BOISE STATE UNIVERSITY

CS 533
INTRO TO DATA SCIENCE

Instructor: Jun Zhuang

Source: https://cs533.ekstrandom.net/f22/

DESCRIBING DATA

Data and Dataset

Data

Photo by Dan-Cristian Păduret on Unsplash

Data set

Photo by Brian Kostiuk on Unsplash

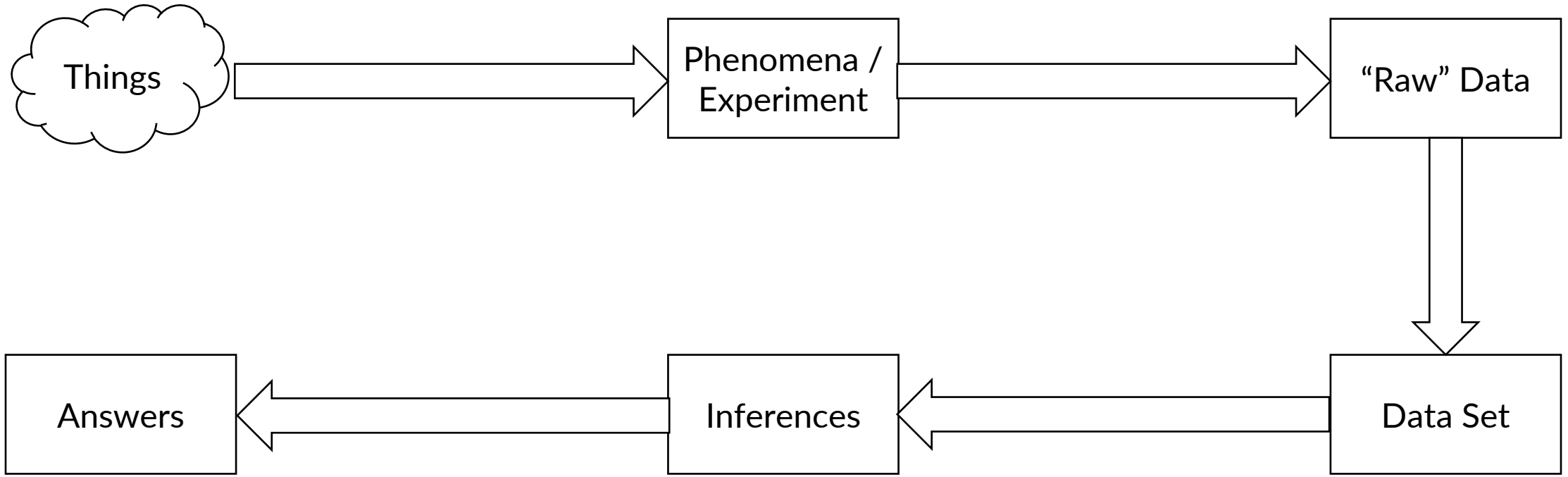
Data sets

Many definitions, see the reading.

A set of data that is

- Collected or curated
- For a purpose
- Mostly ready-to-use
- Documented

Where does it come from? Where does it go?



What We Need To Know

- How much data do we have?
- What kind(s) of data do we have?
- What is the data about?
- How was the data collected?
- How was the data recorded?
- What biases might the data have?
- What do we know about the *data generating process*?

Reading discusses more!

Wrapping Up

Data sets arise from curating or collecting data, resulting from observations, for a purpose.

There are layers between what we want to study and the data we have.

Do the reading!

Photo by Rick Mason on Unsplash

DESCRIPTIVE STATISTICS

Statistic

A **statistic** is a value computed from a collection of data

Often summarizes (observations of) a variable

Descriptive Questions

- Where is the variable centered?
  - How large does it tend to be?
  - Called *meaure of central tendency*
- How spread out is it?

Mean

The **mean** $\bar{x}$ of a series $x_1, x_2, \ldots, x_n$ is

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

This is a **measure of central tendency**

Often *informally* called ‘average’, but average is not specific

Pts
8
15
0
45
8
7
0
2
0
2
$\bar{x} = 8.7$

What is a Mean?

What does the mean measure?

- If every instance had the same value, what would it be?
  - “Points per player”

How do you change it?

- Increase total (score more points)
- It doesn’t matter where – one person can do all of it!

Spread - Variance and Standard Deviation

How spread out are the values?

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x - \bar{x})^2}{n - 1}}$$

This is the *sample standard deviation* – used for empirical data

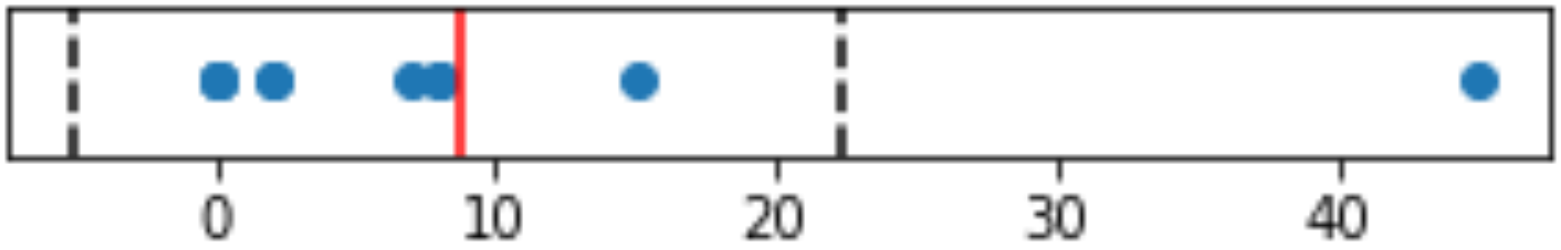
The *sample variance* is s^2

What is Variance?

What does variance measure?

- The mean squared distance from the mean
  - If mean is center, how far away do values tend to be?
  - If mean is expectation, how far off does it tend to be?
  - Square penalizes large differences more

Standard deviation translates back to original units



Computing Statistics

Pandas:

- Series.mean()
- Series.std()
- Series.var()

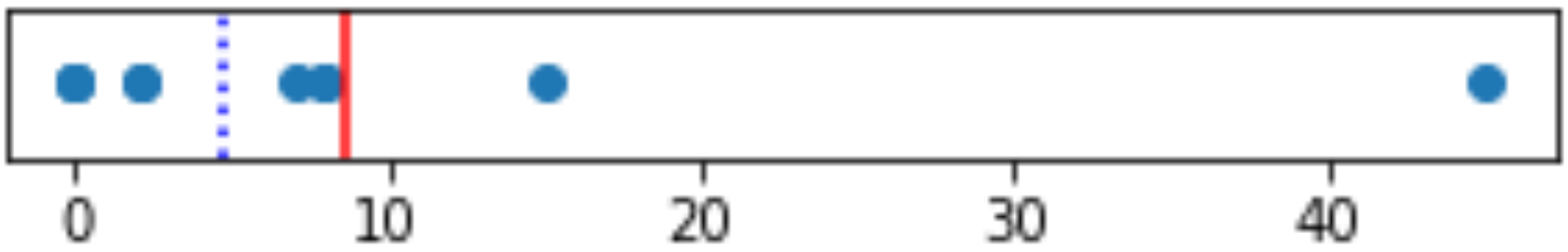
Note: np.std and np.var compute *population* std and var, not *sample*

- Change with `ddof=1`

Outliers

Outliers are particularly large or small values

Outliers draw the mean towards them! (also affect SD)



Median

What value divides upper half from lower half?

- Sort values
- Pick middle one
  - If even number: take mean of middle 2

0 0 0 2 **2** **7** 8 8 15 45

How to increase? Increase small values.

Pts
8
15
0
45
8
7
0
2
0
2
$\tilde{x} = 4.5$

Spread – Range and IQR

The **range** is max – min

The **inter-quartile range** is distance between 1st and 3rd quartiles (width of “middle 50%”)

0 0 **0** 2 2 | 7 8 **8** 15 45

Quick Summary

```
>>> movie_info['count'].describe()
count    59047.000000
mean       423.393144
std       2477.885821
min          1.000000
25%          2.000000
50%          6.000000
75%         36.000000
max      81491.000000
Name: count, dtype: float64
```

Mean, Median, and Skew

Mean works well for **centered** values

- No excessively large or small values
  - $\bar{x} \approx \tilde{x}$
- Critical for computation and prediction
- Total (and mean) deviation is 0

Median more *robust to outliers*

- Good for heavily-skewed data, large outliers
- Divides high/low, but limited for prediction

What question do we want to answer?

If we distributed the points equally, how many would each have?

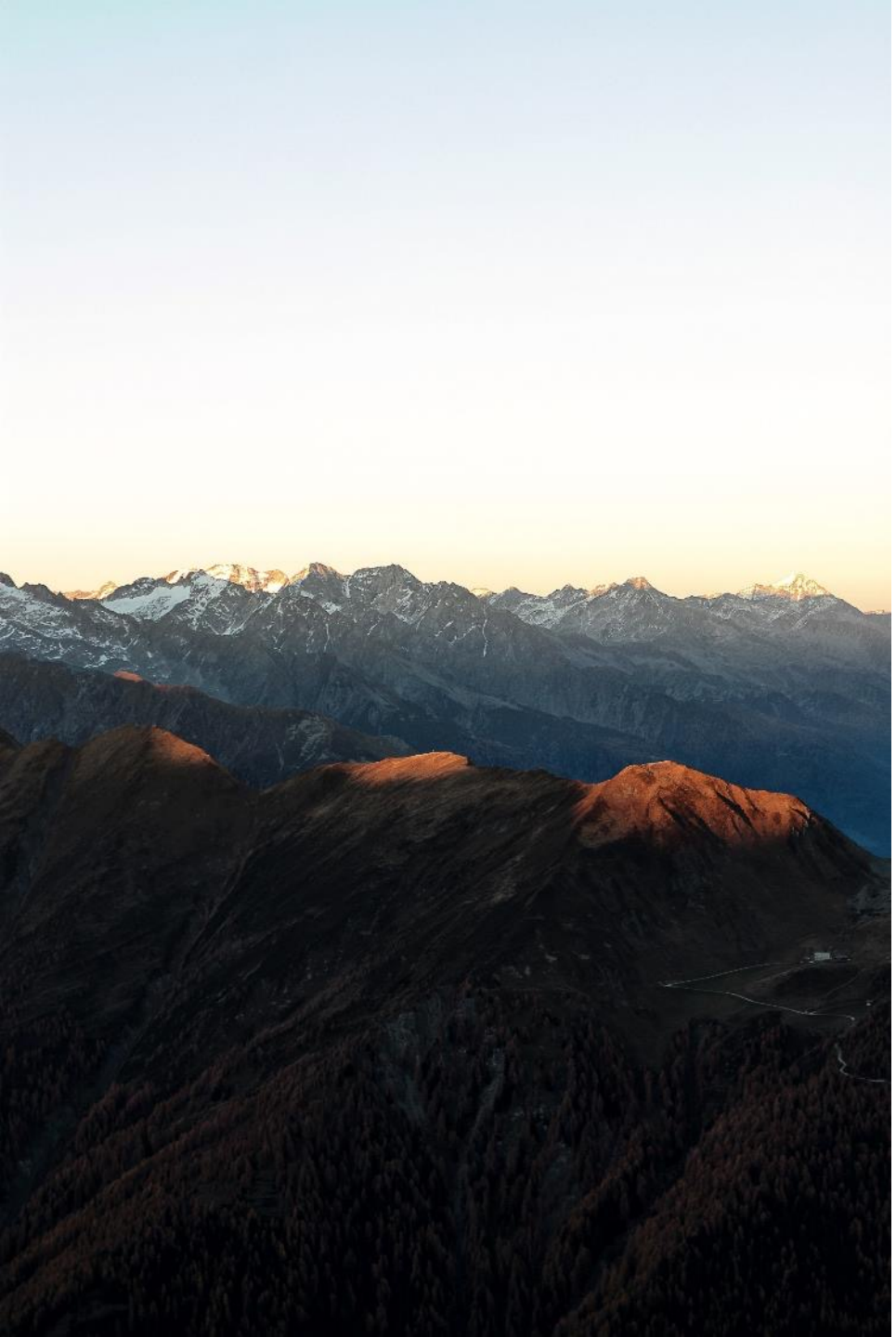
If we randomly selected a player, are they equally likely to have more or less points?

Mode

The most common value

- doesn’t work great for continuous values
- fantastic for categorical variables!

Pts
8
15
0
45
8
7
0
2
0
2
0



Wrapping Up

Mean and median describe where a value tends to be.

Standard deviation, variance, range, and IQR measure how spread out it is.

Mean-based computationally useful; median-based robust to outliers.

Photo by Brylee Hawkins on Unsplash

GROUPS AND AGGREGATES

Data Frame

In [8]:

```
ratings
```

Out[8]:

	userId	movieId	rating	timestamp
0	1	296	5.0	1147880044
1	1	306	3.5	1147868817
2	1	307	5.0	1147868828
3	1	665	5.0	1147878820
4	1	899	3.5	1147868510
...	...	...	...	...
25000090	162541	50872	4.5	1240953372
25000091	162541	55768	2.5	1240951998
25000092	162541	56176	2.0	1240950697
25000093	162541	58559	4.0	1240953434
25000094	162541	63876	5.0	1240952515

25000095 rows × 4 columns

Aggregates

What's the mean rating?

```
ratings['rating'].mean()
```

- Use data frame
- Select column `rating`
- Compute mean

Aggregate Functions

Many aggregates:

- `mean`
- `median`
- `mode`
- `min / max`
- `sum`
- `count`
- `std` (standard deviation)
- `var` (variance)

All are **methods** on pandas `Series`.

Count and Length

Total size of series, *including* missing values:

- `series.size`
- `len(series)`

Number of values, *not including* missing data:

- `series.count()`

Quantiles

`series.quantile(p)`

- $0 \leq p \leq 1$ (also written $p \in [0,1]$)
- Value p of the way from min to max in sorted order
  - 0 – min
  - 1 – max
  - 0.5 – median

Grouped Aggregates

Remember: # of ratings per movie could be a movie variable?

```
ratings.groupby('movieId')['rating'].count()
```

Yields a Series:

- indexed by movieId
- Values are # of ratings per movie
- We’re only aggregating one series
  - If we did the other way, we wouldn’t have a movieid to group by!

Sorting

Sometimes we want to sort data:

- `sort_values(col)` re-sorts the data

Sometimes we just need that to get the largest or smallest:

- `nlargest(n, col)` gets *n* rows with largest values of *col*
- `nsmallest(n, col)` gets the smallest

Joining

Joining two frames lets us combine data

- Simplest: join on common index
- `set_index` sets the index
- `join` joins two frames (by default on their indexes)
  - Can specify join columns

More later (including in readings)

Aggregating Aggregates

Our movie-level rating statistics are just more variables!

That can be aggregated.

Practice: compute the mean # of ratings per movie

Wrapping Up

Aggregates combine a series or array into a single value

Can compute over whole series, or over groups.

Join combines frames

Photo by Dwayne Legrand on Unsplash

DESCRIBING DISTRIBUTIONS

What Questions?

- What is the average value?
  - Mean, median
- How spread out is the data?
  - Standard deviation, IQR
- Is it skewed?
- What does it look like?

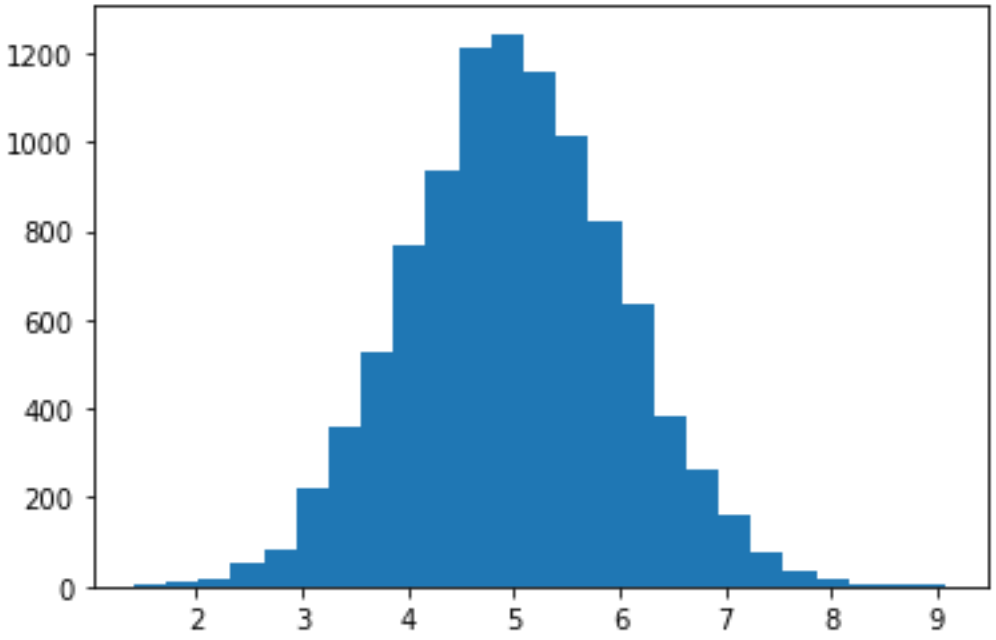
Numeric Descriptions

Previous video!

```
>>> numbers.describe()
count    10000.000000
mean         5.006208
std          0.998461
min          1.410392
25%          4.330468
50%          4.990636
75%          5.678746
max          9.081399
dtype: float64
```

Histograms

```
>>> plt.hist(numbers, bins=25)
```



This data is *symmetrical* (not skewed)

Shows how common different values are

- X-axis: values
- Y-axis: frequency (count or relative)

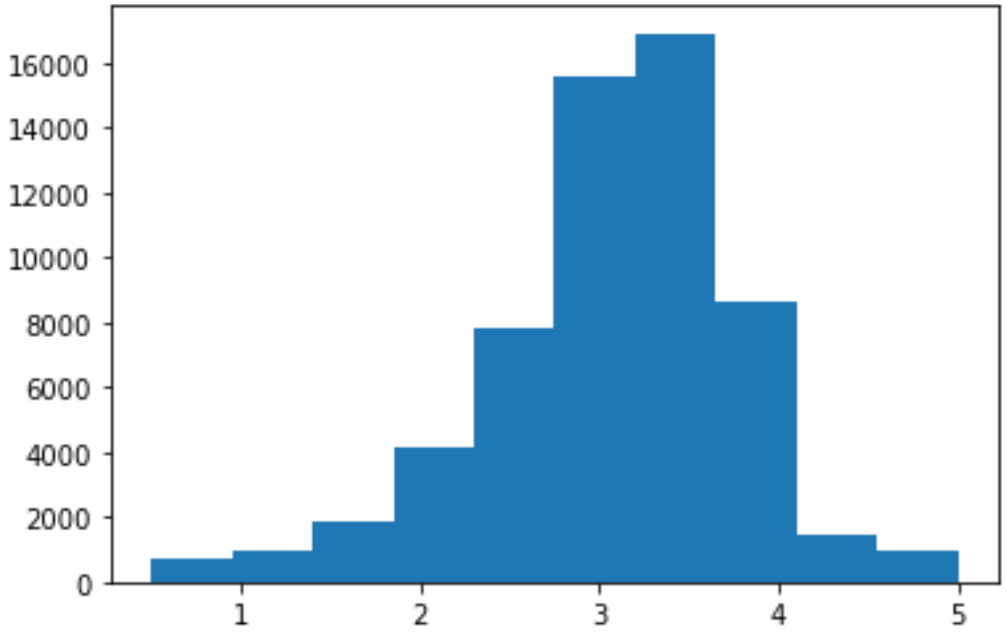
Bins controls division points
Pick for clarity and integrity!

```
import matplotlib.pyplot as plt
```

Real Data – Average Movie Rating

```
>>> movie_info['mean'].describe()
count    59047.000000
mean         3.071374
std          0.739840
min          0.500000
25%          2.687500
50%          3.150000
75%          3.500000
max          5.000000
Name: mean, dtype: float64
```

```
>>> plt.hist(movie_info['mean'])
```



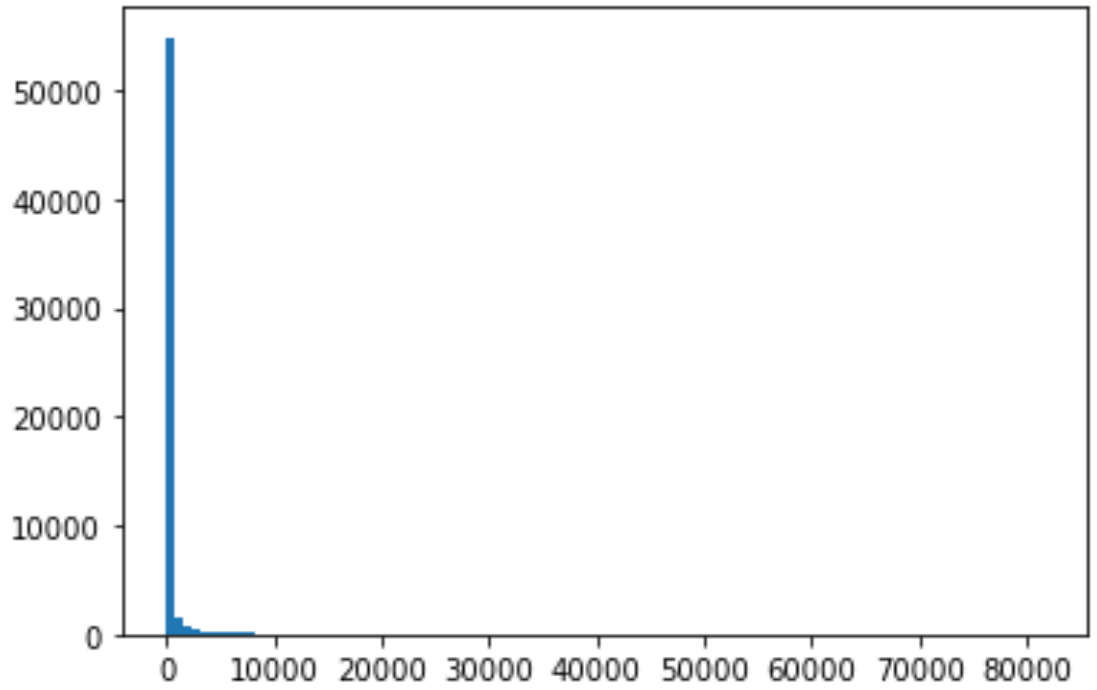
Slight left skew

- Mean less than median
- Longer tail on the left of the histogram

Real Data – Movie Rating Count

```
>>> movie_info['count'].describe()
count    59047.000000
mean       423.393144
std       2477.885821
min          1.000000
25%          2.000000
50%          6.000000
75%         36.000000
max      81491.000000
Name: count, dtype: float64
```

```
>>> plt.hist(movie_info['mean’], bins=100)
```



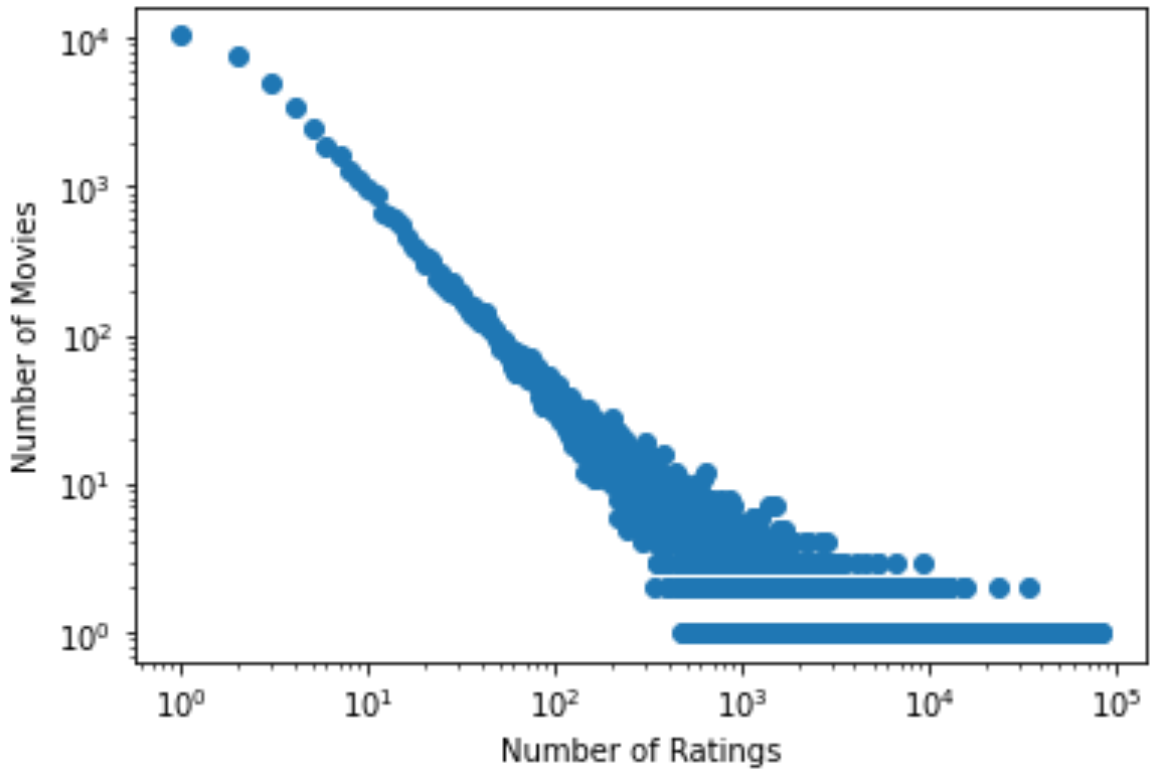
Very strong right skew

- Mean much greater than median
- Most movies have far fewer ratings than mean!
- Hard to really see in a histogram

Alternate Histogram

```
hist = movie_info['count'].value_counts()
plt.scatter(hist.index, hist)
plt.yscale('log')
plt.ylabel('Number of Movies’)
plt.xscale('log')
plt.xlabel('Number of Ratings’)
```

- Points rather than bars
- One point per rating count
- Plotted on logarithmic axis
- The mode is 1 (10^0)



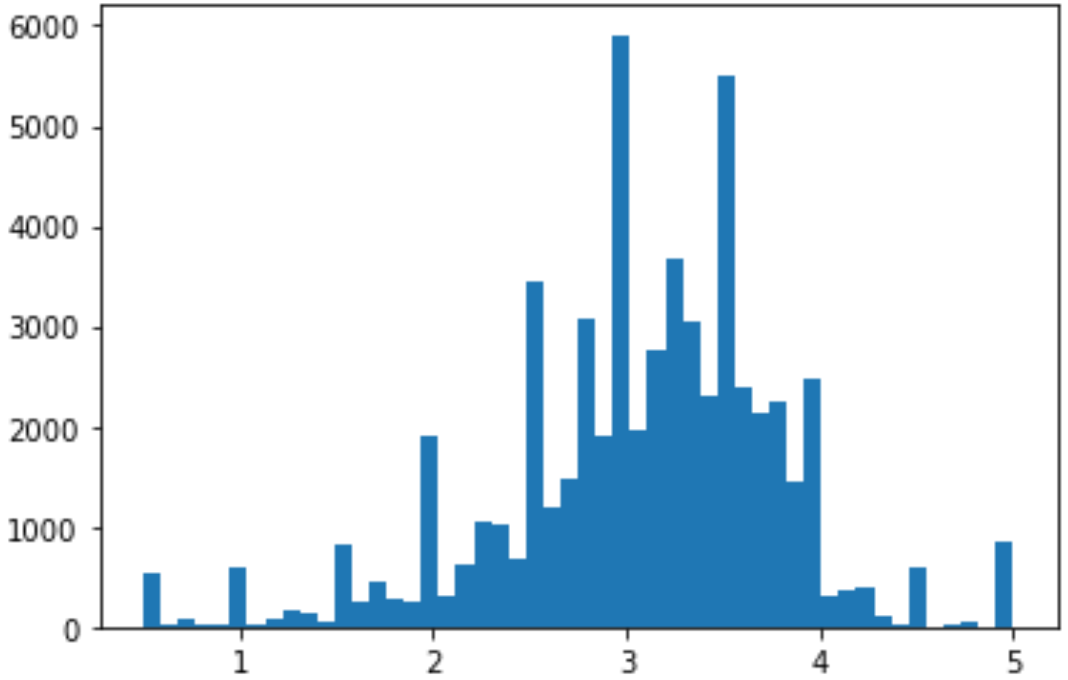
Power law distribution (almost)

Artifact

Movie mean rating, more bins

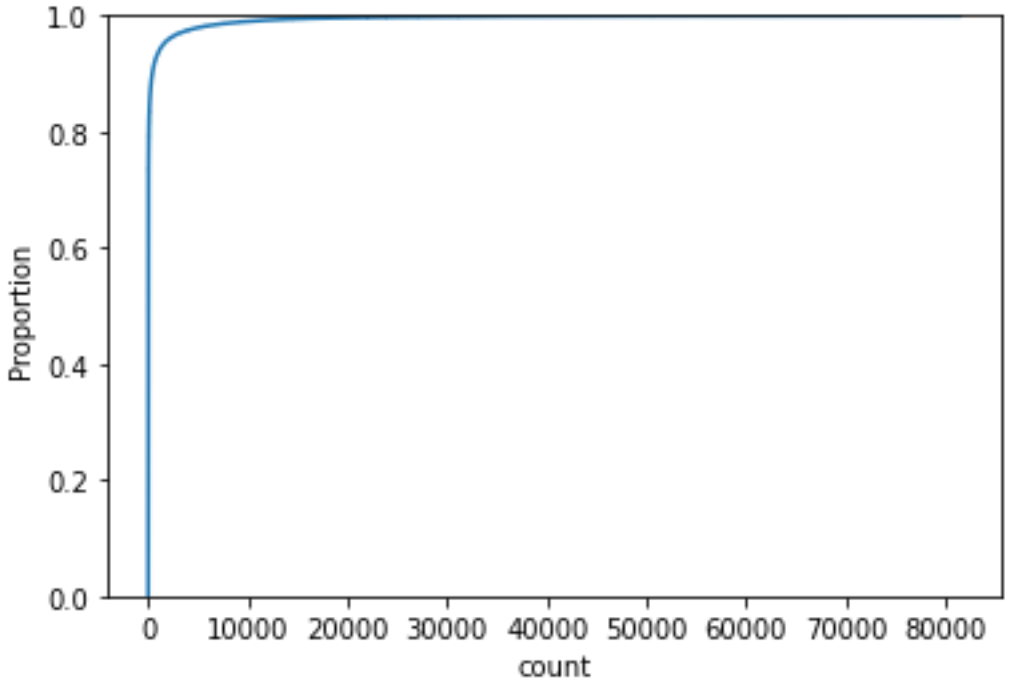
- Certain values more common!
  - 1, 2, 3, 4, 5, 2.5, 3.5…
- These are exact rating values
  - Because so many movies have only one rating!

```
>>> plt.hist(movie_info['mean'], bins=50)
```



Cumulative Distribution

```
>>> sns.ecdfplot(movie_info['count'])
```



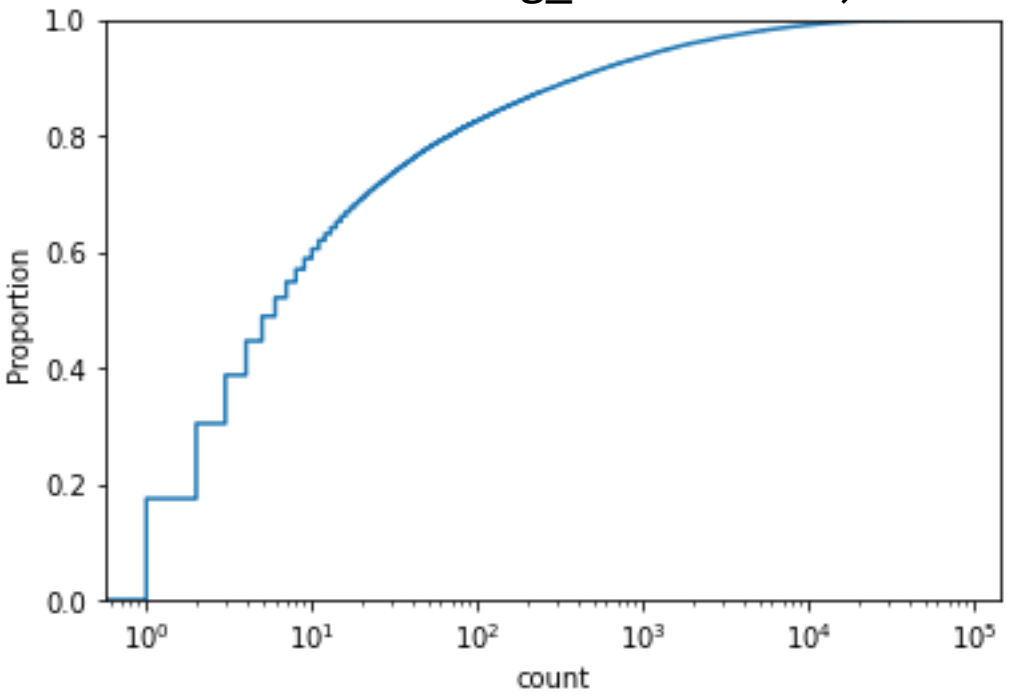
Shows how observations accumulate as we go through the range of values.

This data is super skewed…

```
import seaborn as sns
```

Cumulative Distribution

```
>>> sns.ecdfplot(movie_info['count'],
                 log_scale=True)
```



Shows how observations accumulate as we go through the range of values.

Log scaling: accumulation as we increase order of magnitude.

Approximately ½ of movies have 10 or fewer ratings.

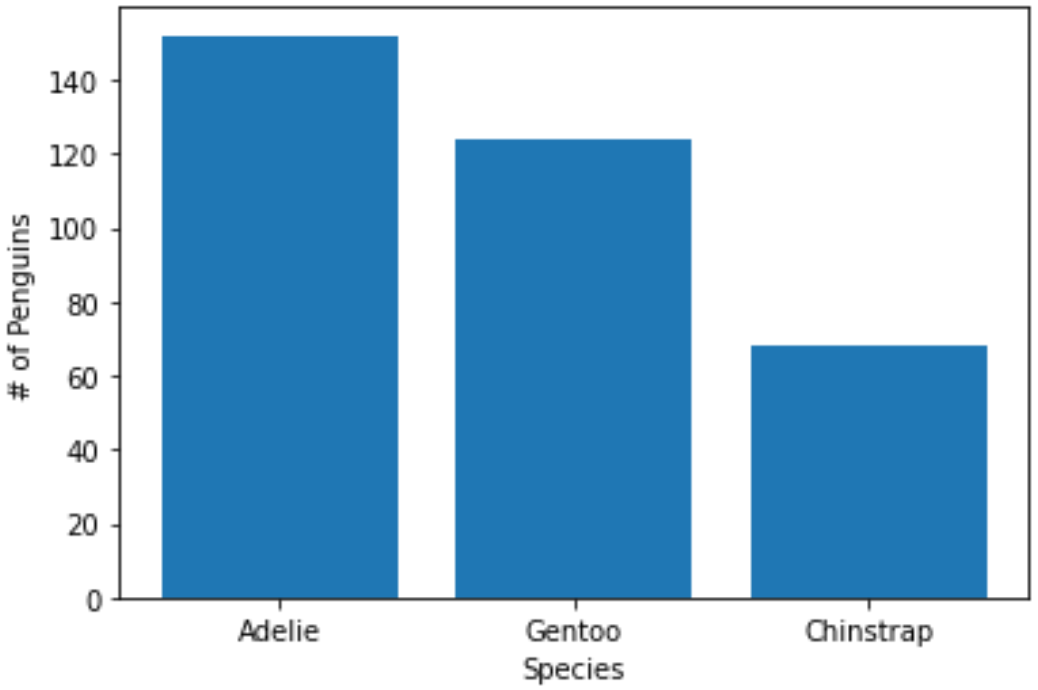
- Useful for seeing skewed data
- Useful for comparing distributions
- Obscures mean, median, etc.

Categorical Distributions

Categorical distributions with *bar charts*

```
spec_counts = penguins['species'].value_counts()
plt.bar(spec_counts.index, spec_counts)
plt.xlabel('Species’)
plt.ylabel('# of Penguins’)
```

- Adelie is most common



Wrapping Up

- In this section, we learn how to describe and visualize a distribution.

Photo by Leohoho on Unsplash

SOURCES AND BIAS

Estimating

Goal: **estimate** the value of a **parameter**

- True value in the world or population
- Estimated by a *statistic*

Example: estimate approval of our company

- Parameter: net approval (% of people who have a positive opinion)
- Statistic: % of a sample of people who say they have a positive opinion
- Goal: statistic is approximately parameter

A Few Sources of Bias

Selection bias

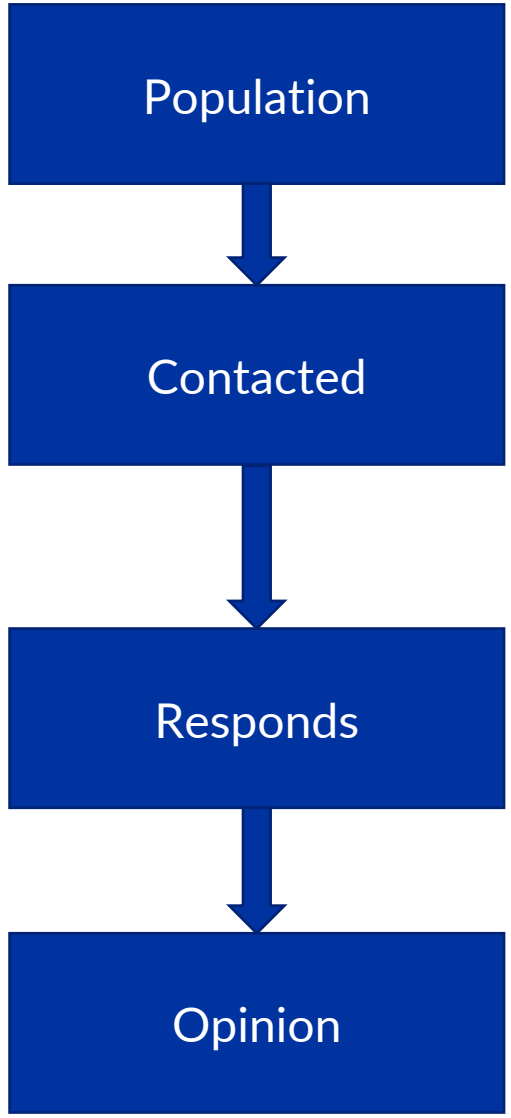
- Some people more likely to be contacted

Response bias

- Some people are more likely to respond

Measurement bias

- Measurement skews one way or another



Non-Uniform Bias

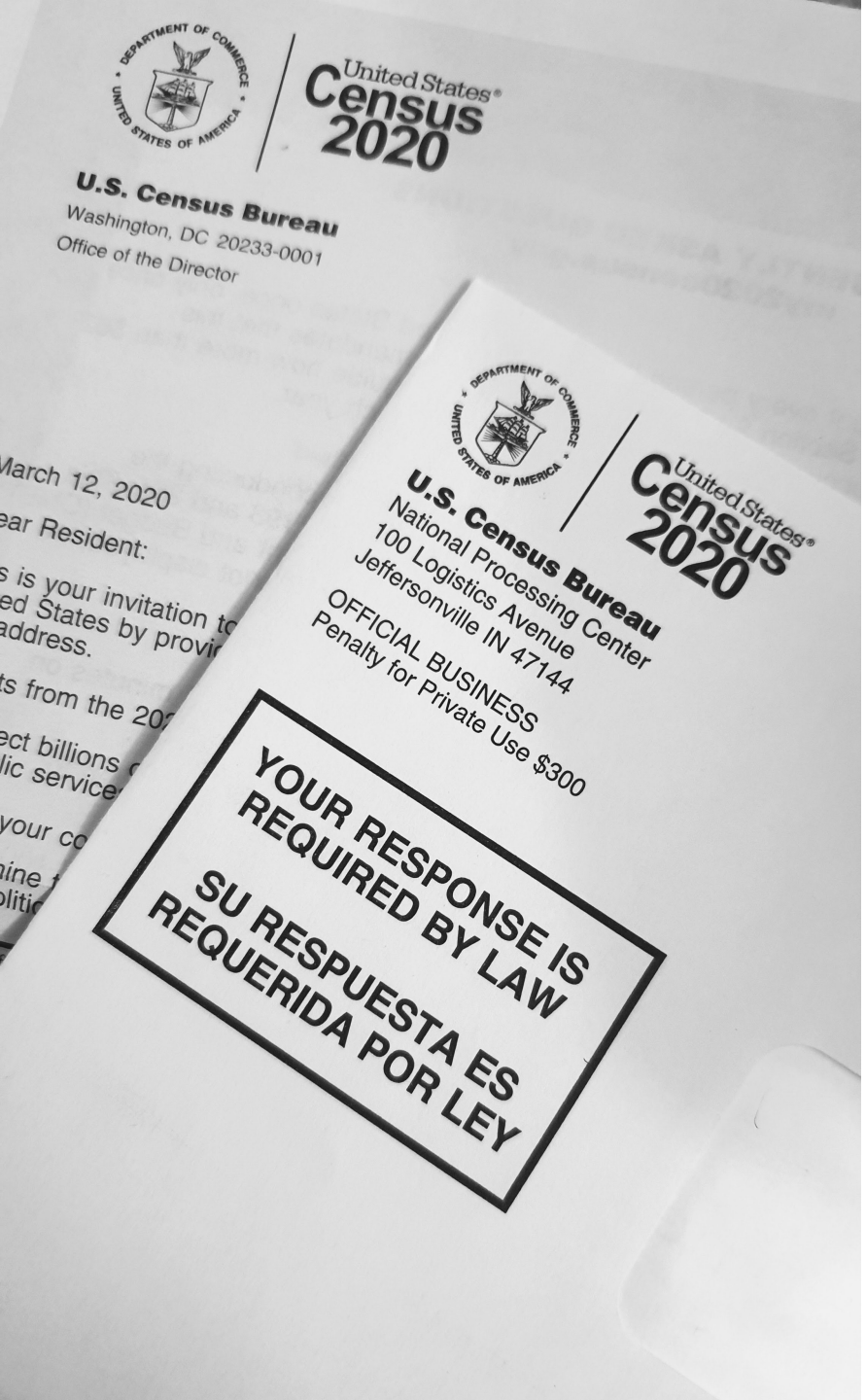
Bias may not affect all groups equally

- Under-representation
- Over-representation
- Measurement skews
  - E.g. standardized test scores reflect socio-economic status

Documenting

Clearly and **fully** document data collection process!

- See Datasheets for Datasets reading
- Enables further and future analysis of collection process!
- Helps future users assess if biases affect their problem



Wrapping Up

Goal is for data to accurately reflect population, and statistics to approximate parameters.

Various sources of bias can impede this.

Photo by Enayet Raheem on Unsplash