Lecture 3 – Statistics I

January 29th, 2026

Spring 2026

Today’s Learning Outcomes

1. Be able to identify sources of variation in datasets

2. Be able to explain the differences between the term pairs accuracy/precision and sample/population

3. Be able to explain the relationship between sample size, bias, and accuracy

Variation In The Universe

- With very few exceptions (ex. the speed of light) most measurable properties vary in the real world
  - Height of individuals
  - Length of time it takes to eat a bowl of cereal
  - Percent of feldspar in a granite

- We gather and analyze data because the vast majority of systems are variable
  - i.e. we cannot look at one instance and safely assume that this will be the case every time

Variation In The Universe

- For example, if a student does extremely poorly on the first exam in a class it's no guarantee they will do poorly on the next exam
  - Maybe the first assignment was hard to weed out people before the drop/add period was over?
  - Maybe they slept poorly the night before?
  - Maybe they didn't attend many lectures or study?

- To figure out which, if any, of the reasons above account for differences among student performance we need to collect more data

Variation In The Universe

- Gathering lots of data to look for consistent patterns of variation between information is the realm of statistics

- Statistics is concerned with estimating and explaining observed variation
  - How variable is sleep among students the night before an exam?
  - What is the source of the observed variation?
  - How strongly do length of sleep and exam scores track one another?
  - What ago do young people start using social media in the United States?

Collecting Data

- First step in any study to gather data
  - Field observations
  - Literature search
  - Online database query

- What data are needed to answer the question?
  - Get back to this part later…

- What are the sources of variation in a dataset?
  - Ex. mugs in UB Earth Sciences kitchenette

Mug Dataset

Measure 4 variables of 24 mugs in the Cooke kitchenette

Cause(s) of variation?

Height_cm	OpeningDiameter_cm	Color	Text
13	7.5	blue	no
10	8	black	no
9.5	7.5	blue	yes
10	8.5	brown	yes
9	7.5	white	yes
9.5	9.5	grey	yes
10	8.5	white	no
10	7	blue	yes
8.5	9.5	red	yes
9.5	8.5	white	yes
9	7.5	white	yes
10	7.5	mixed	no
14.5	8	silver	yes
9	10.5	yellow	no
12.5	8.5	red	no
11.5	10.5	white	yes
10.5	8	orange	yes
11	7.5	white	no
8.5	9	blue	yes
9	10	yellow	yes
8.5	7.5	white	no
6.5	7	white	no
9.5	7.5	brown	no
10	8.5	brown	no
17	7.5	white	no

Collecting Data

- Types of data
  - Continuous
    - Numerical data with theoretically infinite values
    - Ex. height of the mug
  - Discrete
    - Numerical data with specific finite values
    - Ex. number of children in a family

Collecting Data

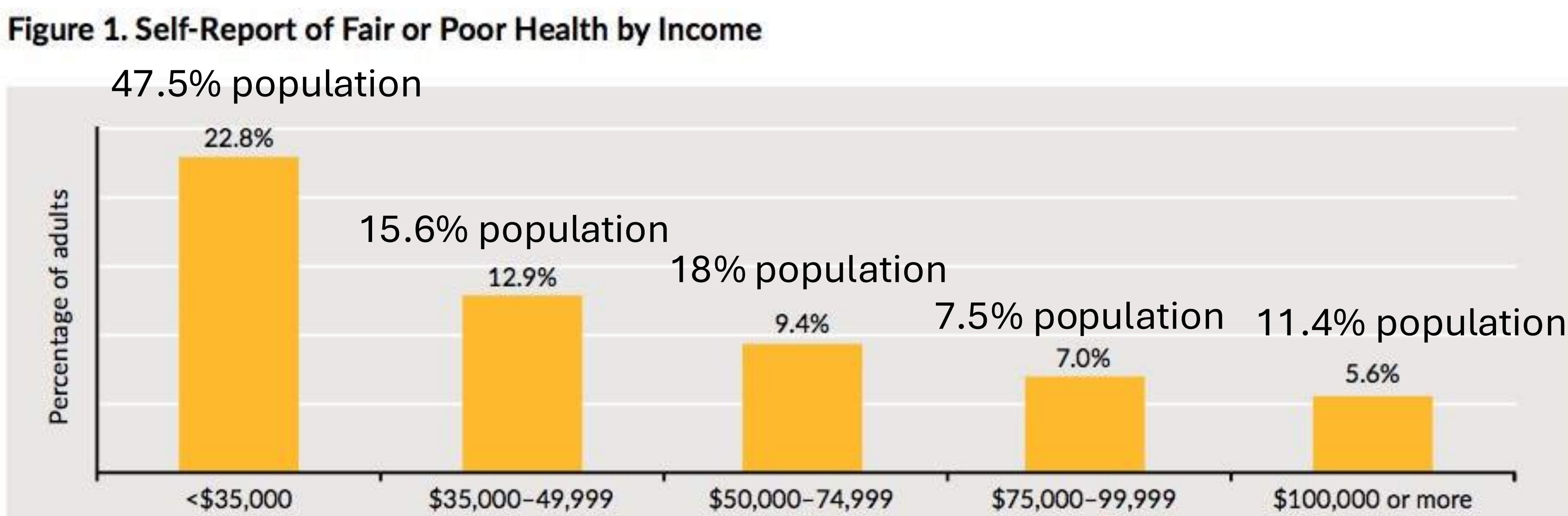
- Types of data
  - Ordered
    - Data which are inherently ranked by magnitude, but the differences between values need not be equal
    - Ex. Income brackets (<$20,000; $20,000-50,000; $50,000-100,000; >$100,000)

  - Categorical
    - Data which can be non-numeric and are unranked
    - Ex. Color, sex, volcano type, presence/absence

*Data can belong to multiple categories
(mostly combinations of continuous/discrete and order)

Misleading Binning

- Can split continuous data into bins (converting to discrete ordered data)
  - Used to create evenly sized samples (sometimes desirable)
  - Can also be used to skew data towards a particular narrative or obscure variation by averaging



Sources of Variation

- What data was collected?
  - Was a variable of interest not recorded?
  - Was one type of data more likely to be collected?

- How was the data measured?
  - What lab the sample was analyzed at?
  - What tool was used to take a measurement?
  - How precise was the measurement?
  - Was the data collected consistently?*

- How much data was collected?
  - Random sample error

*Generally a bad idea to adjust your data collection in the middle of a project as it introduces an additional source of variation

Sample Versus Population

- The total theoretical set of observations/samples is known as a <u>population</u>

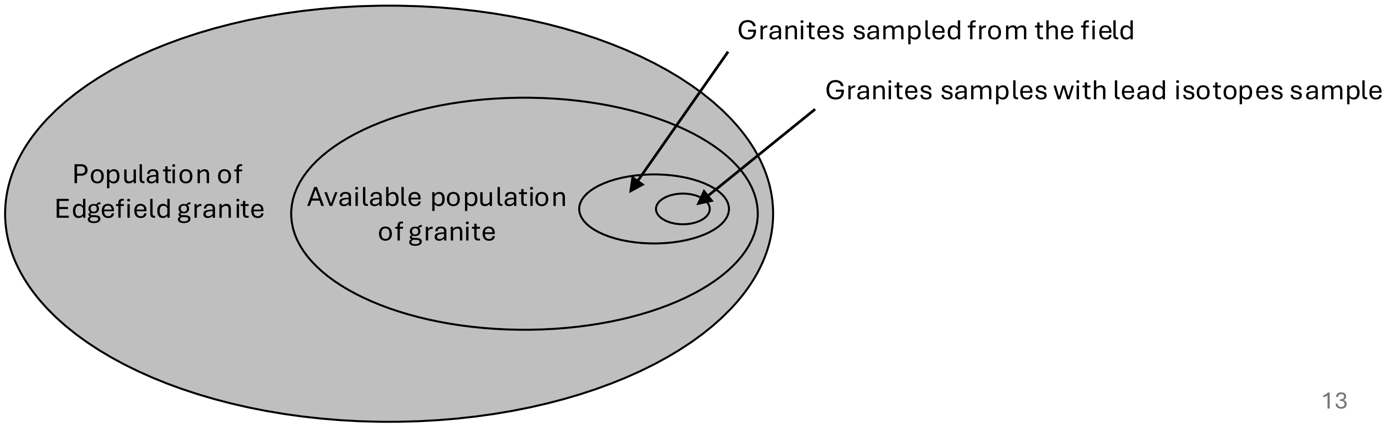
- In general, we cannot gather every piece of informatic for a given system
  - Ex. I cannot measure every fossil snail shell

- Instead we take a subset of data from the population extrapolate the observations
  - This subset is known as a <u>sample</u>

Sample Versus Population

- In many geology problems the true population no longer exists and has been lost through time
  - Instead we have an available population

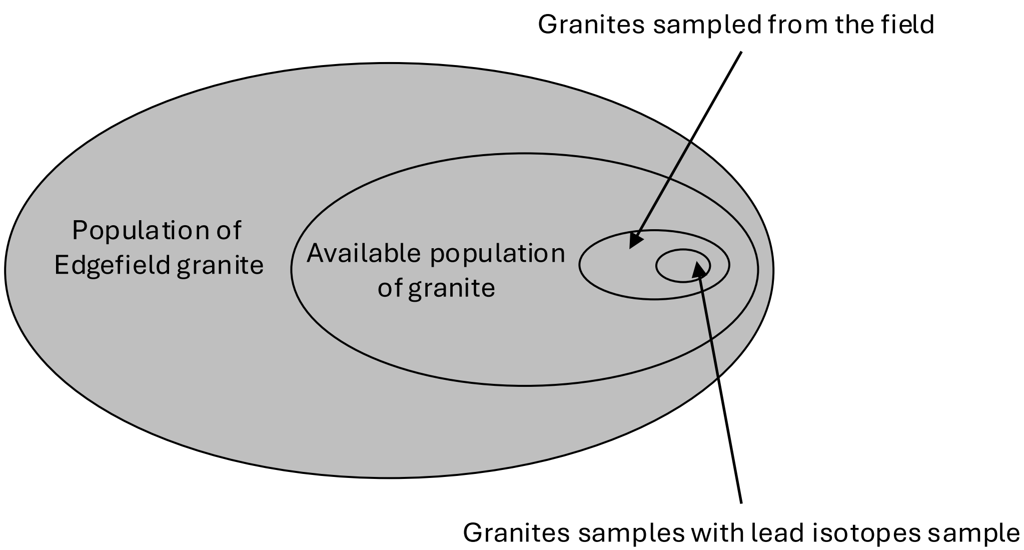
- Let's say we want to get a radiometric date of a specific granite (Edgefield) so that our sample data will be Pb isotope values

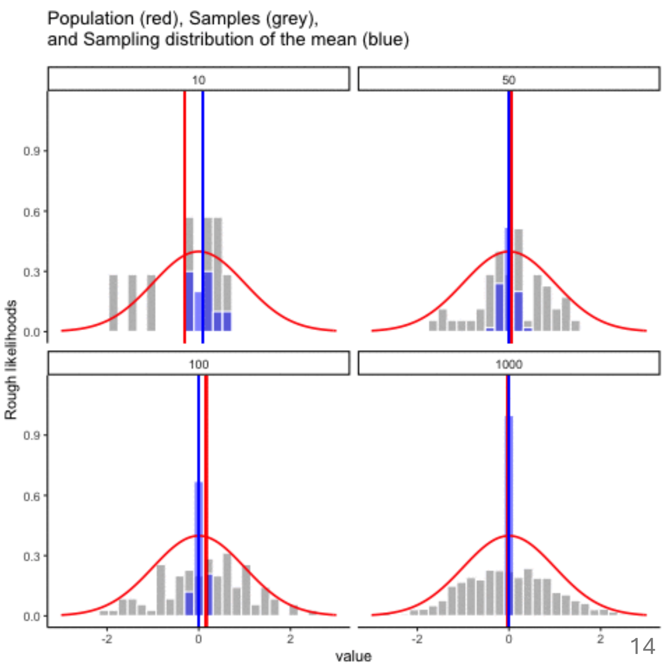


Sample Versus Population

- The hope when sampling is that the sample is representative of the population

- But even if it is you can get misleading results due to random chance





Sample Size (n)*

- General perception that increasing sample size will makes a study/measure/result more reliable (i.e. closer to the reality of the population)
  - Ex. flipping a coin 20 times and getting 15 heads is not unusual, but flipping a coin 2000 times and getting 1500 heads would be extremely unlikely for a fair coin

- That correlation is only true <u>IF</u> your sample is truly representative of the population

*Robust sample size varies depending on the effect size in question

Sampling Bias

- If our sample tends to select (or not select) certain parts of a population it is no longer representative
  - This overrepresentation of certain parts of a population is referred to as bias

- Ex. We want to characterize the global distribution of igneous rock types on the planet
  - I choose to use Macrostrat for this purpose
  - Only includes terrestrial deposits (database limitation)
  - Older igneous rocks are more likely to have been destroyed
  - Equatorial igneous rocks are more likely to have been eroded than higher latitude deposits

Sampling Bias

1. Only includes terrestrial deposits (database limitation)
   - Limit my conclusions to talking about terrestrial settings (i.e. do not generalize beyond what my sample represents)

2. Older igneous rocks are more likely to have been destroyed
   - Gather data on age to check whether there is evidence for this and if so, account for age as a variable

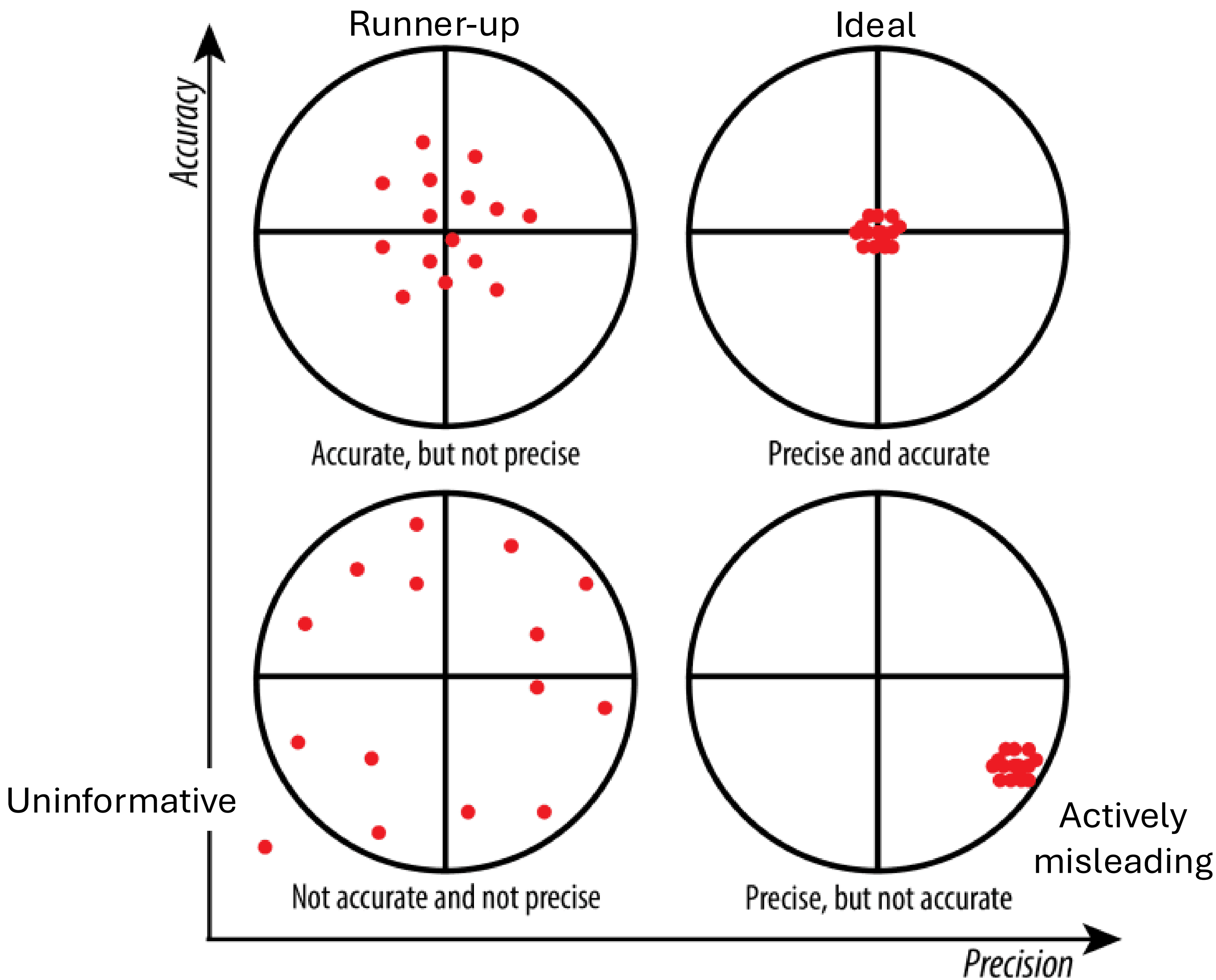
3. Equatorial igneous rocks are more likely to have been eroded than higher latitude deposits
   - Gather data on latitudinal position (or even better erosion rates) of the deposit [paleolatitude complication], check for evidence of relationship, add variable to model

Bias, Accuracy, and Precision

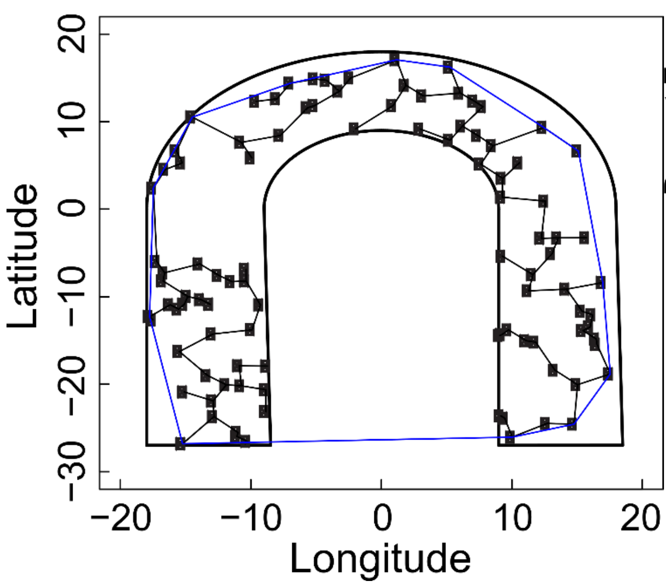
- In statistics values can vary on two axes
  - Accuracy (how close to the true value an estimate is)
  - Precision (how repeatable the estimate is)

- Ideally, as sample size increases will make an estimate more accurate and precise
  - Precision is generally controlled by methods

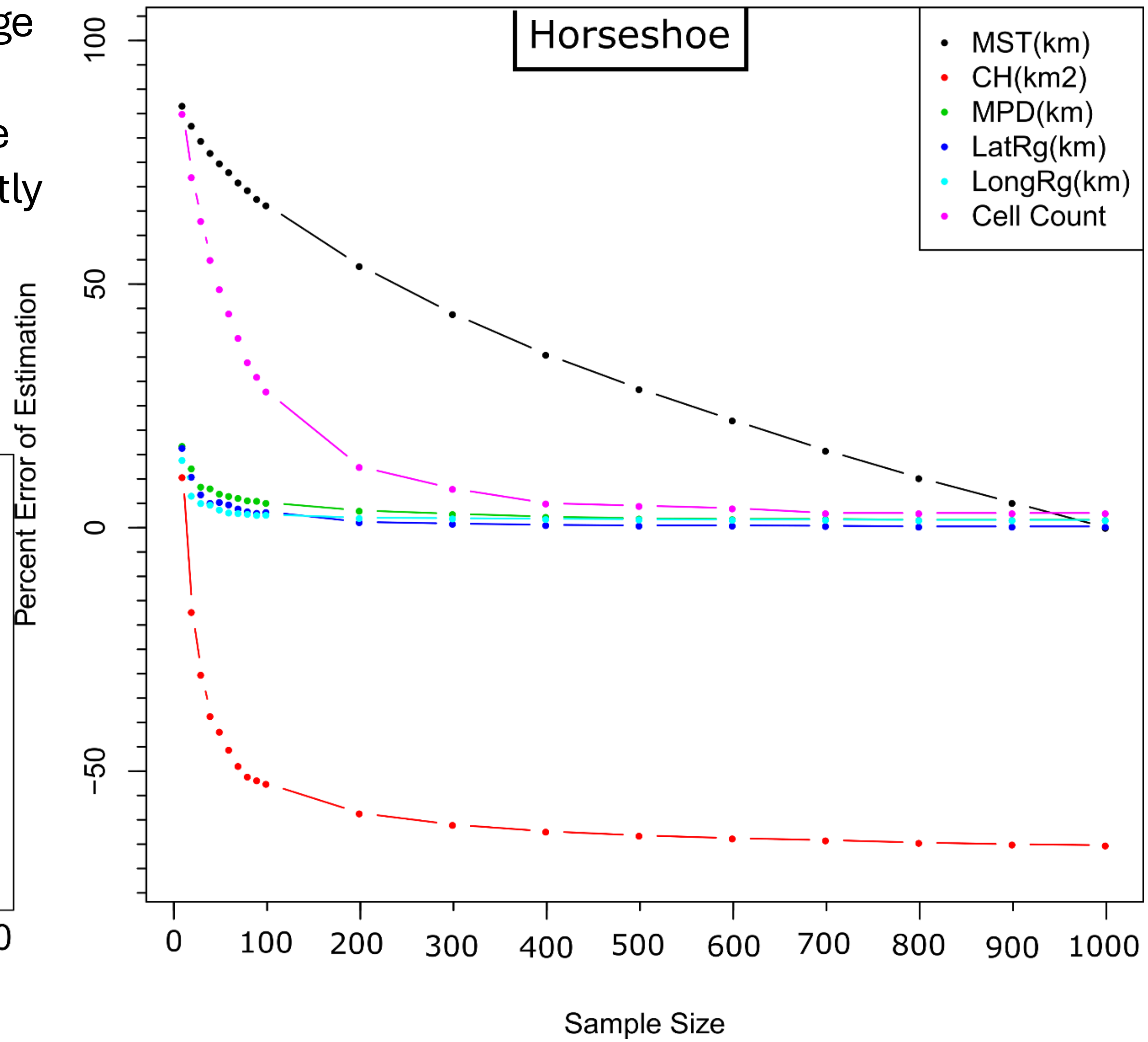
- However, uncontrolled bias will make an estimate less accurate even as precision increases!


Runner-up
Ideal
Accuracy
Accurate, but not precise
Precise and accurate
Uninformative
Actively misleading
Not accurate and not precise
Precise, but not accurate
Precision

Convex hull (CH) is a method of estimating geographic range size. If we have a horseshoe shape increasing sample size increases accuracy, but greatly overestimate actual geographic range



Y-axis is % difference from actual value
(negative values are overestimates)



General Rules of Data Collection

- Always consider what your population is for the question at hand
- Identify what & how data are going to be gathered prior to actually gathering any data
- For most projects gathering and/or cleaning data is the most time/labor intensive part
  - It’s very rare that gathering data quickly and making adjustments afterwards saves you time

General Rules of Data Collection

- Identify possible biases in data collection and what additional data might be required to control for them

- Larger sample sizes are good, but very large sample sizes are **not** a substitute for high-quality data
  - GIGO (garbage in, garbage out)
  - No model/technique can correct for fundamentally flawed data

Summarizing Data

- Since we want large sample sizes of variable data there's a need to summarize the data (i.e. reduce it to a few easily understood numbers)

- Variance (focus of next lecture)
  - Standard deviation
  - Confidence intervals
  - Max/Min values

Range of variance

- Shape of a distribution
  - Mean
  - Mode
  - Percentiles
    - Median
    - Quartiles
    - Quintiles

Which values are most common relative to other values

Summarizing Data

- As with any summary these numbers flatten out variation

- Always useful to plot your actual data and make sure they are appropriate and actually represent what you think they do

Load the dataset

```
%matplotlib inline
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
```

[1] ✓ 3.0s

```
df = pd.read_csv('MugData.csv')
df.info()
```

[3] ✓ 0.0s

```
<class 'pandas.DataFrame'>
RangeIndex: 25 entries, 0 to 24
Data columns (total 4 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   Height_cm           25 non-null     float64
 1   OpeningDiameter_cm  25 non-null     float64
 2   Color               25 non-null     str
 3   Text                25 non-null     str
dtypes: float64(2), str(2)
memory usage: 932.0 bytes
```

Do some quick plotting

```
mug_color = df['Color'].values
mug_heigh = df['Height_cm'].values
```

✓ 0.0s

```
mug_color_fix = []

for color in mug_color:
    if color=='mixed':
        mug_color_fix.append('gray')
    else:
        mug_color_fix.append(color)
```

✓ 0.0s

```
plt.scatter(df['Height_cm'].values,
            df['OpeningDiameter_cm'].values,
            c=mug_color_fix)
```

✓ 0.1s

Simple statistical analysis

```
df['Height_cm'].mean()
```

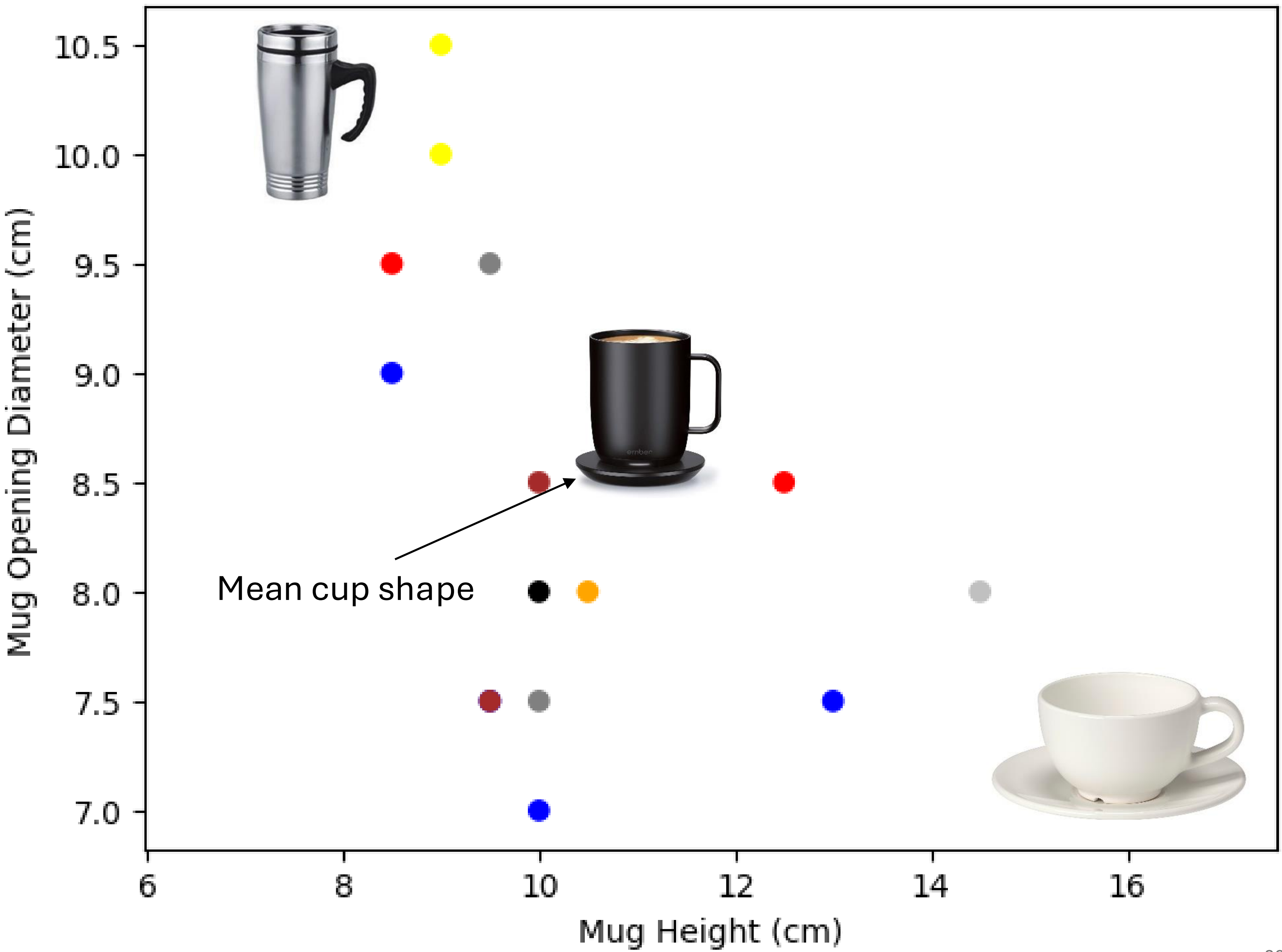
✓ 0.0s

```
np.float64(10.24)
```

```
df['Height_cm'].median()
```

✓ 0.0s

```
np.float64(10.0)
```


10.5
10.0
9.5
9.0
8.5
8.0
7.5
7.0
Mug Opening Diameter (cm)
Mean cup shape
6
8
10
12
14
16
Mug Height (cm)

Today's Learning Outcomes

1. Be able to identify sources of variation in datasets

2. Be able to explain the differences between the pairs of terms accuracy/precision and sample/population

3. Be able to explain the relationship between sample size, bias, and accuracy

Demo: how to submit your homework