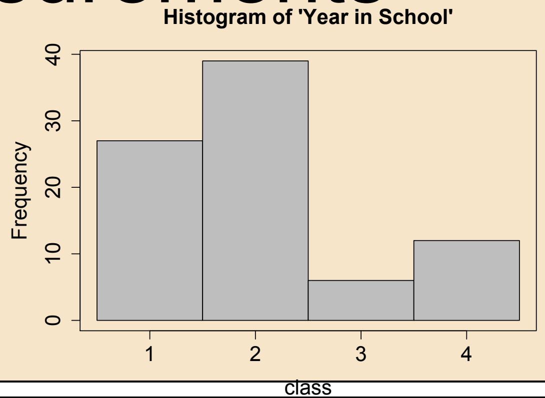
2.1 Data:

Types of Data
and
Levels of Measurement

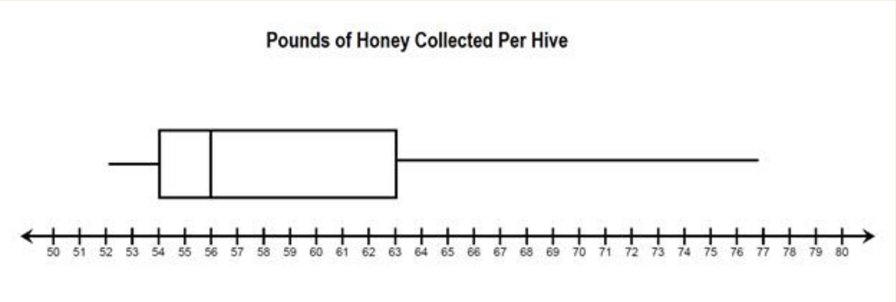
Quantitative or Qualitative?

- **Quantitative data** consist of values representing counts or measurements
  - Variable: Year in school



- **Qualitative** (or **non-numeric**) **data** consist of values that can be placed into *nonnumeric* categories.
  - Variable: Political affiliation (rep, dem, ind)

Types of Data



- **Quantitative**
  - Numerical values representing counts or measures.
  - Something we can `measure’ with a tool or a scale or count.
  - We can compare these values on a number line.
    - 2 pounds is less than 4 pounds
  - You can take a mathematical ‘average’ of these values, i.e. can be used in computations.
    - e.g. weight
    - e.g. number of students in a class

Types of Data

- **Qualitative** (or **non-numeric)**
  - Non-numerical in nature (but could be `coded' as a number, so be careful).
    - e.g. low=1, med=2, high=3 (still qualitative)
  - Could be considered a label in some cases.
    - e.g. Political affiliation (dem, rep, ind)
    - e.g. Numbers on a baseball uniform
      - #90 isn't "larger than" #45 in the mathematical sense. They're just a label.
    - e.g. ID (34B, 67AA, 19G, …)
    - e.g. Education level (HS, 2-yr, 4-yr, MS, PhD)

Types of Data

- **Qualitative** (or **non-numeric)**
  - Can't use meaningfully in a computation…
    - Can you take the average of the observed political affiliations? No, it's non-numerical.
      - Dem, Dem, Rep, Ind, Dem, Rep…
    - e.g. ID #s 56, 213, 788,… Average ID? no.
      - If variable is represented by numbers (as with IDs), ask yourself if an average makes sense… if not, then it's qualitative not quantitative.

Types of Data

- **Quantitative**
  - Number of medals won by U.S. in a given year.

- **Qualitative**
  - Medal Type: Gold/Silver/Bronze

Summer Olympic USA medalists 1896-2008

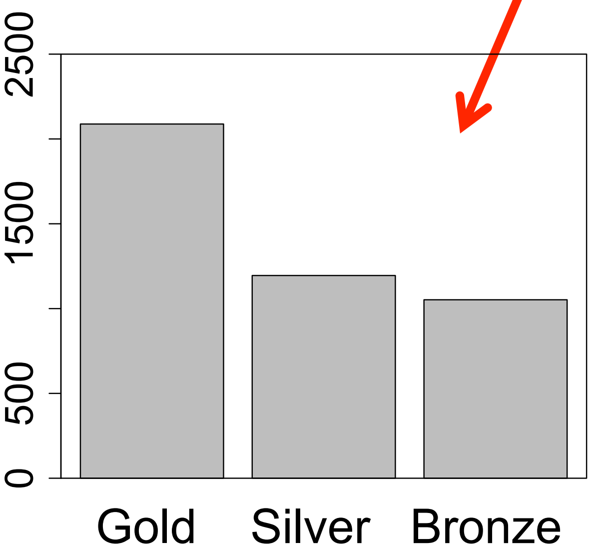
City	Edition	Sport	Discipline	Athlete	NOC	Gender	Event	Event_gende	Medal
Athens	1896	Athletics	Athletics	LANE, Franci:	USA	Men	100m	M	Bronze
Athens	1896	Athletics	Athletics	BURKE, Thon	USA	Men	100m	M	Gold
Athens	1896	Athletics	Athletics	CURTIS, Thor	USA	Men	110m hurdle	M	Gold
Athens	1896	Athletics	Athletics	BLAKE, Arthu	USA	Men	1500m	M	Silver
Athens	1896	Athletics	Athletics	BURKE, Thon	USA	Men	400m	M	Gold
Athens	1896	Athletics	Athletics	JAMISON, He	USA	Men	400m	M	Silver

Types of Data

- **Quantitative**
  - Number of medals won by U.S. in a single year.

- **Qualitative**
  - Medal Type: Gold/Silver/Bronze
  - Summarized with a table or chart.

Gold	Silver	Bronze
2088	1195	1052



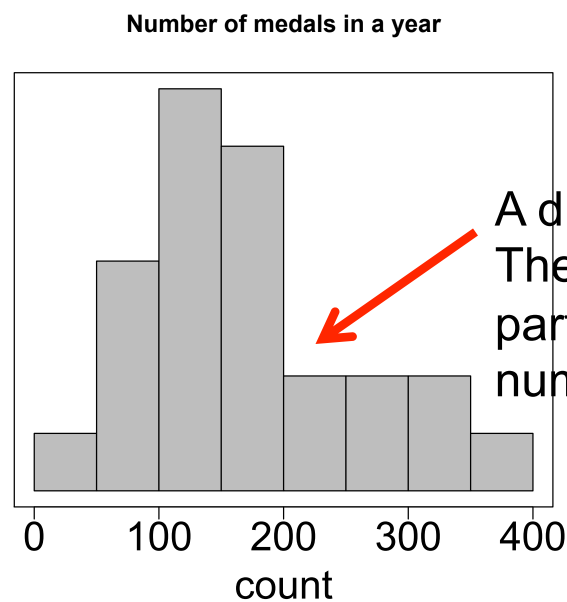
Types of Data

- **Quantitative**
  - Number of medals won by U.S. in a given year.
  - Can be shown with a distribution, or summarized with an average, etc.

- **Qualitative**
  - Medal Type: Gold/Silver/Bronze
  - Summarized with a table or chart.

With some reformatting of the earlier data, we can get a count of medals for each year.

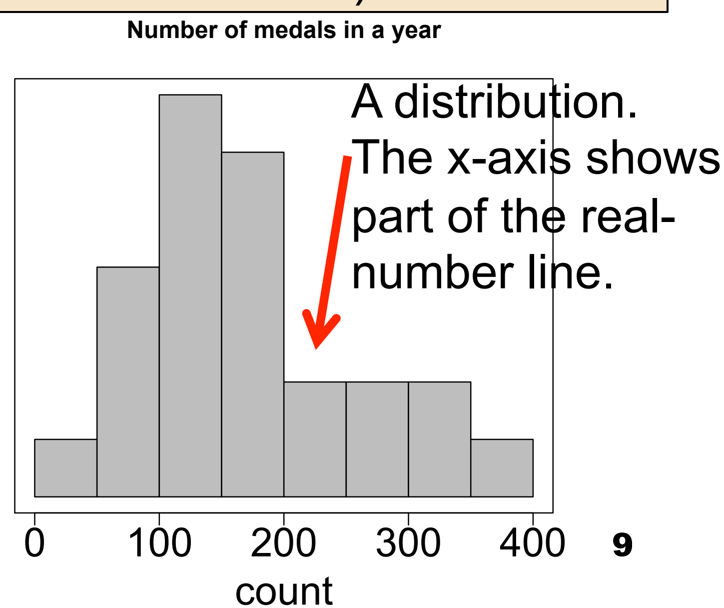
Year	Count
1896	20
1900	55
1904	394
1908	63
1912	101
1920	193



A distribution. The x-axis shows part of the real-number line.

Types of Data

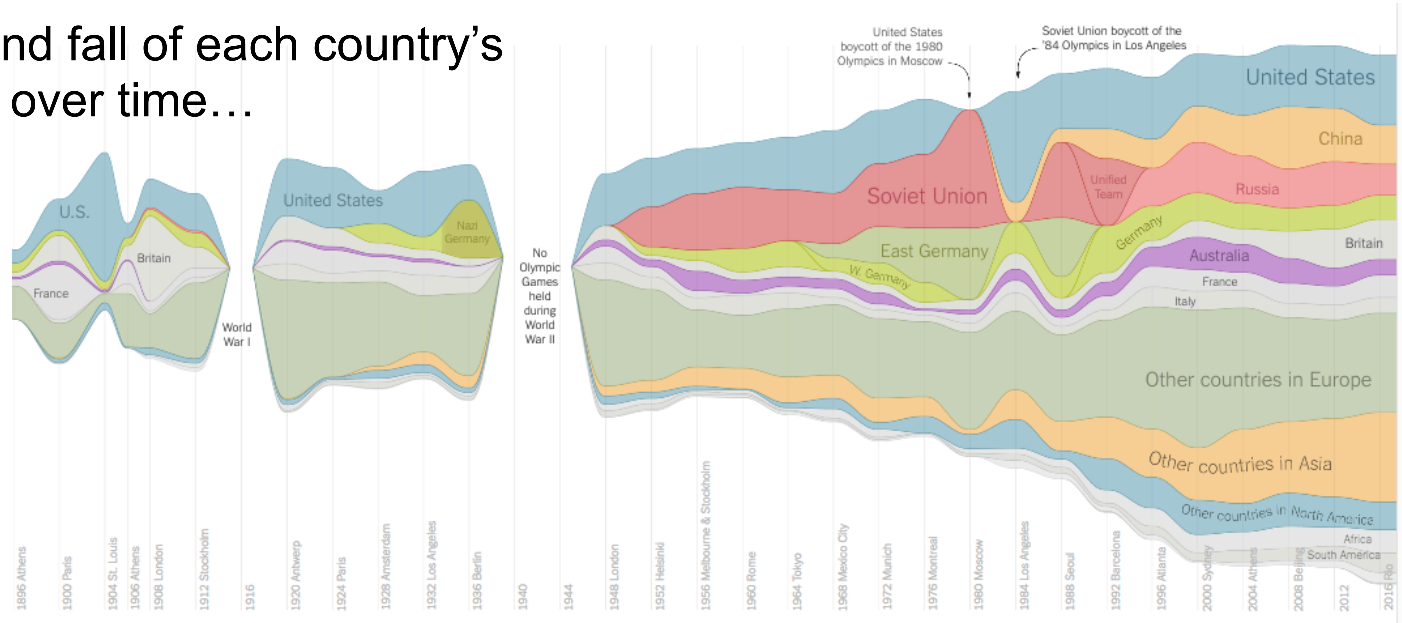
- **Quantitative**
  - Can be shown with a distribution, or summarized with an average, etc.
  - Commonly used summaries:
    - Average value
    - Maximum or Minimum value
    - Standard deviation (a measure of spread of the data)
  - Summarizing a distribution with a single value can be very useful.
  - But be aware that ‘averaging’ (or pooling, or aggregating) can potentially hide some interesting information (next slide).



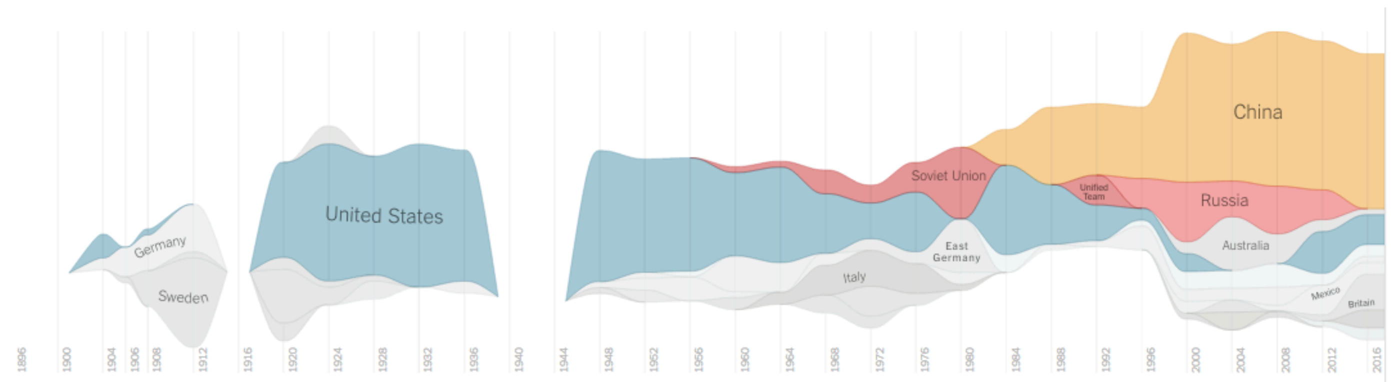
Pooling (or Aggregating) Data

Tracing the rise and fall of each country's total medal count over time…

All sports:



Only diving:



A Visual History of Which Countries Have Dominated the Summer Olympics, New York Times, Aug. 22, 2016

Levels of Measurement for Qualitative Data

- **Qualitative (two levels of qualitative data)**
  - Nominal level (by name)
    - No natural ranking or ordering of the data exists.
    - e.g. political affiliation (dem, rep, ind)
  - Ordinal level (by order)
    - Provides an order, but can't get a precise mathematical difference between levels.
      - e.g. heat (low, medium, high)
      - e.g. movie ratings (1-star, 2-star, etc.)
        - Watching **two** 2-star** movies isn't the same as watching **one** 4-star**** movie (the math not relevant here).
      - Could be coded numerically, so again, be careful.

Levels of Measurement for **Qualitative** Data

Political affiliation (dem, rep, ind) Nominal

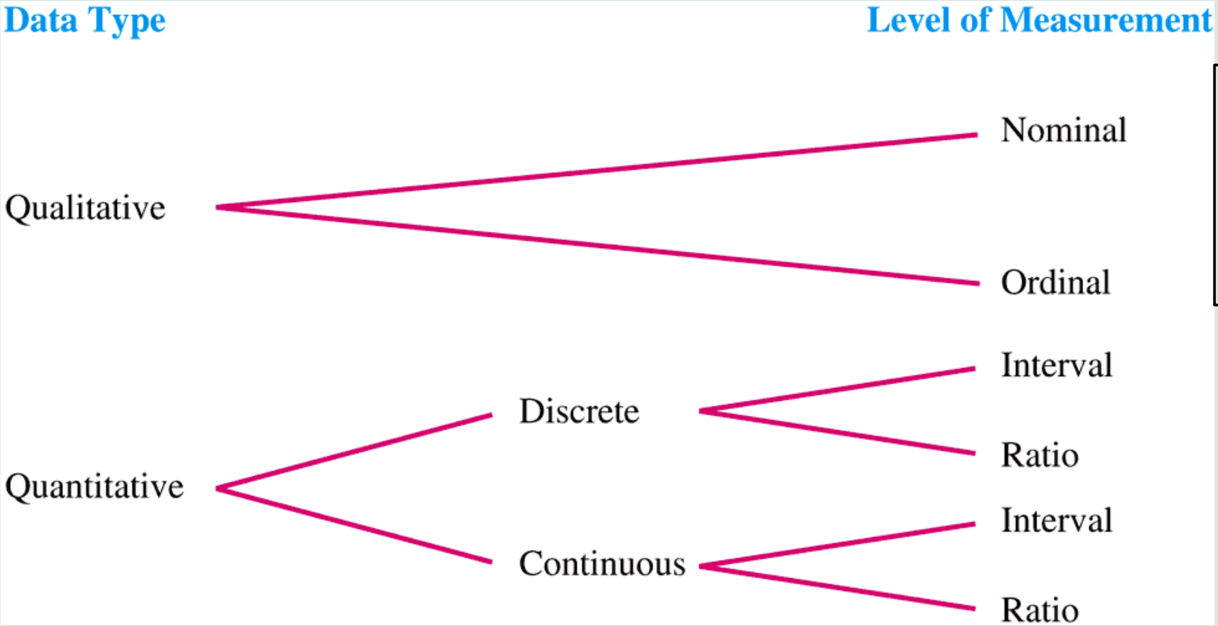
Level of pain (low, med, high) Ordinal

Answer to survey:
(strongly disagree, disagree, agree, strongly agree) Ordinal

Eye color (blue, green, brown, etc.) Nominal

Levels of Measurement

(Another way to characterize data)



Qualitative data is either *Nominal* or *Ordinal* (only 2 options)

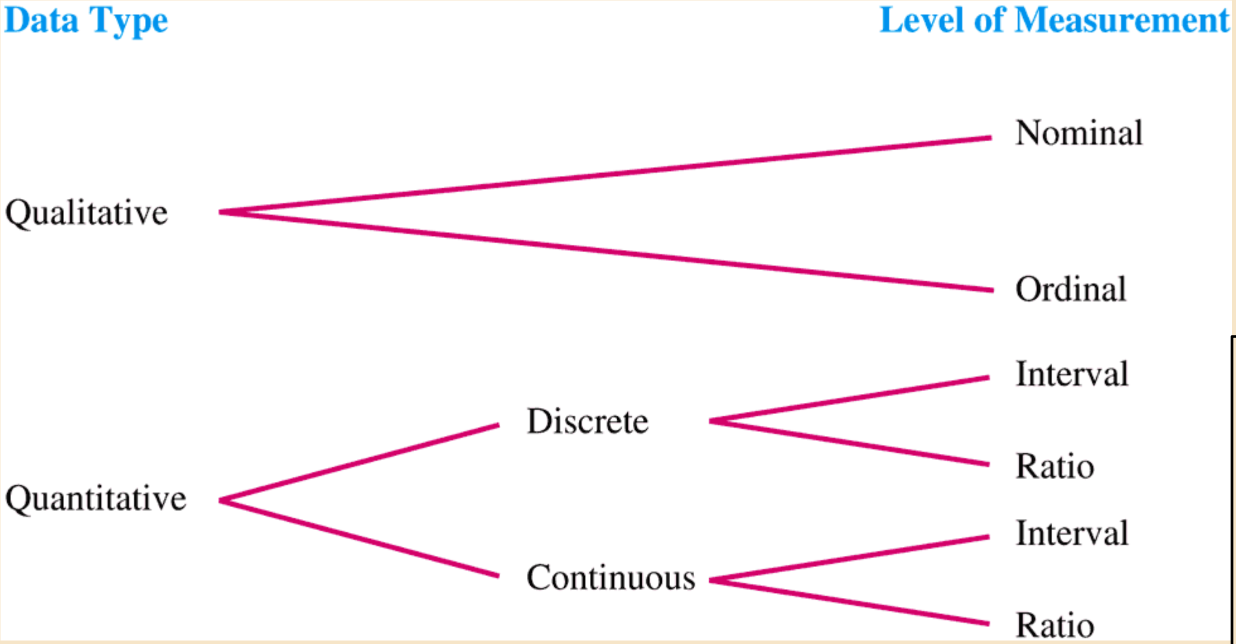
Two kinds of **Quantitative** Data

- **Continuous or Discrete?**
  - Continuous
    - Can take on **any** value in an interval
    - Could have any number of decimals
      - e.g. weight, home value, height
      - 2.45, 7.63454, 4.0, π, etc.

  - Discrete
    - Can take on <u>only particular values</u>
      - e.g. number of prerequisite courses (0, 1, 2, …)
      - e.g. number of students in a course
      - e.g. shoe sizes (7, 7-1/2, 8, 8-1/2,…)

Levels of Measurement

(Another way to characterize data)



Quantitative data can be either *Discrete* or *Continuous* and either *Interval* or *Ratio*

Levels of Measurement for **Quantitative** Data

- **Interval or Ratio?**
  - Interval level (a.k.a differences or subtraction level)
    - Intervals of equal length signify equal differences in the characteristic.
      - The difference in 90° and 100° Fahrenheit is the same as the difference between 80° and 90° Fahrenheit.
    - Differences make sense, but ratios do not.
      - 100° Fahrenheit is not twice as hot as 50° Fahrenheit.
    - Occurs when a numerical scale does not have a ‘true zero’ start point (i.e. it has an arbitrary zero).
      - Zero does not signify an absence of the characteristic.
      - Does 0° Fahrenheit represent an absence of heat?
    - Designates an equal-interval ordering.
      - 1 to 2 has the same meaning as 3 to 4.

Levels of Measurement for Quantitative Data

- **Interval or Ratio?**
  - Interval level (a.k.a differences or subtraction level)
    - May initially look like a qualitative ordinal variable (e.g. low, med, high), but levels are quantitative in nature and the differences in levels have consistent meaning.
      - Scale for evaluation:

5 Strongly Agree	4 Agree	3 Neither Agree nor Disagree	2 Disagree	1 Strongly Disagree	N/A Does Not Apply

      - If a change from 1 to 2 has the same strength as a 4 to 5, then we would call it an interval level measurement (**if not, then it's just an ordinal qualitative measurement**).
      - To be an interval measurement, each sequential difference should represent the same quantitative change.
      - But a "5" is not 5 times a "1" (ratios don't make sense here).
      - This could have been on a 6 to 10 scale (arbitrary start).

See comment on slide 20.

Levels of Measurement for Quantitative Data

- **Interval or Ratio?**
  - Interval level (a.k.a differences or subtraction level)
    - IQ tests (interval scale).
      - We don't have meaning for a 0 IQ.
      - A 120 IQ is not twice as intelligent as a 60 IQ.
    - Calendar years (interval scale).
      - An interval of one calendar year (2005 to 2006, 2014 to 2015) always has the same meaning.
      - But ratios of calendar years do not make sense because the choice of the year 0 is arbitrary and does not mean "the beginning of time."
      - Calendar years are therefore at the interval level of measurement.

Levels of Measurement for **Quantitative** Data

- **Interval or Ratio?**
  - Ratio level (even *more* meaning than interval level)
    - At this level, both differences <u>and</u> ratios are meaningful.
      - **Two** 2 oz glasses of water IS equal to **one** 4 oz glass of water
      - 4 oz of water is twice as much as 2 oz of water.
    - Occurs when scale *does* have a 'true zero' start point.
      - 0 oz of water is a 'true zero' as it is empty, absence of water.
    - Ratios involve division (or multiplication) rather than addition or subtraction.

Levels of Measurement for **Quantitative** Data

- **Quantitative** – Interval level example
  - Temperature used to cook food*.

A brownie recipe calls for the brownies to be cooked at 400 degrees for 30 minutes.

Would the results be the same if you cooked them at 200 degrees for 60 minutes? How about at 800 degrees for 15 minutes?

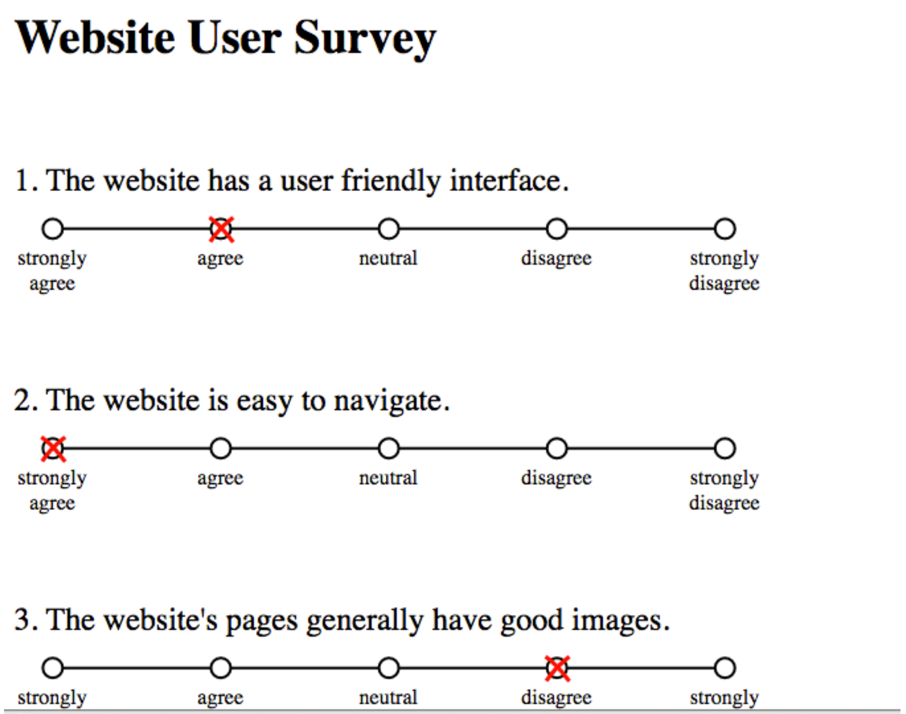
200 degrees is not half as hot as 400 degrees. The ratio of temperatures does not make sense here.

Levels of Measurement for **Quantitative** Data

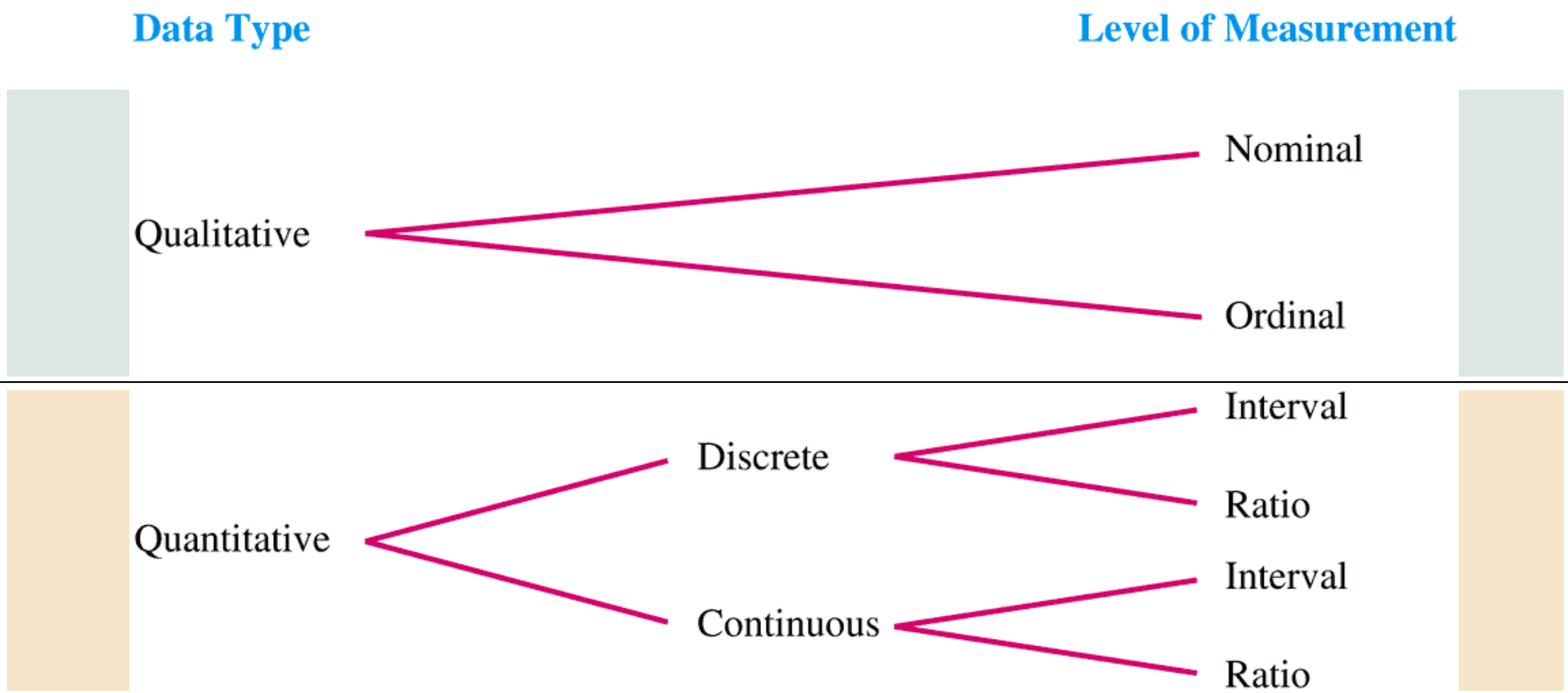
- **Quantitative** - Ratio level examples
  - Centimeters
    - Difference of 40 cm (an interval) makes sense and has the same meaning anywhere along the scale.
    - 10cm is twice as long as 5 cm (put two 5 cm items together and they are equivalent to 10 cm). Ratios make sense.
    - 0cm truly represents ‘no length’ or absence of length.
  - Mass
  - Length
  - Time

Likert scale (sometimes unclear)

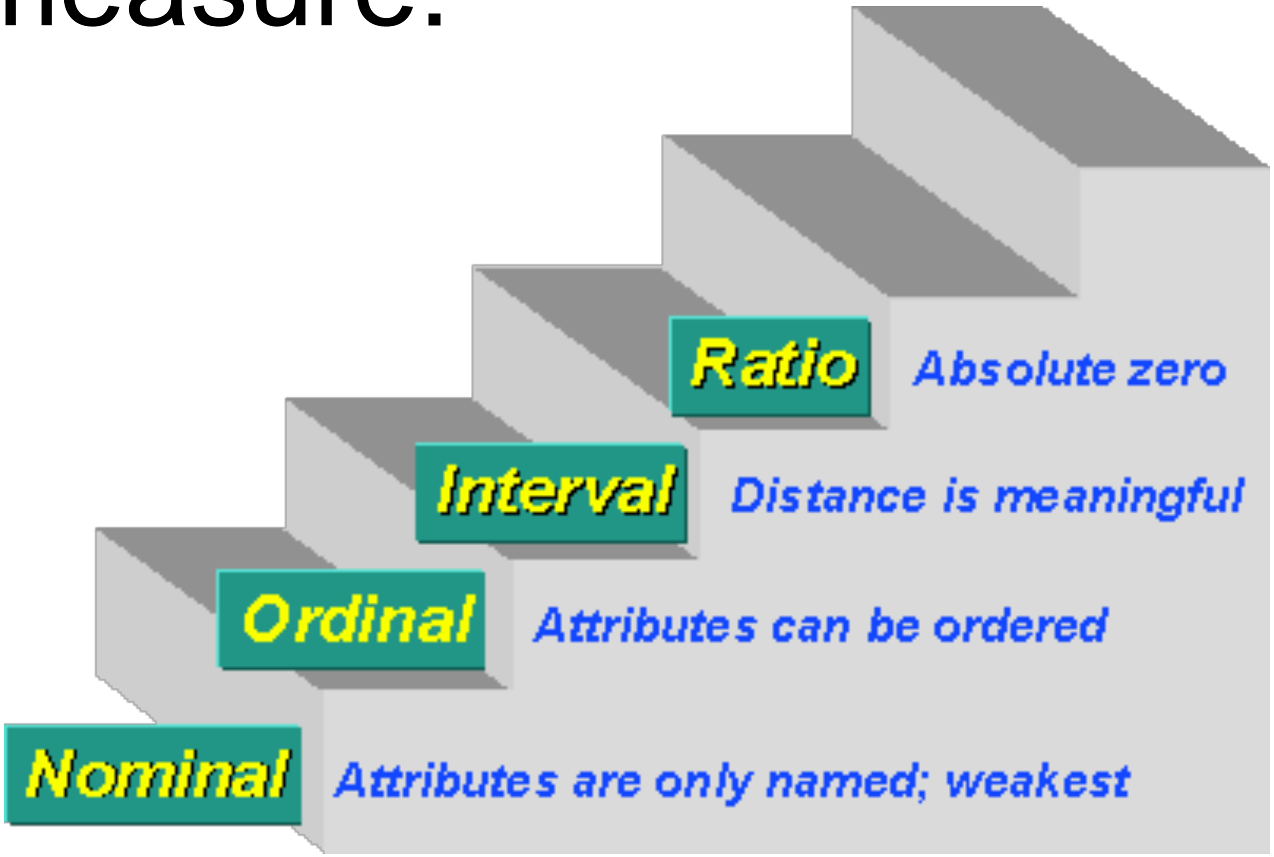
- **Is it Interval (Quantitative) or Ordinal (Qualitative) scale?**
  - I think, most of the time, these surveys are just ordering responses lowest to highest and NOT fulfilling the interval scale requirements.
  - Difference of opinions on this though.



Possible data types and levels of measure.



Possible data types and levels of measure.



As a statistician, the type of data that I have dictates the type of analysis I will perform.

2.2 Dealing with Errors

- Types of errors:
  - Random vs. Systematic errors
- Size of Errors:
  - Absolute vs. Relative
- Describing Results:
  - Accuracy and Precision

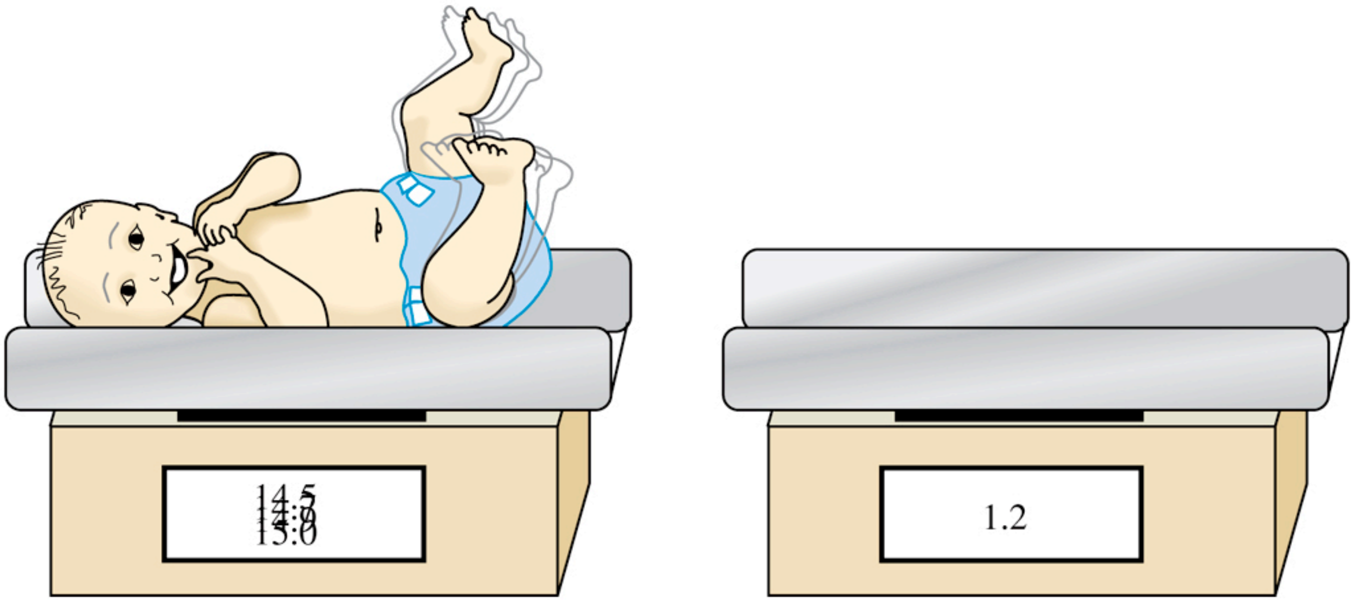
Types of Errors

- Random errors:
  - Affects measurement in an unpredictable manner
    - Baby squirming on a scale
    - may cause error above or below truth
    - Introduces random noise to your measurement
- Systematic errors:
  - Error that affects all measurements in a similar fashion
    - Scale systematically weighs all babies a little too high (scale needs to be *calibrated*).

Types of Errors

- Random errors:
  - Just part of the process we have to deal with, sometimes called noise
  - We can measure object numerous times and take an average to reduce the effect of the random error
- Systematic errors:
  - We may be able to remove the error if the source can be detected (e.g. recalibrating)
  - After data collected, can be corrected if error is detected and quantified.

Types of Errors



(a) (b)

Picture (a) on the left represents a baby's motion, which introduces **random errors** to the measurement process.

Picture (b) on the right shows the scale reads 1.2 pounds when empty, introducing a **systematic error** that makes all measurements 1.2 pounds too high.

Size of Errors

- Consider a clerk that made a mistake and overcharged you $1.
  - What if you had just bought…
    - 1) A $1 piece of pie.
    - 2) A $30,000 car.
  - Would you see the $1 discrepancy differently?
    - Should we consider the mistake relative to the price?

Size of Errors

- 1) \$1 overcharge on a \$1 piece of pie:
  - **Absolute** value of overcharge: \$1.00
  - **Relative** value of overcharge: $\frac{1}{1} = 1$ or 100%
- 2) \$1 overcharge on a \$30,000 car:
  - **Absolute** value of overcharge: \$1.00
  - **Relative** value of overcharge: $\frac{1}{30000} = 0.00003$ or 0.003%

Size of Errors

- This idea can be applied to measurement errors…
- **Absolute errors** are expressed as a difference in units
- **Relative errors** are expressed as a ratio with the true value in the denominator and the error in the numerator

Size of Error: Absolute versus Relative

Absolute and Relative Errors

The **absolute error** describes how far a claimed or measured value lies from the true value:

absolute error = claimed or measured value – true value

The **relative error** compares the size of the absolute error to the true value. It is often expressed as a percentage:

$$\text{relative error} = \frac{\text{absolute error}}{\text{true value}} \times 100\%$$

- Example: True weight is 25 pounds, but the scale reads 26.5
  - Absolute error:

    26.5 pounds -25 pounds = 1.5 pounds
  - Relative error:

$$\frac{1.5 \text{ pounds}}{25 \text{ pounds}} \text{ X } 100\% = 6\%$$

Accuracy vs. Precision

- If a measured value is close to the truth, it has **accuracy**.
  - We usually quantify 'close' in relative terms rather than absolute terms.
- **Precision** describes the amount of detail (or resolution) in a measurement.
  - Suppose your true salary is $47,500…
    - Telling someone your salary is $49,546 sounds more precise (to a specific dollar) than saying it is $49,000, but the $49,000 statement is more accurate (closer to truth).
    - Precision doesn't necessarily coincide with accuracy.

Accuracy vs. Precision

- Suppose that your true weight is 102.4 pounds. The scale at the doctor's office, which can be read only to the nearest quarter pound, says that you weigh 102¼ pounds. The scale at the gym, which gives a digital readout to the nearest 0.1 pound, says that you weigh 100.7 pounds.
  - Which scale is more **precise**? Which is more **accurate**?

Summary: Dealing with Errors

- Errors can occur in many ways, but generally can be classified into one of two basic types: **random errors** or **systematic errors**.
- Whatever the source of an error, its size can be described in two different ways: as an **absolute error** or as a **relative error.**
- Once a measurement is reported, we can evaluate it in terms of its **accuracy** and its **precision**.