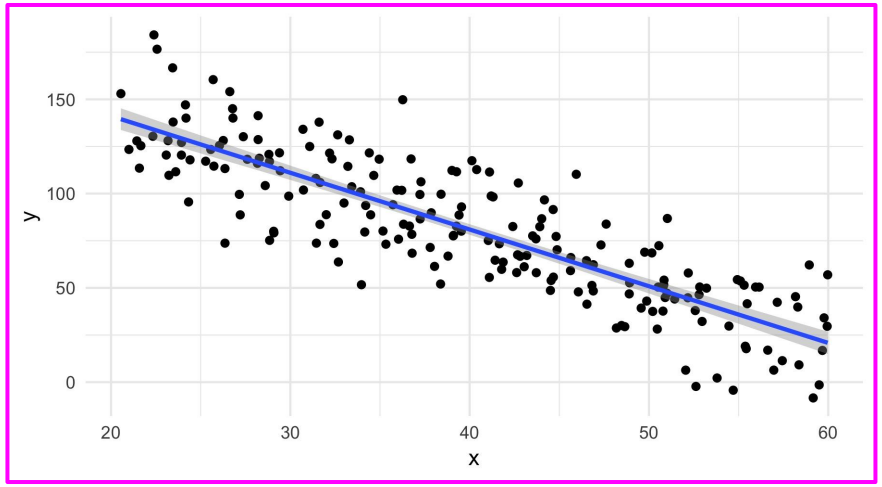
Are poverty and test scores related?



Correlation

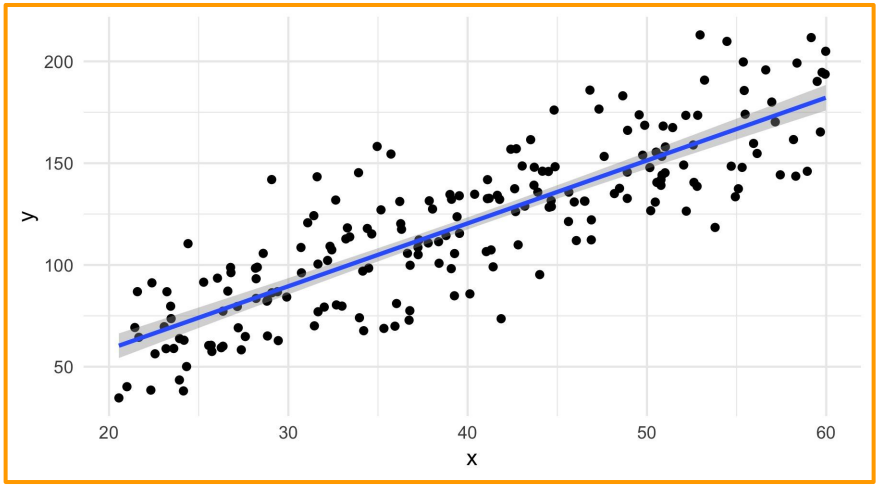
Negative correlation:

as one variable goes up, the other goes down

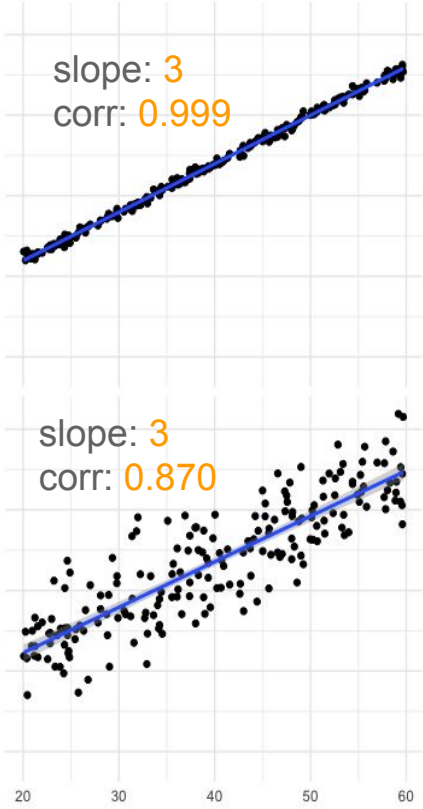


Positive correlation:

as one variable goes up, the other also goes up



Slope and correlation are different concepts



Using a scatterplot for reporting

1. What is the relationship between the variables overall?
    a. Positive or negative correlation?
    b. Very correlated or not so correlated?

2. Are there ‘outliers’, or observations that don’t fit the pattern? This can be the source of good reporting leads.