GENERAL ASSEMBLY

STATISTICS FUNDAMENTALS

Stefan Jansen

DAT-NYC

LEARNING OBJECTIVES

- Use NumPy and Pandas to analyze datasets using basic summary statistics: mean, median, mode, max, min, quartile, inter-quartile range, variance, standard deviation, and correlation
- Create data visualizations - including: line graphs, box plots, and histograms- to discern characteristics and trends in a dataset
- Identify a normal distribution within a dataset using summary statistics and visualization
- ID variable types and complete dummy coding by hand

OPENING

STATISTICS FUNDAMENTALS

LET'S REVIEW THE DATA SCIENCE WORKFLOW

The steps:

1. Identify the problem
2. Acquire the data
3. Parse the data
4. Mine the data
5. Refine the data
6. Build a data model
7. Present the results

TODAY

- We’re going to begin to talk about step 3:
  - Parsing the Data

- We’ll talk about the fundamentals of Statistics
  - Probability & Frequency Distributions
  - Descriptive Statistics of Distributions

- We’ll take another look at data types and introduce categorical variables

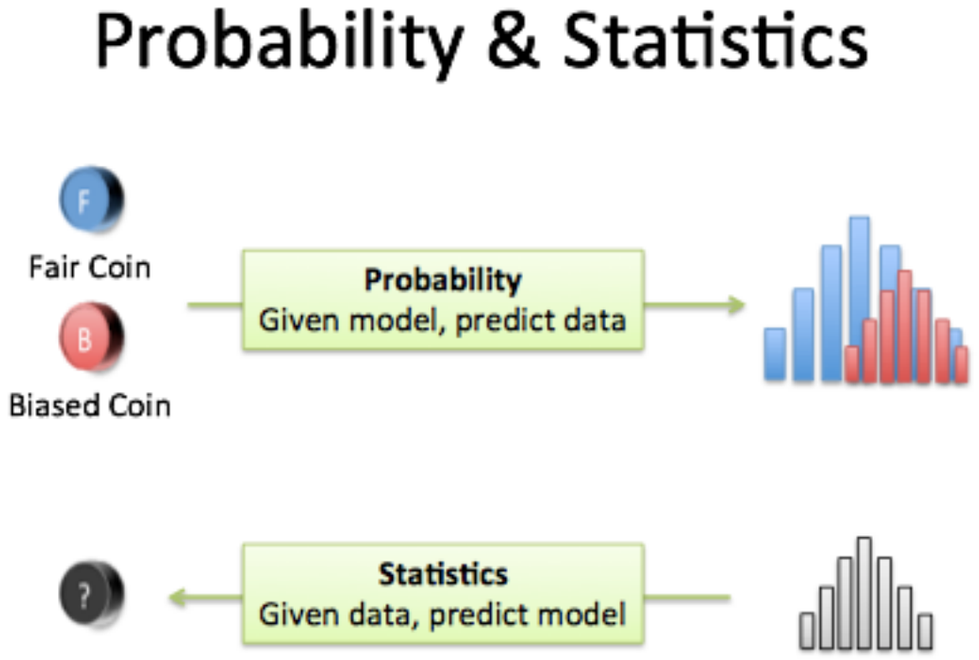
LAYING THE GROUND WORK

WHY STATISTICS?

Statistics is the science of **learning from data**, and of measuring, controlling, and communicating **uncertainty**; and it thereby provides the navigation essential for controlling the course of scientific and societal advances.

Marie Davidian and Thomas A. Louis, Science, 2012

- **Probability** is starting with an animal, and figuring out what footprints it will make.
- **Statistics** is seeing a footprint, and guessing the animal.



WHY STATISTICS?

- Data Scientists use data to **learn – infer and predict** - facts about the real world.
- Data are usually samples, i.e. they only provide a partial perspective on the subject of interest, and the perspective may be biased (as opposed to representative).
- The partial information obtained from a sample is uncertain, and so are the conclusions.
- Before we start building models to make predictions, we need to characterize the empirical, i.e. observed, distribution of the data for each variable, as well as the relationships between them.

SEVERAL WAYS TO CHARACTERIZE DISTRIBUTIONS

Central Tendency

- Mean
- Median
- Mode

Dispersion

- Max
- Min
- Quartile
- Interquartile Range
- Variance
- Standard Deviation
- Coefficient of Variation

Relationships

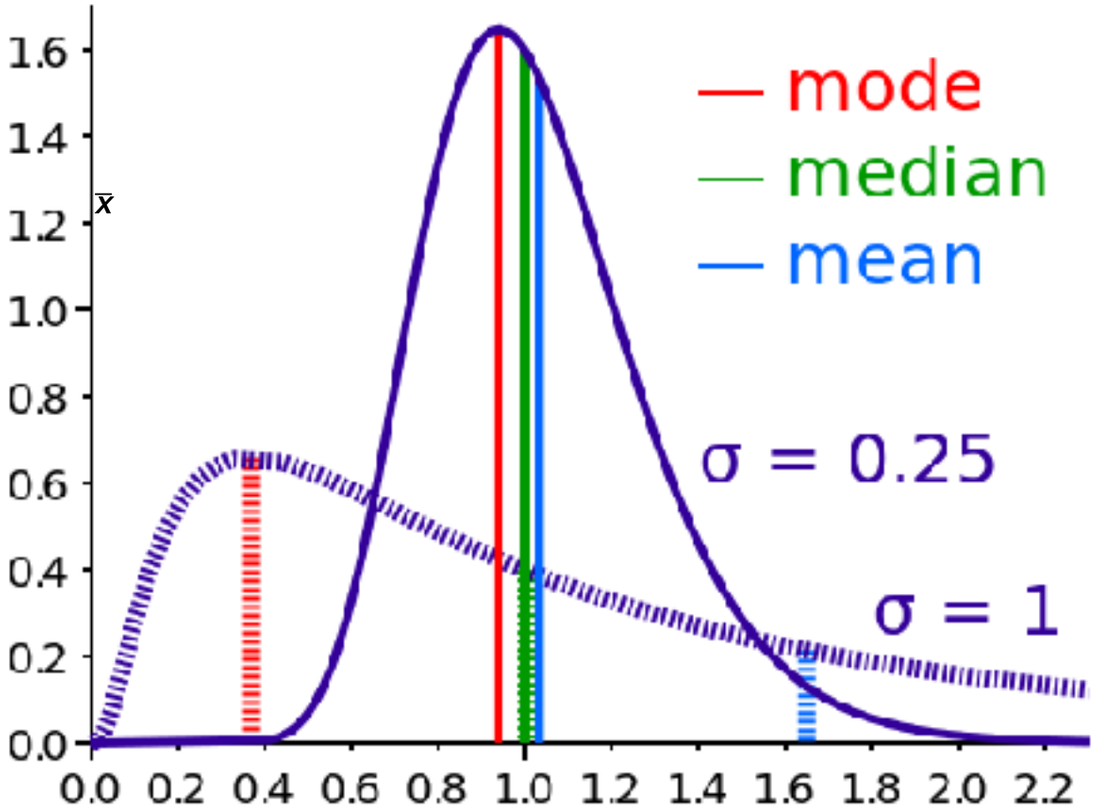
- Correlation

MEAN

- The mean of a set of values is the sum of the values divided by the number of values. It is also called the sample average.

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

N = # observations



Note: population, as opposed to sample concepts are referred to by Greek symbols – for the mean, it's μ.

MEAN EXAMPLE

- Find the mean of 19, 13, 15, 25, and 18.

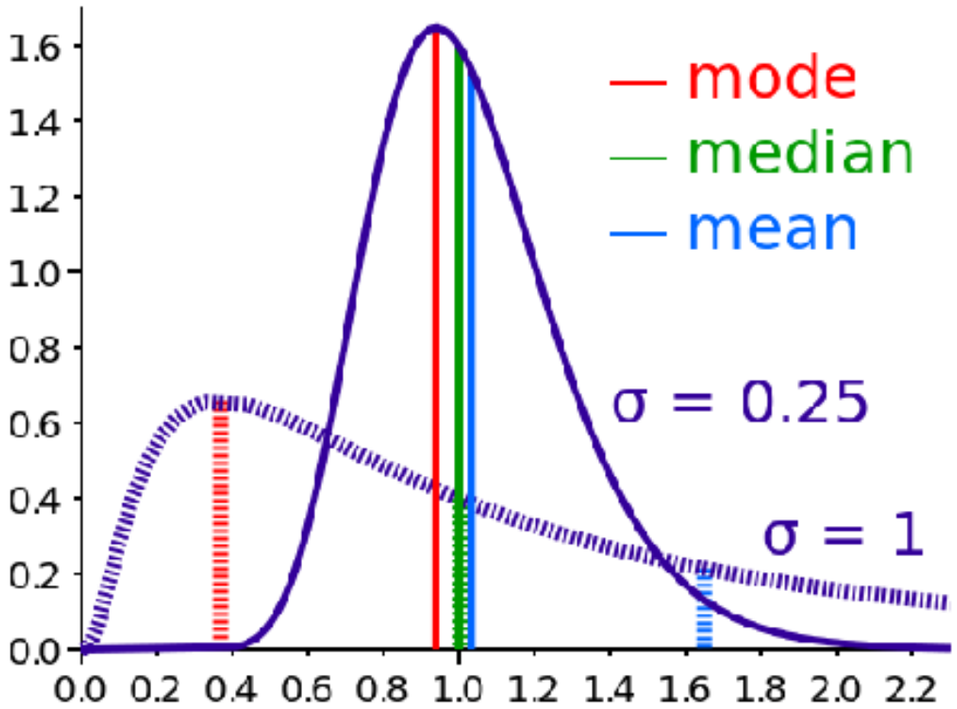
MEAN EXAMPLE

- Find the mean of 19, 13, 15, 25, and 18.

$$\frac{19 + 13 + 15 + 25 + 18}{5} = \frac{90}{5} = 18$$

MEDIAN

- The median refers to the midpoint in a series of numbers.
- To find the median
  - Arrange the numbers in order smallest to largest.
  - If there is an odd number of values, the middle value is the median.
  - If there is an even number of values, the average of the middle two values is the median.



MEDIAN EXAMPLE

- Find the median of 19, 29, 36, 15, and 20.

MEDIAN EXAMPLE

- Find the median of 19, 29, 36, 15, and 20.

Ordered Values:

15, 19, 20, 29, 36

20 is the median

MEDIAN EXAMPLE

- Find the median of 67, 28, 92, 37, 81, 75.

MEDIAN EXAMPLE

- Find the median of 67, 28, 92, 37, 81, 75.

Ordered Values:

28, 37, 67, 75, 81, 92

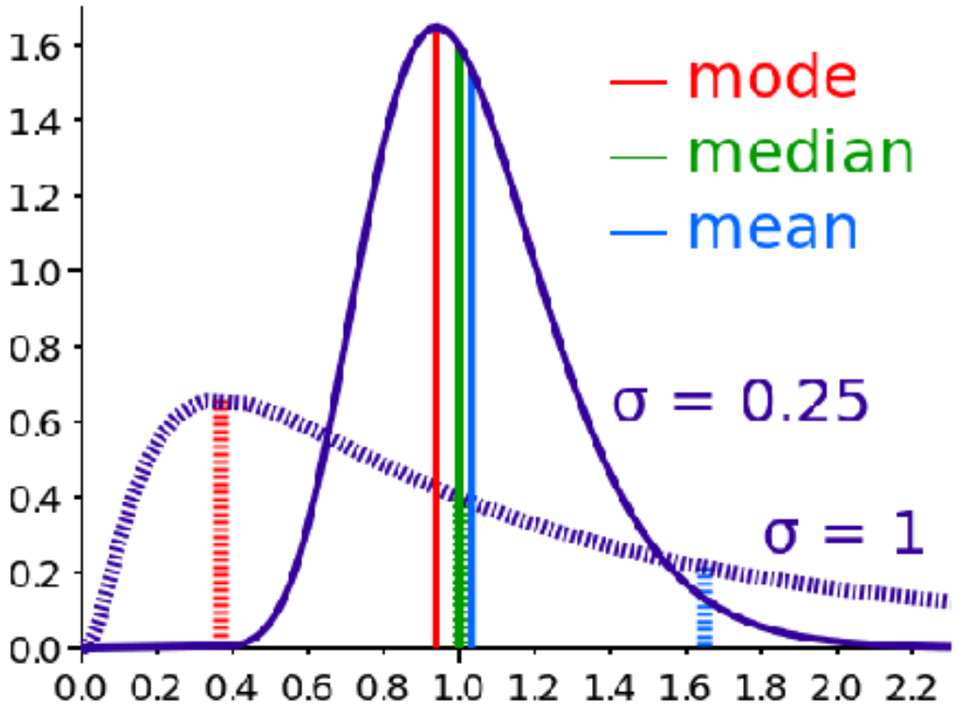
67 and 75 are the middle values.

$$\frac{67 + 75}{2} = \frac{142}{2} = 71$$

71 is the median.

MODE

- The mode of a set of values is the value that occurs most often.
- A set of values may have more than one mode or no mode.

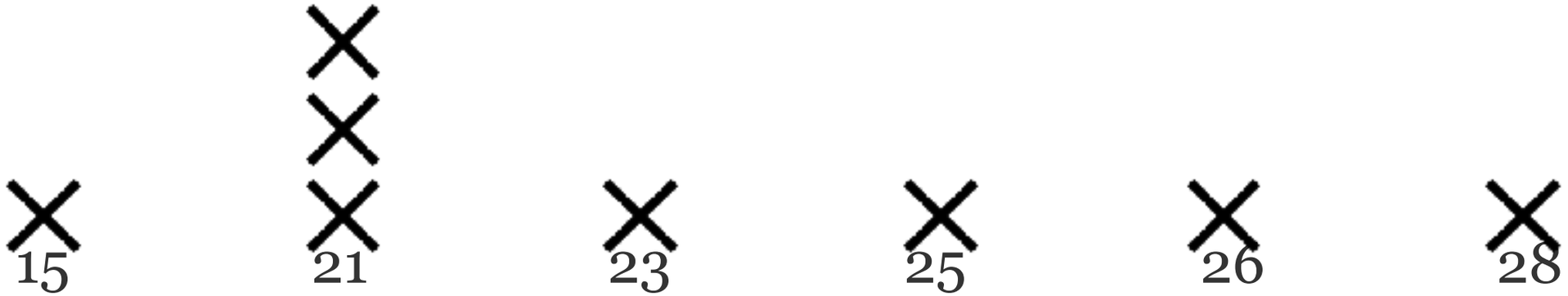


MODE EXAMPLE

- Find the mode of 15, 21, 26, 25, 21, 23, 28, and 21.

MODE EXAMPLE

- Find the mode of 15, 21, 26, 25, 21, 23, 28, and 21.



21 is the mode because it occurs most frequently

MODE EXAMPLE

- Find the mode of 12, 15, 18, 26, 15, 9, 12, and 27.

MODE EXAMPLE

- Find the mode of 12, 15, 18, 26, 15, 9, 12, and 27.

12 and 15 are the modes since the both occur twice.

MODE EXAMPLE

- Find the mode of 4, 8, 15, 21, and 23.

MODE EXAMPLE

- Find the mode of 4, 8, 15, 21, and 23.

There is no mode since all values occur the same number of times.

ACTIVITY: KNOWLEDGE CHECK

ANSWER THE FOLLOWING QUESTIONS (5 minutes)

1. For the following groups of numbers, calculate the mean, median and mode by hand. Also determine the min and max.

   a. 18, 24, 17, 21, 24, 16, 29, 18

   b. 75, 87, 49, 68, 75, 84, 98, 92

   c. 55, 47, 38, 66, 56, 64, 44, 39

DELIVERABLE

Answers to the above questions

CODEALONG

SUMMARY STATISTICS IN PANDAS

CODEALONG: SUMMARY STATISTICS IN PANDAS

- Open the starter-code notebook in the class repo at:
    - /lesson-03/code/stats-code-along-class-03.ipynb

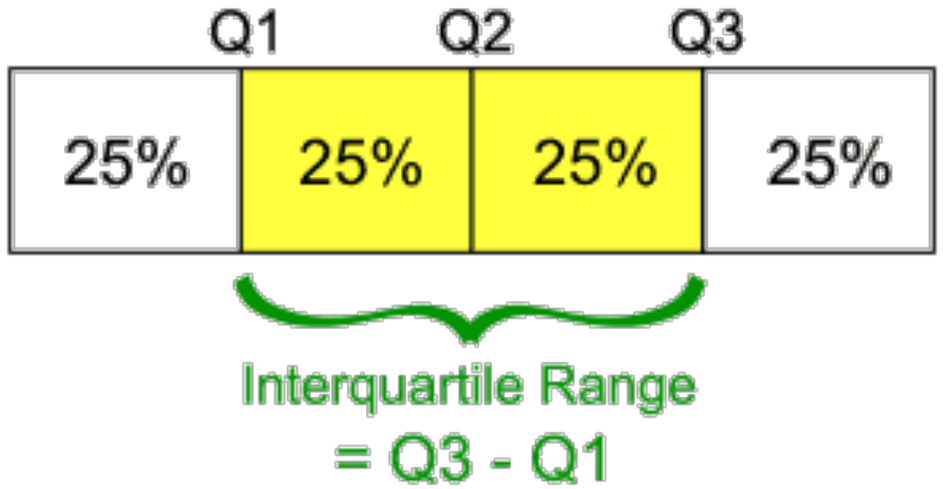
CODEALONG PART 1: BASIC STATS

- We can use Pandas to calculate the mean, median, mode, min, and max.

```
Methods available include:
.min() - Compute minimum value
.max() - Compute maximum value
.mean() - Compute mean value
.median() - Compute median value
.mode() - Compute mode value
.count() - Count the number of observations
```

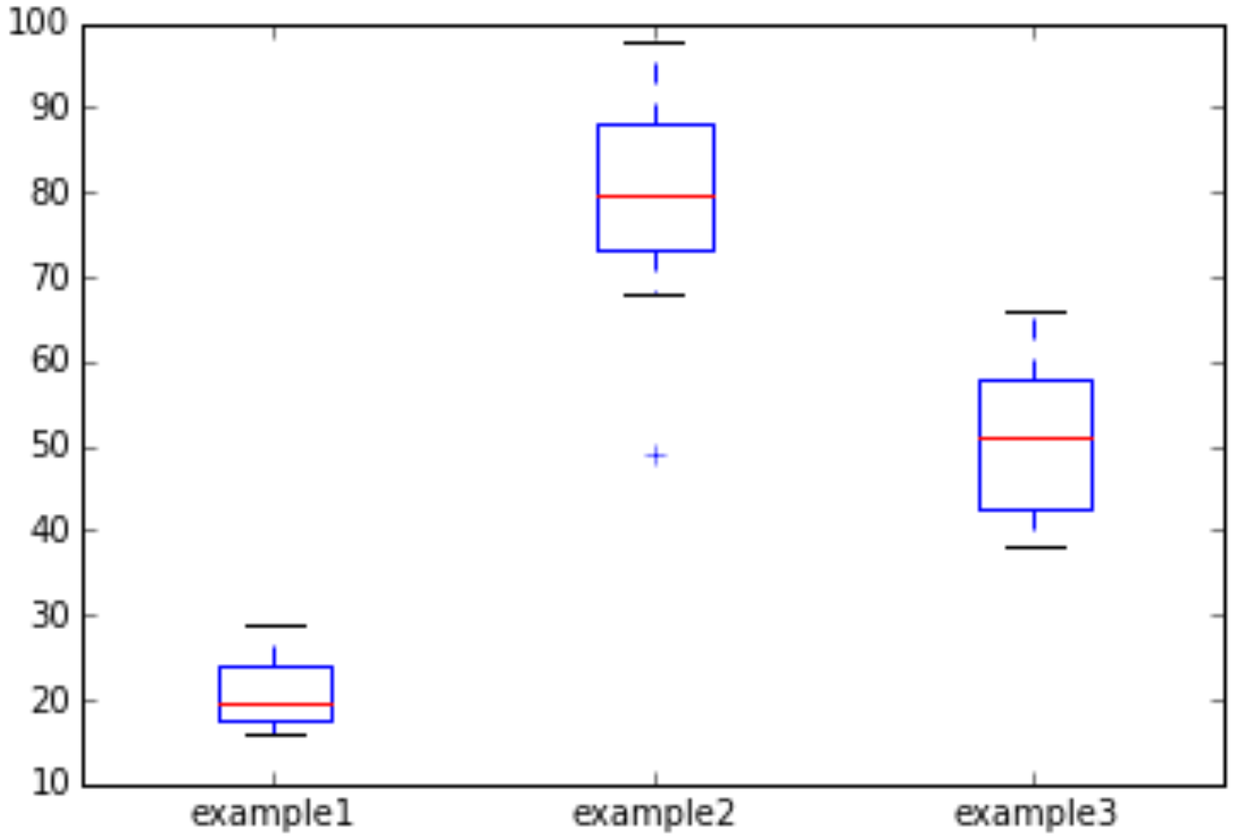
QUARTILES AND INTERQUARTILE RANGE

- Quartiles divide a rank-ordered data set into four equal parts.
- The values that divide each part are called first, second, and third quartiles, denoted Q1, Q2, and Q3, respectively.
- The interquartile range (IQR) is Q3 - Q1, a measure of variability.



CODEALONG PART 2: BOX PLOT

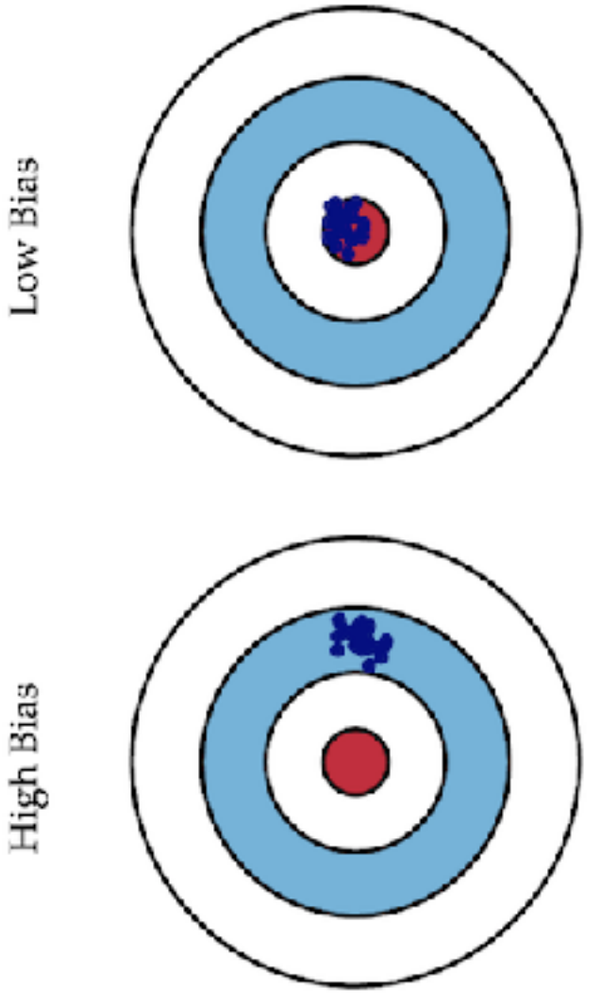
- Box plots give a nice visual of min, max, mean, median, and the quartile and interquartile range.



TWO SOURCES OF ERROR: BIAS VS. VARIANCE

- Error due to **bias** is calculated as the difference between the *expected prediction* of our model and the *correct value* we are trying to predict.

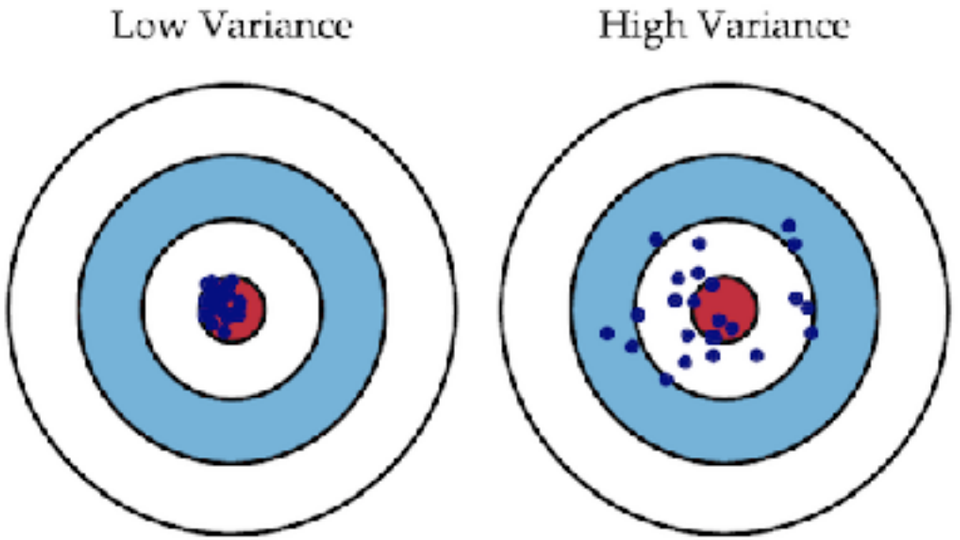
- Imagine creating multiple models on various datasets. **Bias** measures *how far off in general* models' predictions are from the correct value.



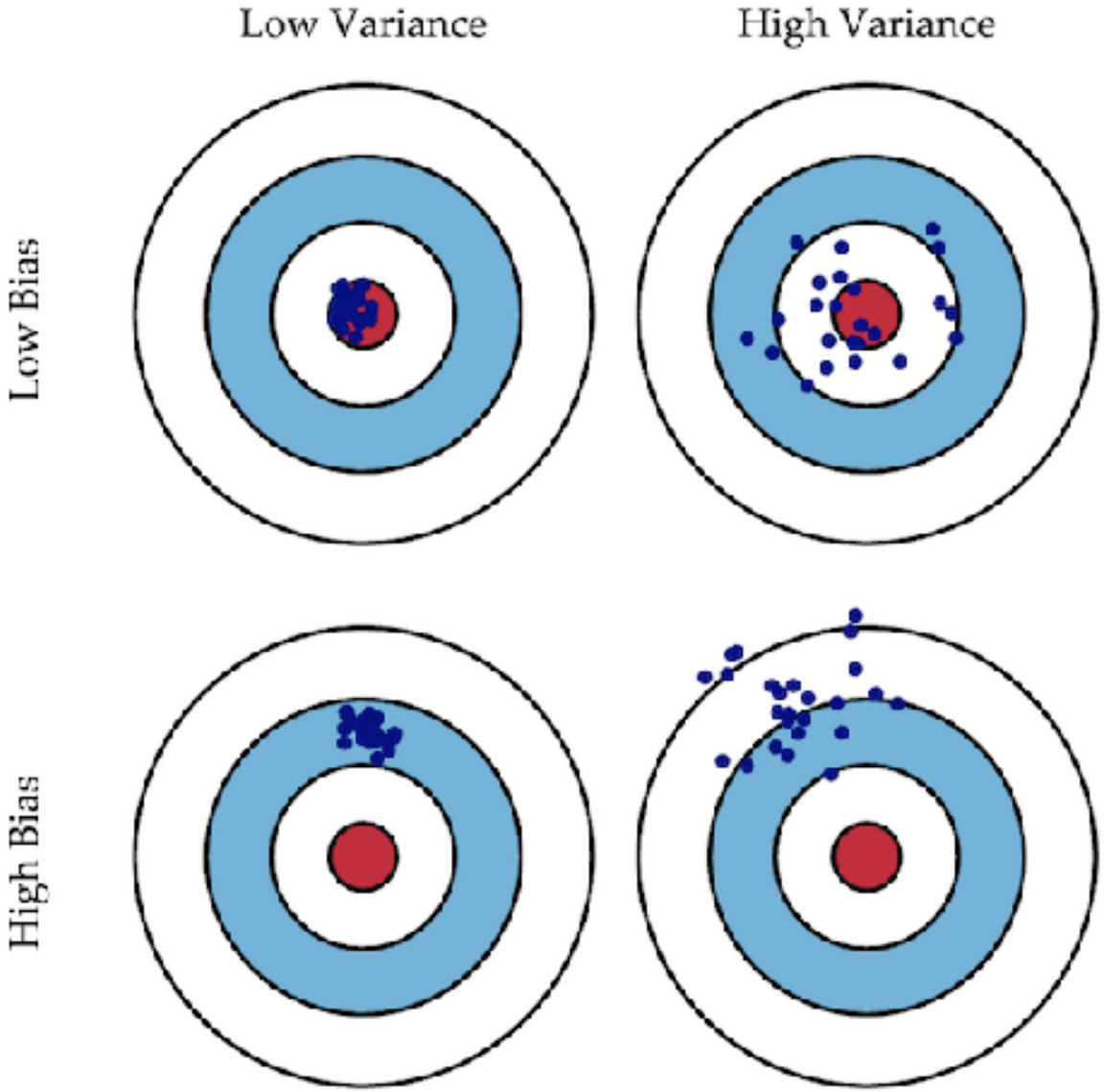
BIAS VS. VARIANCE

- Error due to **variance** is taken as the variability of a model prediction for a given point, i.e., how sensitive predictions are to the sample data.

- Imagine creating multiple models on various datasets. The **variance** is *how much the predictions for a given point vary* between different realizations of the model.



BIAS VS. VARIANCE



For more detail, see: http://scott.fortmann-roe.com/docs/BiasVariance.html

STANDARD DEVIATION

- Standard deviation (SD, σ for population, s for sample) is a measure that is used to quantify the amount of variation or dispersion of a set of data values.

- Standard deviation is the square root of variance.

$$S = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \bar{x})^2}{N-1}}$$

Why dividing by N-1?

Using the mean to calculate SD costs one degree of freedom!

- The sum of all deviations from the mean is always zero.
- So if we know both the mean and any n-1 values, the n-th value is determined, and cannot be freely ‘chosen’ (or vary randomly).

STANDARD ERROR

- **Test statistics** are metrics calculated for a sample, like the mean. Statisticians derive sampling distributions for test statistics.

- The **standard deviation** of the **sampling distribution** of a test statistic is called the Standard Error of this test statistic.

- The **standard error of the mean** (SEM) quantifies the precision, or reliability, of our estimate of the population mean.

- In other words, it is a **measure of how far your sample mean is likely to be from the true population mean**.

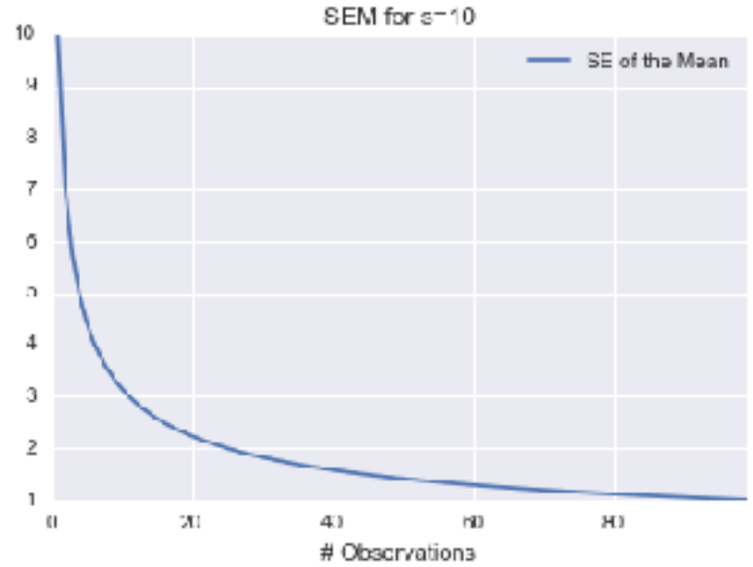
STANDARD ERROR

- The Standard Error of the Mean increases with the standard deviation of the underlying data, and decreases with the square root of the sample size.

- In other words, the population mean estimate will become more reliable as the sample size increases – assuming the sample standard deviation remains constant.

- In practical terms, decreasing the uncertainty of a mean estimate 2x requires 4x as many observations in the sample. Decreasing SE 10x requires 100x as many observations.

$$SE_{\bar{x}} = \frac{s}{\sqrt{N}}$$



CODEALONG PART 3: STANDARD DEVIATION & VARIANCE

- You can calculate variance and standard deviation easily in Pandas.

```
Methods include:
.std() - Compute Standard Deviation
.var() - Compute variance
.describe() - short cut that prints out count, mean, std, min,
quartiles, max
```

CORRELATION

- The correlation coefficient measures the

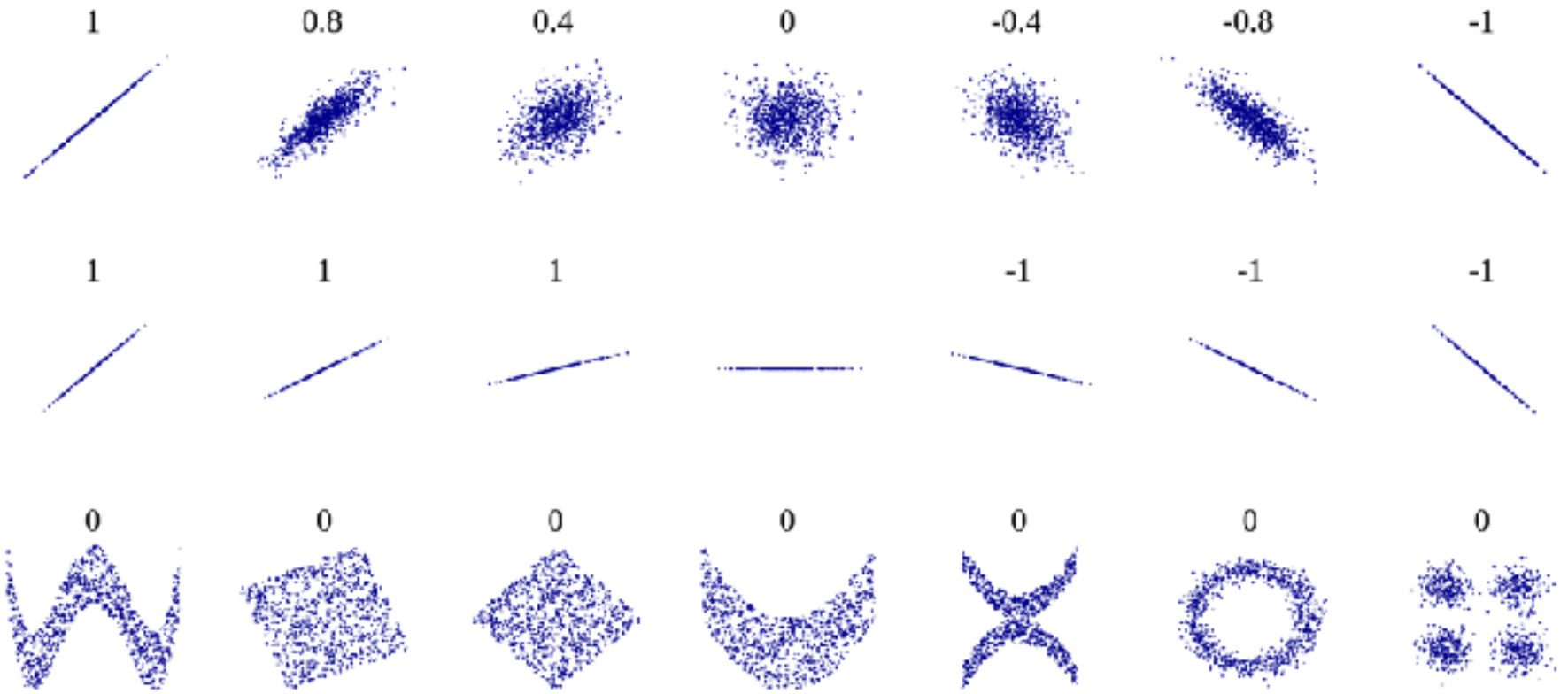
STRENGTH OF A LINEAR RELATIONSHIP

- among two variables.

$$r_{xy} = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{s_x s_y}$$

The sample correlation coefficient is a standardized version (divided by the respective standard deviations) of the sample covariance.

EXAMPLE CORRELATION VALUES



The correlation reflects

- the noisiness and direction of a linear relationship (top row),
- but not the slope of that relationship (middle),
- nor many aspects of nonlinear relationships (bottom).

CONTEXT

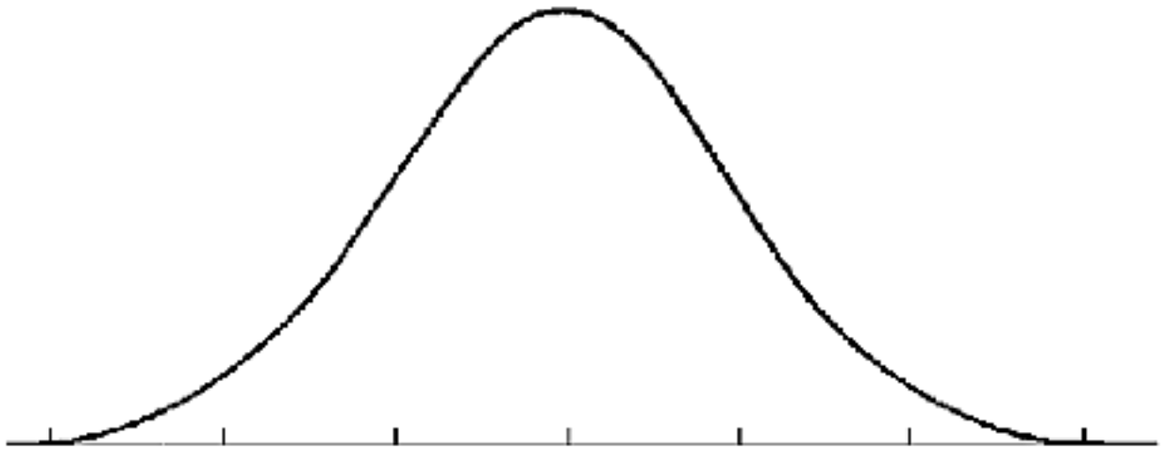
- For most projects, descriptive stats will come first. These help you get to know your dataset better.

- Sometimes, descriptive stats may be all you need to answer your question.

IS THIS NORMAL?

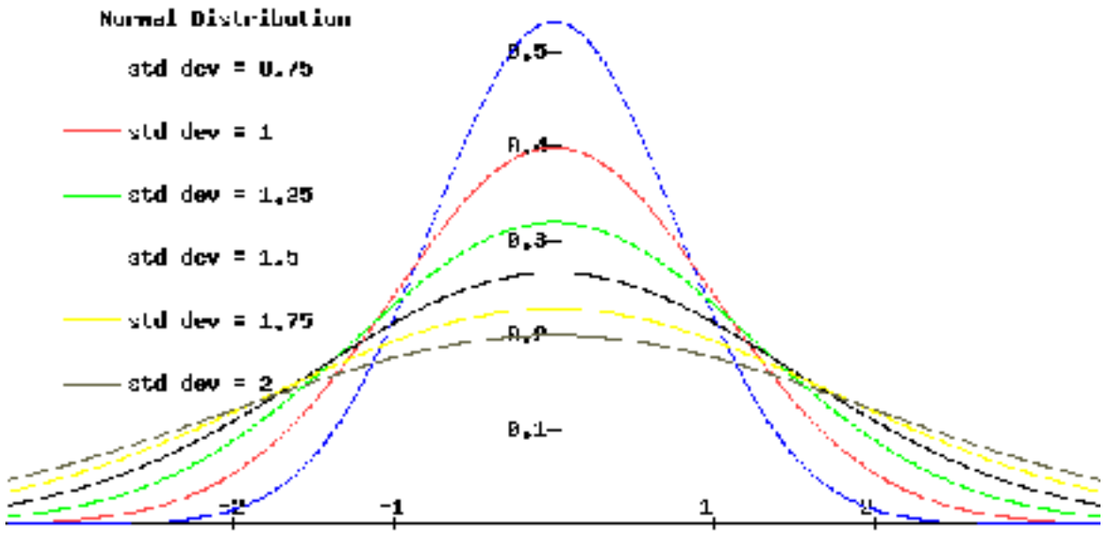
THE NORMAL DISTRIBUTION

- A normal distribution is often a key assumption to many models.
- The normal distribution depends ONLY on *mean* and *standard deviation*.
- The *mean* determines the center of the distribution. The *standard deviation* determines the height and width of the distribution.

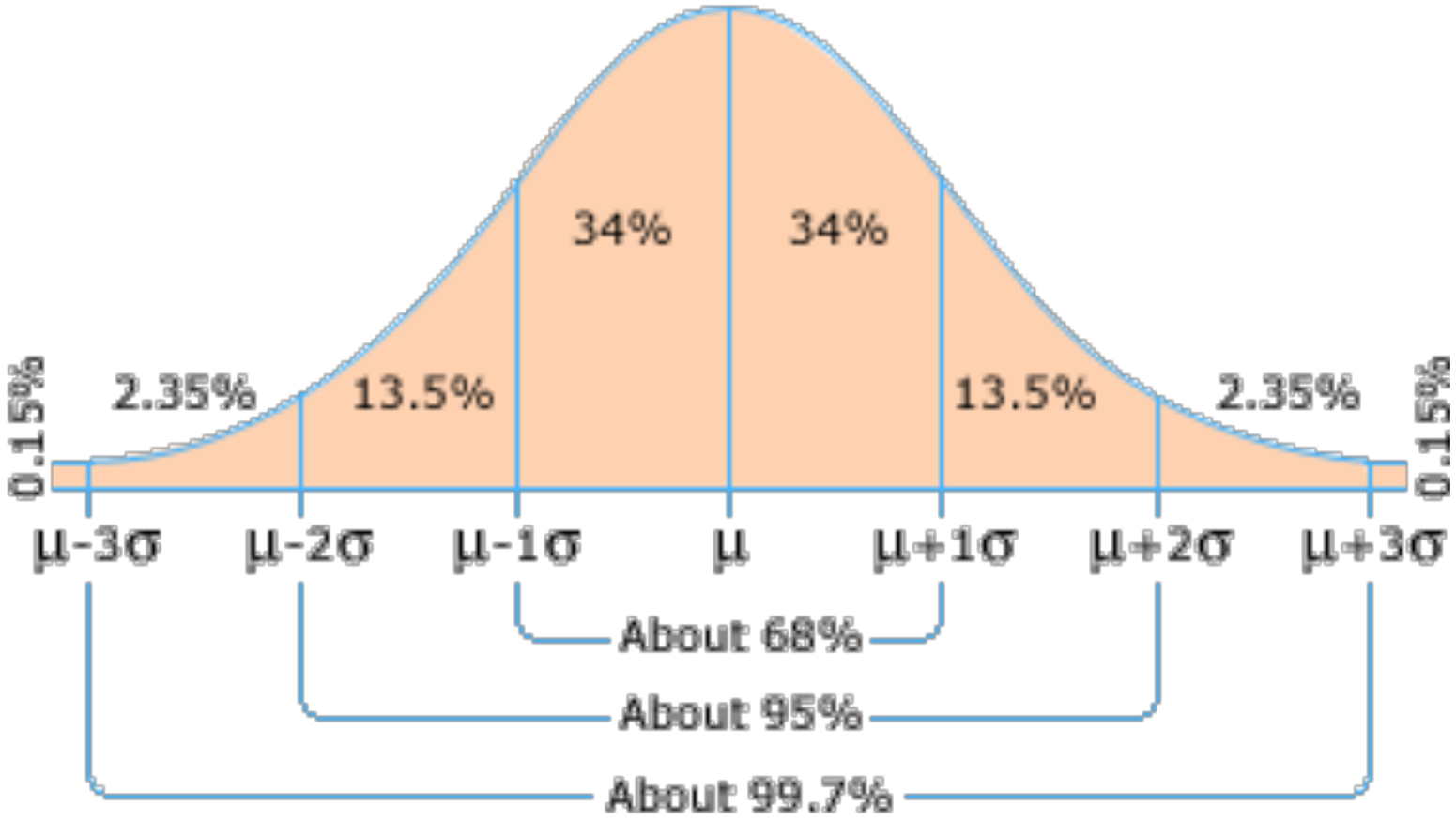


THE NORMAL DISTRIBUTION

- Normal distributions are symmetric, bell-shaped curves.
- When the standard deviation is large, the curve is short and wide.
- When the standard deviation is small, the curve is tall and narrow.



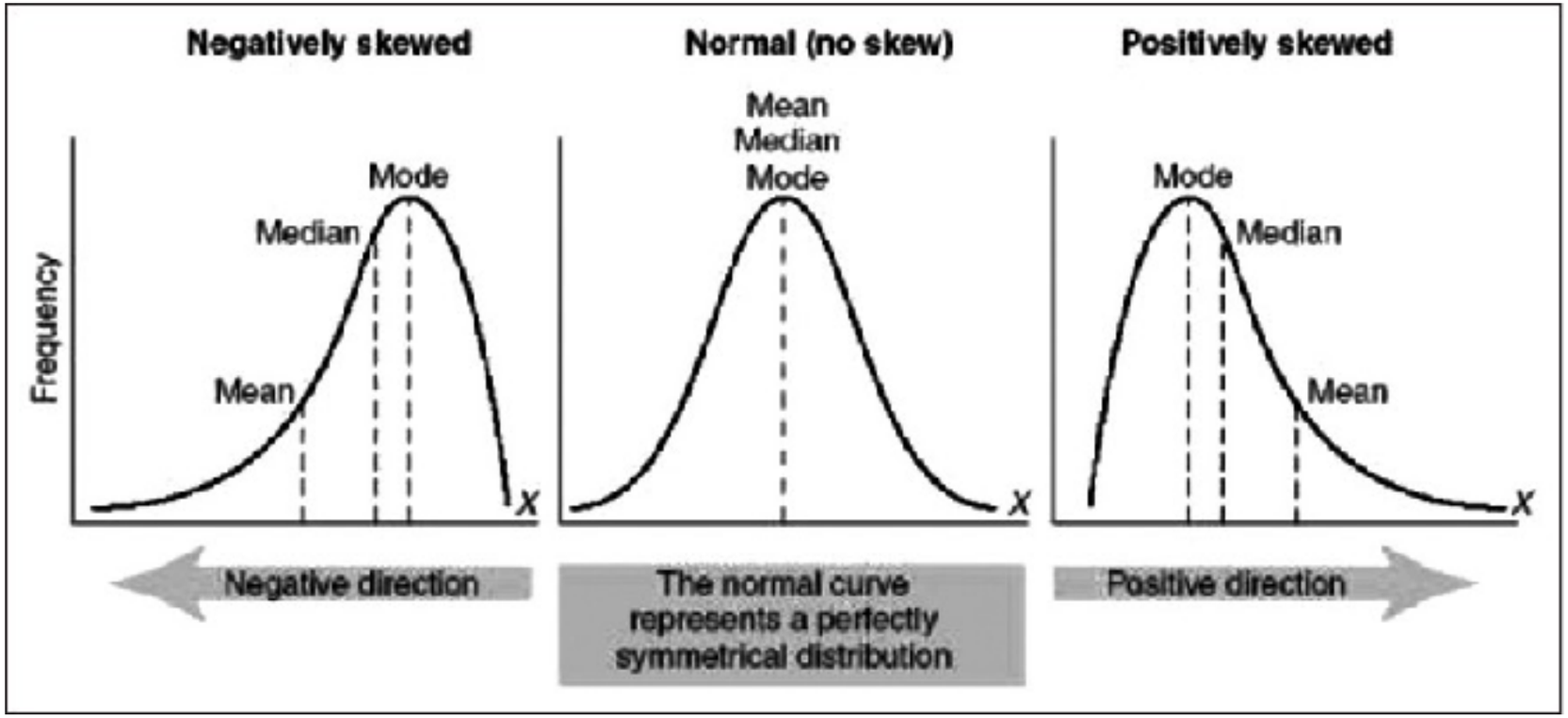
THE 68 – 95 – 99.7% RULE



SKEWNESS

- Skewness is a measure of the asymmetry of the distribution of a random variable about its mean.

- Skewness can be positive or negative, or even undefined.



Look for the 'long tail'

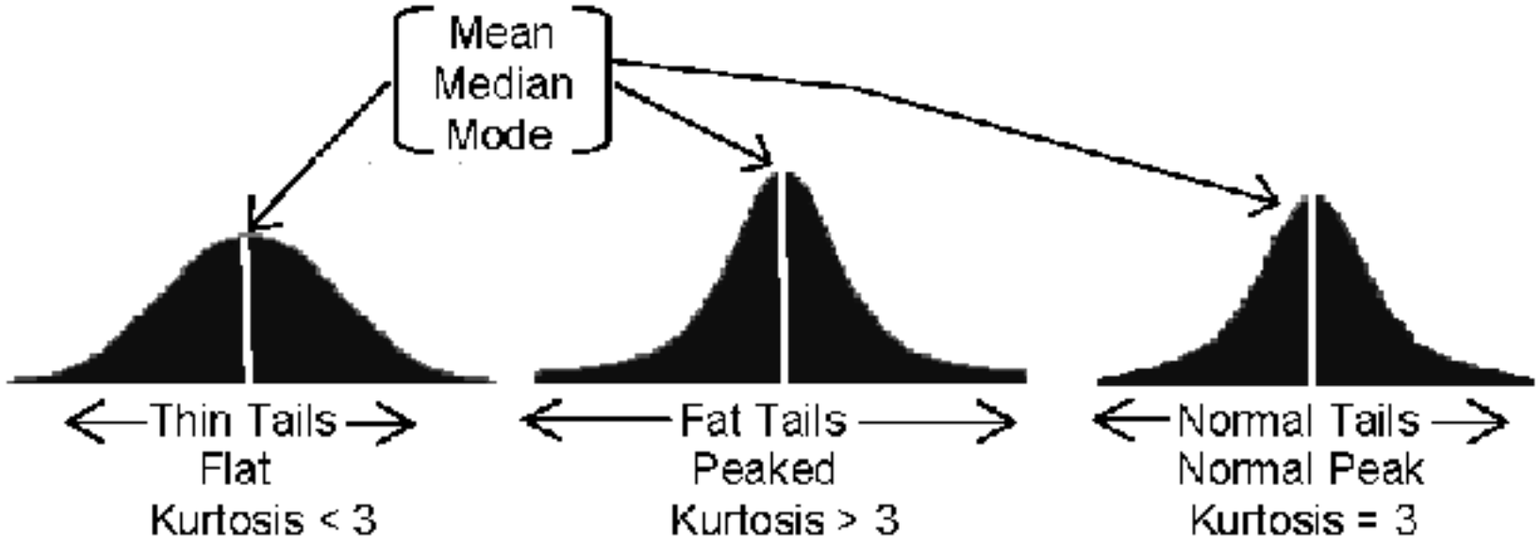
KURTOSIS

- Kurtosis is a measure of whether the data have ‘fat tails’ relative to a normal distribution, i.e., have more outliers.

- It is calculated as the sum of the deviations of the values from the mean raised to the 4^{th} power, divided by the squared variance.

Benchmark
Normal distribution:
Kurtosis = 3

Also occurs as ‘excess kurtosis’ => kurtosis - 3



DEMO

DETERMINING THE DISTRIBUTION OF YOUR DATA

DETERMINING THE DISTRIBUTION OF YOUR DATA

- Follow along as we walk through this in an iPython Notebook.

GUIDED PRACTICE

IS THIS SKEWED?

ACTIVITY: IS THIS SKEWED?

DIRECTIONS (10 minutes)

1. We're going to walk through several images of datasets.
2. For each image, vote on whether the image is:
   a. Normal
   b. Positively, negatively, or not skewed
   c. Has positive, negative, or zero kurtosis
3. Determine how you would correct the issue with each dataset to return it to the normal distribution.

VARIABLE TYPES

VARIABLE TYPES & STATISTICS

- **Numeric variables** can take on a large range of non-predetermined, quantitative values. These are things such as height, income, etc.
  - Absolute vs relative: fixed zero value?
  - Continuous vs integer values

- **Categorical variables** can take on a specific set of variables. These are things such as race, gender, paint colors, movie titles, product ID.
  - Nominal vs ordinal: ordered, but not a measurable intervals

VARIABLE TYPES

Type of Variable	Best measure of central tendency
Nominal	Mode
Ordinal	Median
Interval/Ratio (not skewed)	Mean
Interval/Ratio (skewed)	Median

DEMO

CLASSES

CLASS/DUMMY VARIABLES

- Let's say we have the categorical variable `area`, which takes on one of the following values: `rural`, `suburban`, and `urban`.

- We need to represent these numerically for a model. So how do we code them?

CLASS/DUMMY VARIABLES

- How about `0=rural`, `1=suburban`, and `2=urban`?

CLASS/DUMMY VARIABLES

- But this implies an ordered relationship - is `urban` twice `suburban`? That doesn't make sense.

- However, we can represent this information by converting the one `area` variable into two new variables:

  `area_urban` and `area_suburban`.

CLASS/DUMMY VARIABLES

- We'll draw out how categorical variables can be represented without implying order.
- First, let's choose a reference category. This will be our "base" category.
- It's often good to choose the category with the largest sample size and a criteria that will help model interpretation. If we are testing for a disease, the reference category would be people without the disease.

CLASS/DUMMY VARIABLES

- Step 1: Select a reference category. We'll choose `rural` as our reference category.
- Step 2: Convert the values `urban`, `suburban`, and `urban` into a numeric representation that does not imply order.
- Step 3: Create two new variables: `area_urban` and `area_suburban`.

CLASS/DUMMY VARIABLES

- Why do we need only two dummy variables?

`rural`	`urban`	`suburban`

- We can derive all of the possible values from these two. If an `area` isn't `urban` or `suburban`, we know it must be `rural`.

- In general, if you have a categorical feature with k categories, you need to create k-1 dummy variable to represent all of the information.

CLASS/DUMMY VARIABLES

- Let's see our dummy variables.

	area_urban	area_suburban
rural	0	0
suburban	0	1
urban	1	0

- As mentioned before, if we know `area_urban=0` and `area_suburban=0`, then the area must be `rural`.

CLASS/DUMMY VARIABLES

- We can do this for a `gender` variable with two categories: `male` and female.
- How many dummy variables need to be created?

CLASS/DUMMY VARIABLES

- # of categories - 1 = 2 -1 = 1

CLASS/DUMMY VARIABLES

- We will make `female` our reference category. Thus, `female=0` and `male=1`.

	gender_male
female	0
male	1

- This can be done in Pandas with the `get_dummies` method.

INDEPENDENT PRACTICE

DUMMY COLORS

ACTIVITY: DUMMY COLORS

DIRECTIONS (15 minutes)

It's important to understand the concept before we use the Pandas function `get_dummies` to create dummy variables. So today, we'll create our dummy variables by hand.

1. Draw a table like the one on the white board.
2. Create dummy variables for the variable "colors" that has 6 categories: blue, red, green, purple, grey, and brown. Use grey as the reference.

DELIVERABLE

Dummy variables table for colors

TOPIC REVIEW

REVIEW

- Let's go through the process for creating dummy variables for "colors".
  - We talked about several different types of summary statistics, what are they?
  - We covered several different types of visualizations; which ones?
  - We talked about the normal distribution; how do we determine your data's distribution?
- Any other questions?

BEFORE NEXT CLASS

BEFORE NEXT CLASS

DUE DATE: 09/29

- Start Project: Unit Project 2

LESSON

Q & A

LESSON

EXIT TICKET

DON’T FORGET TO FILL OUT YOUR EXIT TICKET