

The background features abstract, overlapping green geometric shapes, primarily triangles and polygons, in various shades of green, creating a modern and dynamic visual effect.

# Data Science for Engineers

# Course Objectives

- ▶ Develop relevant programming abilities.
- ▶ Proficiency with statistical analysis of data.
- ▶ Build and assess data based models.
- ▶ Execute statistical analysis.

- ▶ Text book:
- ▶ Joel Grus, Data Science from Scratch- First Principles with Python, O'Reilly Publications, 2<sup>nd</sup> Edition, 2019, ISBN: 978-9352138326

# What is data Science?

## ► Big Data is Everywhere:

- Websites
- Smart phones, social network
- Quantified selfers ( store heart rates, diet, sleep patterns,)
- Smart Devices
- Smart market, e-commerce
- Point of sales at stores.
- Internet( cross references encyclopaedia, movies, sport results, games, memes, government statistics).

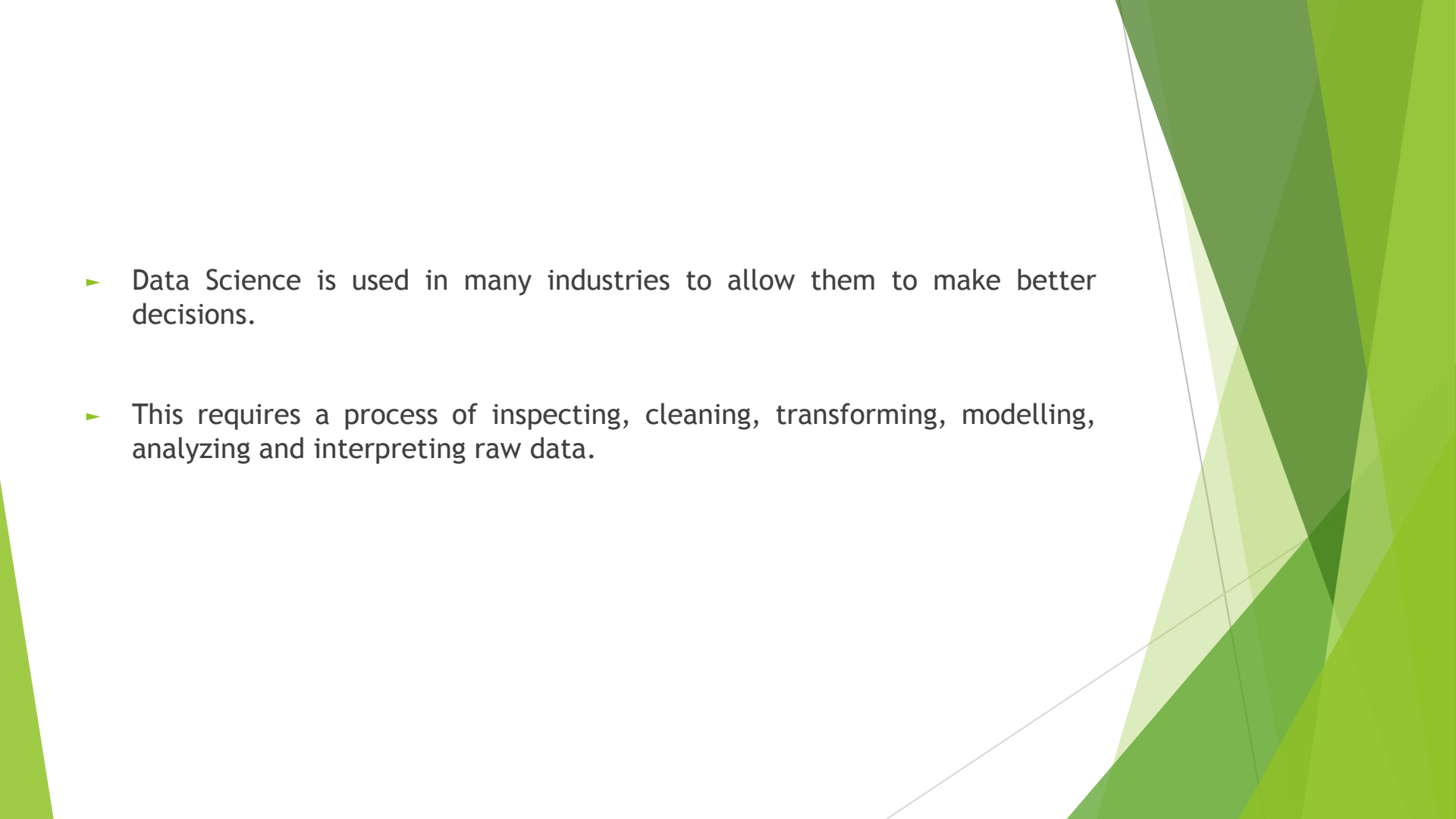
< Buried in this data are answers to countless questions that no one ever thought to ask >

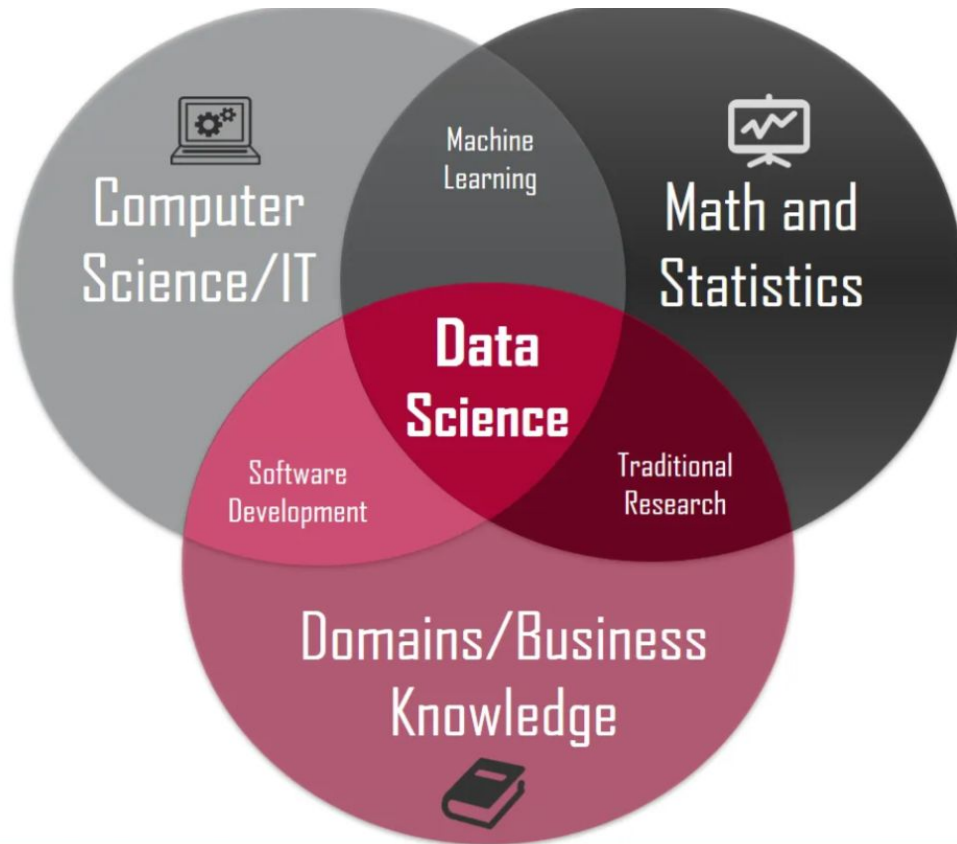
- ▶ Google handle over 2.5 exabytes (2,500,000,000 gigabytes) of data every single day.
- ▶ Facebook uses 500+ terabytes of data everyday.
- ▶ Amazon owns 1.4 million servers. They generate 2.5 exabytes bytes of data every day.

Data Storage	Units
Bit	1 or 0
Byte	8 bits
Kilobyte	1,000 bytes
Megabyte	1,000 kilobytes
Gigabyte	1,000 megabytes
Terabyte	1,000 gigabytes
Petabyte	1,000 terabytes
Exabyte	1,000 petabytes
Zettabyte	1,000 exabytes
Yottabyte	1,000 zettabytes

# What is Data Science?

- ▶ Data Science is the science of analyzing the raw data using statistics and machine learning techniques with the purpose of drawing conclusions about that information.
  - ▶ Machine learning is a data analytics technique that teaches computers to do what comes naturally to humans and animals: learn from experience. Machine learning algorithms use computational methods to directly "**learn**" from data without relying on a predetermined equation as a model.

- 
- The background of the slide features an abstract design composed of various shades of green. These shades are arranged in overlapping, angular, and somewhat translucent shapes that create a sense of depth and movement. The colors range from a light, pale green to a deep, forest green. The overall effect is modern and tech-oriented, typical of a corporate or academic presentation.
- ▶ Data Science is used in many industries to allow them to make better decisions.
  - ▶ This requires a process of inspecting, cleaning, transforming, modelling, analyzing and interpreting raw data.





# Fundamental Python Libraries for data scientists

- ▶ NumPy provides tools for
  - ▶ Basic operations
  - ▶ Multidimensional arrays
  - ▶ Linear Algebraic functions
- ▶ SciPy
  - ▶ signal processing,
  - ▶ optimization,
  - ▶ statistics,

# Visualizing Data

- ▶ **the practice of translating information into a visual context, such as a map or graph, to make data easier for the human brain to understand and pull insights from.**
- ▶ **Two main uses of visualization:**
  - ▶ **To explore data**
  - ▶ **To communicate data**

# matplotlib

- ▶ There are variety of tools for visualising the data. The one we use is **matplotlib**.
- ▶ Matplotlib is open source and we can use it freely.
- ▶ Its not a part of core python library. So we need to install it.
- ▶ Command:  
*python -m pip install matplotlib*
- ▶ And then you need to import the library.

# plot() function

- ▶ The `plot()` function is used to draw points (markers) in a diagram.
- ▶ By default, the `plot()` function draws a line from point to point.
- ▶ The function takes parameters for specifying points in the diagram.
- ▶ Parameter 1 is an array containing the points on the **x-axis**.
- ▶ Parameter 2 is an array containing the points on the **y-axis**.
- ▶ If we need to plot a line from (1, 3) to (8, 10), we have to pass two arrays [1, 8] and [3, 10] to the plot function

# Plotting Without Line

- To plot only the markers, you can use *shortcut string notation* parameter 'o', which means 'rings'.
- `xpoints = array([1, 8])`  
`ypoints = array([3, 10])`

```
plt.plot(xpoints, ypoints, 'o')
```

## Default X-Points

If we do not specify the points on the x-axis, they will get the default values 0, 1, 2, 3 etc., depending on the length of the y-points.

## Marker:

```
plt.plot(ypoints, marker = 'o')
```

## Format Strings fmt

You can also use the shortcut string notation parameter to specify the marker.

This parameter is also called `fmt`, and is written with this syntax:

*marker|line|color*

```
plt.plot(ypoints, 'o:r')
```

Marker size:

`ms`

Marker edge color

`mec`

# linestyle

```
plt.plot(ypoints, linestyle = 'dashed')
```

- ▶ linestyle can be written as `ls`
- ▶ dotted can be written as `:`
- ▶ dashed can be written as `-`

```
plt.plot(ypoints, ls = ':red')
```

- ▶ Line Width
- ▶ Line color

# Label , title

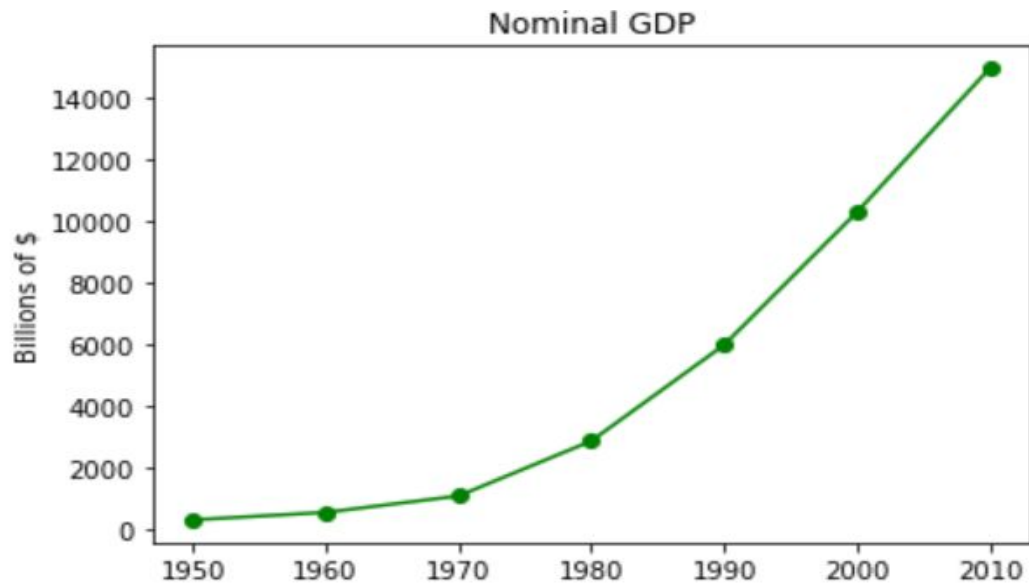
- ▶ you can use the `xlabel()` and `ylabel()` functions to set a label for the x- and y-axis.
- ▶ you can use the `title()` function to set a title for the plot.
  - ▶ We can use the `loc` parameter in `title()` to position the title.
  - ▶ Legal values are: 'left', 'right', and 'center'. Default value is 'center'.



# A simple line chart

```
import matplotlib.pyplot as pyplot  
years = [1950,1960,1970,1980,1990,2000,2010]  
gdp = [300.2, 543.3, 1075.9, 2862.5, 5979.6,  
10289.7, 14958.3]  
  
pyplot.plot(years,gdp, color='green', marker  
='o', linestyle='solid')  
  
pyplot.title("Nominal GDP")  
pyplot.ylabel("Billions of $")  
pyplot.show()
```

# Output



# Bar chart

- ▶ A bar graph is a graphical representation of data in which we can highlight the category with particular shapes like a rectangle.
- ▶ The length and heights of the bar chart represent the data distributed in the dataset.
- ▶ In a bar chart, we have one axis representing a particular category of a column in the dataset and another axis representing the values or counts associated with it.
- ▶ Bar charts can be plotted vertically or horizontally.
- ▶ A vertical bar chart is often called a column chart.

# Bar Charts

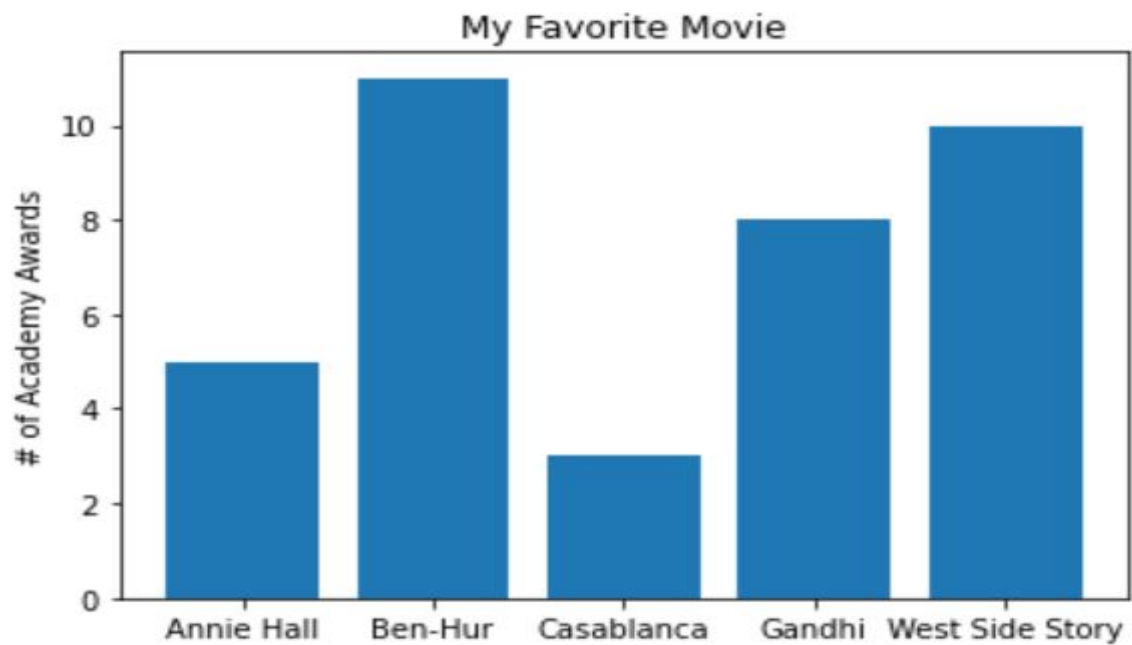
- ▶ Bar chart is a good choice when you want to show how some quantity varies among some discrete set of items.

**bar()**

- ▶ Horizontal bars

**barh()**

- ▶ **Color, width. height**



- Dictionaries are written with curly brackets, and have keys and values:

```
thisdict = {  
    "brand": "Ford",  
    "model": "Mustang",  
    "year": 1964  
}  
print(thisdict)
```

Dictionary items are ordered, changeable, and does not allow duplicates.

- Dictionary items are presented in key:value pairs, and can be referred to by using the key name.

```
print(thisdict["brand"])
```

- Change the "year" to 2018:

```
thisdict = {  
    "brand": "Ford",  
    "model": "Mustang",  
    "year": 1964  
}  
thisdict["year"] = 2018
```

## UPDATE:

Update the "year" of the car by using the `update()` method:

```
thisdict = {  
    "brand": "Ford",  
    "model": "Mustang",  
    "year": 1964  
}  
thisdict.update({"year": 2020})
```

# Finding Key connectors

```
users = [  
  { "id": 0, "name": "Hero" },  
  { "id": 1, "name": "Dunn" },  
  { "id": 2, "name": "Sue" },  
  { "id": 3, "name": "Chi" },  
  { "id": 4, "name": "Thor" },  
  { "id": 5, "name": "Clive" },  
  { "id": 6, "name": "Hicks" },  
  { "id": 7, "name": "Devin" },  
  { "id": 8, "name": "Kate" },  
  { "id": 9, "name": "Klein" }  
]
```



```
friendships = [(0, 1), (0, 2), (1, 2), (1, 3), (2, 3), (3, 4),  
               (4, 5), (5, 6), (5, 7), (6, 8), (7, 8), (8, 9)]
```

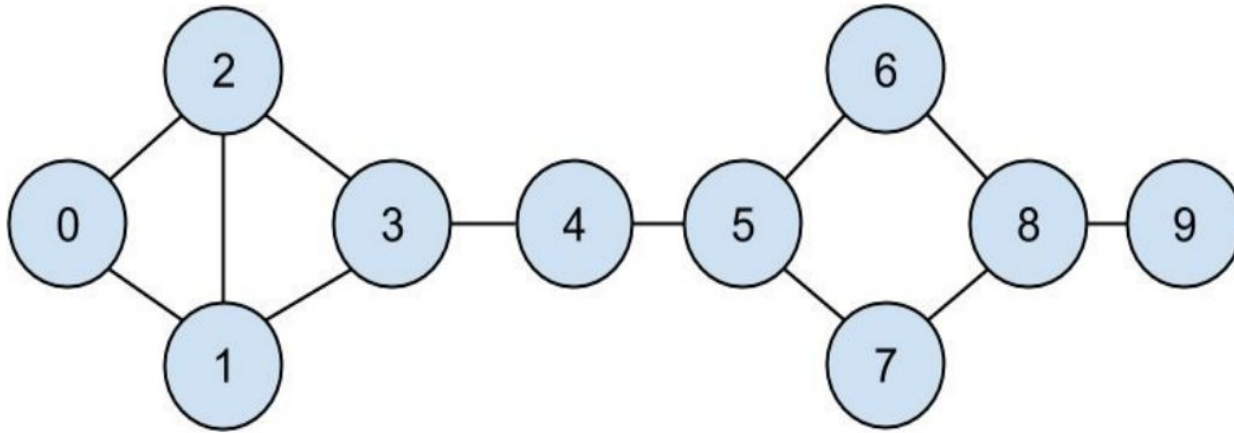


Figure 1-1. The DataSciencecenter network

For example, we might want to add a list of friends to each user. First we set each user's friends property to an empty list:

```
for user in users:  
    user["friends"] = []
```

And then we populate the lists using the friendships data:

```
for i, j in friendships:  
    # this works because users[i] is the user whose id is i  
    users[i]["friends"].append(users[j]) # add i as a friend of j  
    users[j]["friends"].append(users[i]) # add j as a friend of i
```

First we find the *total* number of connections, by summing up the lengths of all the friends lists:

```
def number_of_friends(user):  
    """how many friends does _user_ have?"""  
    return len(user["friends"])           # length of friend_ids list  
  
total_connections = sum(number_of_friends(user)  
                        for user in users) # 24
```

And then we just divide by the number of users:

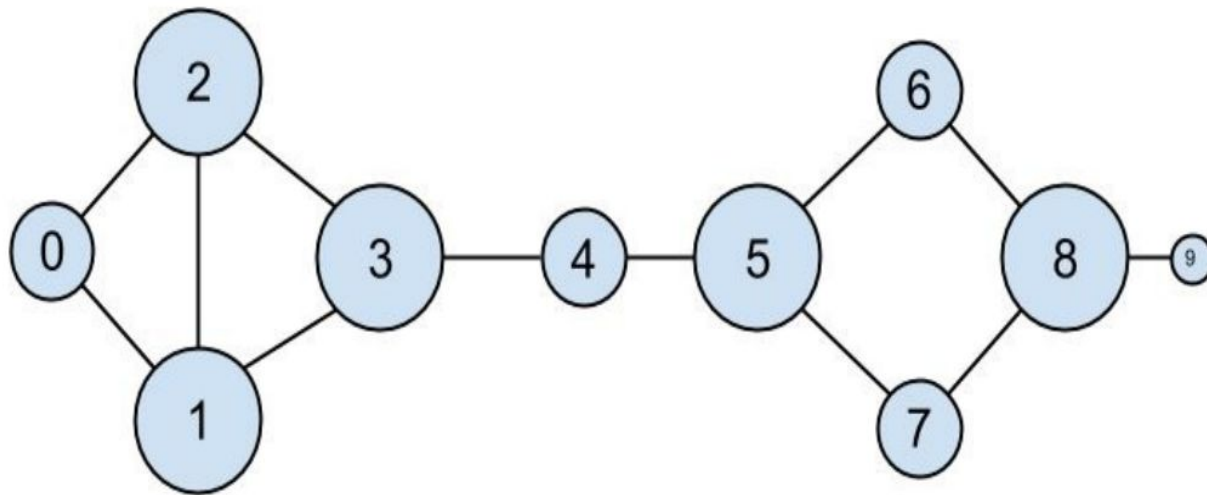
```
from __future__ import division           # integer division is lame  
num_users = len(users)                   # length of the users list  
avg_connections = total_connections / num_users # 2.4
```

It's also easy to find the most connected people — they're the people who have the largest number of friends.

```
# create a list (user_id, number_of_friends)
num_friends_by_id = [(user["id"], number_of_friends(user))
                      for user in users]

sorted(num_friends_by_id,                                # get it sorted
       key=lambda (user_id, num_friends): num_friends, # by num_friends
       reverse=True)                                   # largest to smallest

# each pair is (user_id, num_friends)
# [(1, 3), (2, 3), (3, 3), (5, 3), (8, 3),
#  (0, 2), (4, 2), (6, 2), (7, 2), (9, 1)]
```



*Figure 1-2. The DataSciencecenter network sized by degree*

# Statistics

## Fundamentals of Statistics

- A visual and mathematical portrayal of information is statistics. Data science is all about making calculations with data.
- We make decisions based on that data using mathematical conditions known as models.
- Numerous fields, including data science, machine learning, business intelligence, computer science, and many others have become increasingly dependent on statistics.

- ▶ Researchers heavily rely on statistics to gather relevant data and make informed conclusions.
- ▶ Without proper statistical expertise, valuable resources such as time, money, and data can be wasted.

## Examples of statistics that are used in day-to-day life:

- ▶ Statistics play a vital role in the medical industry as they help determine the effectiveness of drugs before prescribing them. Medical studies rely on statistical analysis to provide accurate and reliable results.
- ▶ In our daily lives, we often make predictions using statistics. For instance, setting an alarm to wake up in the morning is a way of predicting the future based on past patterns.
- ▶ Netflix, for example, uses the number of movies browsed in different genres to recommend new movies based on individual preferences.
- ▶ Similarly, in cricket, the fielding positions are set based on a statistical analysis of a batsman's playing patterns and strengths.



# Central Tendencies

- ▶ Notion of where our data is centered
- ▶ **Mean:** Average value or sum of the value divided by count.
- ▶ If we have two data points, mean is half way between them.
- ▶ **Median:** Middlemost value or average of two middle most values. The median is the middle value in a set of data.
- ▶ Median does not fully depend on every value of the data set.

## Activity:

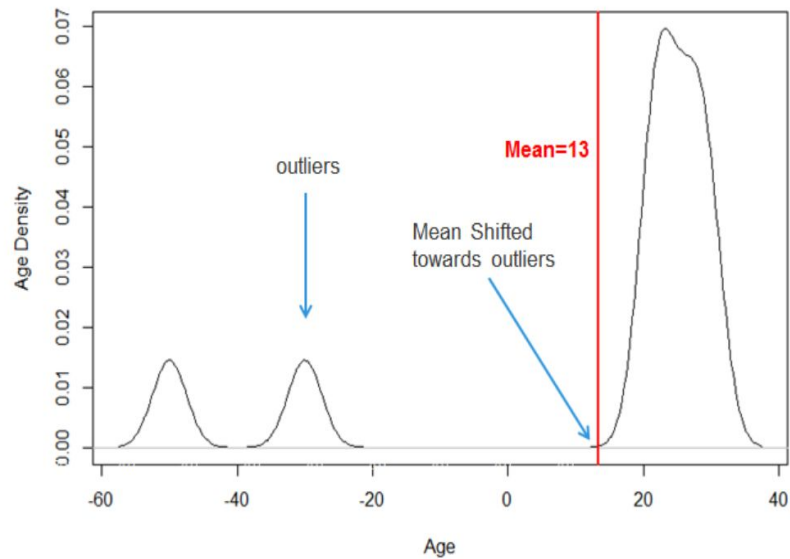
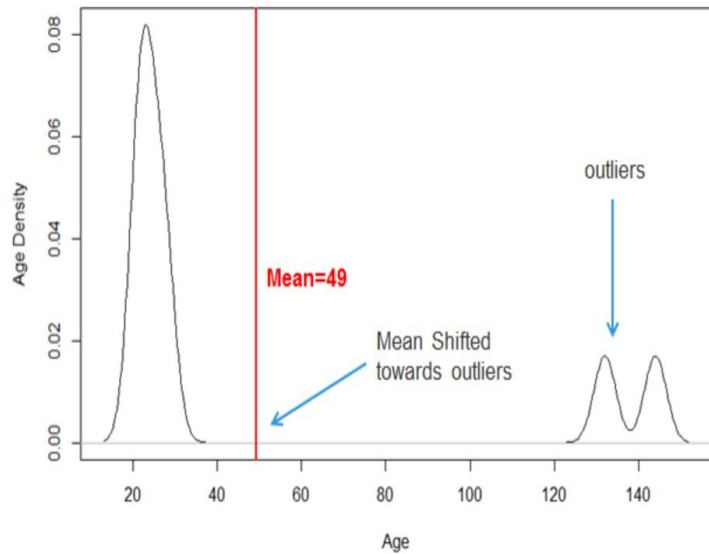
- ▶ Write python program to find mean and median for given set of data.

# Inference

- ▶ Mean is simpler to compute, varies smoothly as our data changes
- ▶ If there are  $n$  data points and one of them increases by small amount  $e$ , then necessarily mean increases by  $e/n$ .
- ▶ Sensitive to outliers of data(bad data)
  - ▶ misleading conclusions.
- ▶ In order to compute median, we need to sort data.
- ▶ changing one of our data points by a small amount  $e$  might increase the median by  $e$ , by some number less than  $e$ , or not at all.

Mean	Median
Simpler to compute	Slightly complex as compared to mean
Varies smoothly as our data changes	Need to sort the data, changing one of the data point might slightly increase median or may not change at all.
Sensitive to outliers of the data(	Median will not be affected much

- ▶ Mean can be misleading:
- ▶ Example:
  - ▶ Planning Develop a shopping app for a particular region
  - ▶ Data of age:
    - ▶ 23,32,34,26,22,27,28,92,90,84
    - ▶ Mean = 45.8
- ▶ 1, 2, 3 mean is 2
- ▶ 1, 2 , 300 mean is 101
- 'Average placement salary of students from our institute is \$120,000'

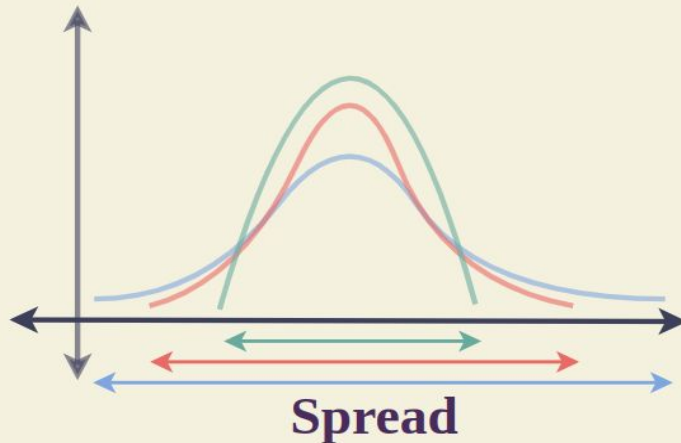


- ▶ Mean can be used where every data point needs to be taken into the calculation
- ▶ The **mode** is the number that occurs most often in a data set

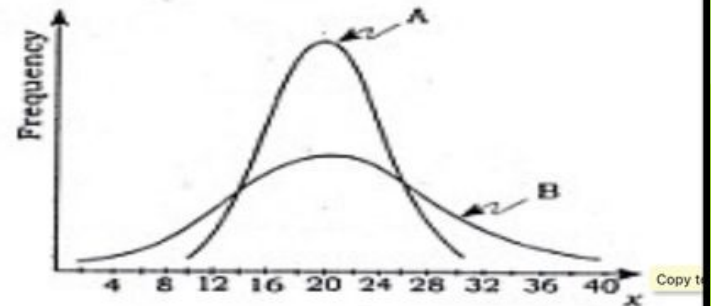
# Dispersion

- ▶ Dispersion is the state of getting dispersed or spread.
- ▶ Statistical dispersion means the extent to which numerical data is likely to vary about an average value.
- ▶ In other words, dispersion helps to understand the distribution of the data.

## Measure of Dispersion



Data A: 10 11 14 20 20 20 22 24 28 31  
Data B: 2 9 13 14 20 20 24 26 32 40



- ▶ **Measure of Dispersion** is the numbers that are used to represent the scattering of the data. There are various measures of dispersion that are used to represent the data that includes,
- ▶ Range
  - Standard Deviation
  - Variance



# Measures of dispersion

- **Range:** It is simply the difference between the maximum value and the minimum value given in a data set.

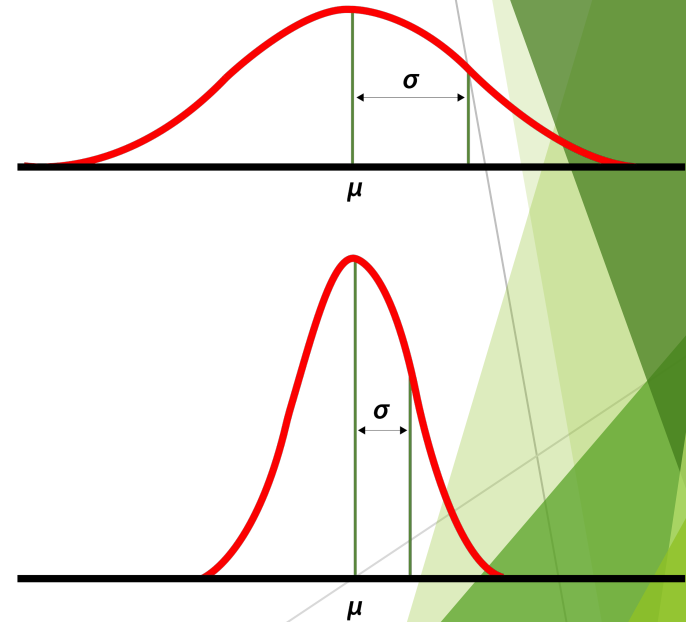
Example: 1, 3, 5, 6, 7  $\Rightarrow$  Range =  $7 - 1 = 6$

Range is 0 when max and min are equal, which can happen only when all the elements in a list are same. Which means data is **undispersed** as possible.

Conversely if range is large, then max is much bigger than min. That means data is more spread out.

# Standard Deviation

- ▶ A standard deviation (or  $\sigma$ ) is a measure of how dispersed the data is in relation to the mean.
- ▶ Low, or small, standard deviation indicates data are clustered tightly around the mean,
- ▶ and high, or large, standard deviation indicates data are more spread out.

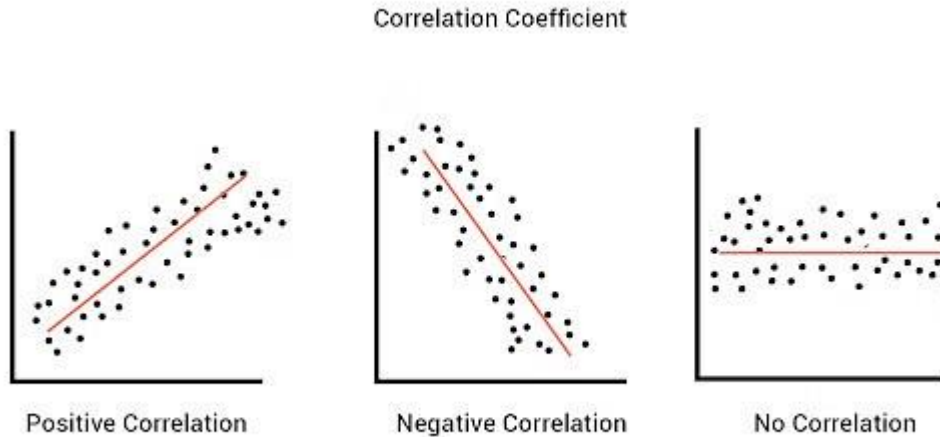


## ► Variance

- variance measures how far each number in the set is from the mean (average), and thus from every other number in the set. Variance is often depicted by this symbol:  $\sigma^2$
- investors use variance to see how much risk an investment carries and whether it will be profitable.

- ▶ **Correlation:** Correlation measures the relationship between two variables.
- ▶ The correlation coefficient can never be less than -1 or higher than 1.
- ▶ It measures the extent to which two variables are linearly related.
- ▶ For example, the height and weight of a person are related, and taller people tend to be heavier than shorter people.

- ▶ You can apply correlation to a variety of data sets.
- ▶ In some cases, you may be able to predict how things will relate, while in others, the relation will come as a complete surprise.
- ▶ It's important to remember that just because something is correlated doesn't mean it's causal.



- ▶ Example 1: Body Fat and Running Time
- ▶ An individual's body fat tends to be lower the more time they spend jogging. In other words, there is a **negative correlation** between the variable body fat and the variable running time. Body fat decreases as running time increases.
- ▶ Example 2: Temperature Vs. Sales of Ice Cream **positive correlation**
- ▶ Example 3: An example of this would be the amount of chocolate someone eats and how many hours they spend on homework. **No correlation**

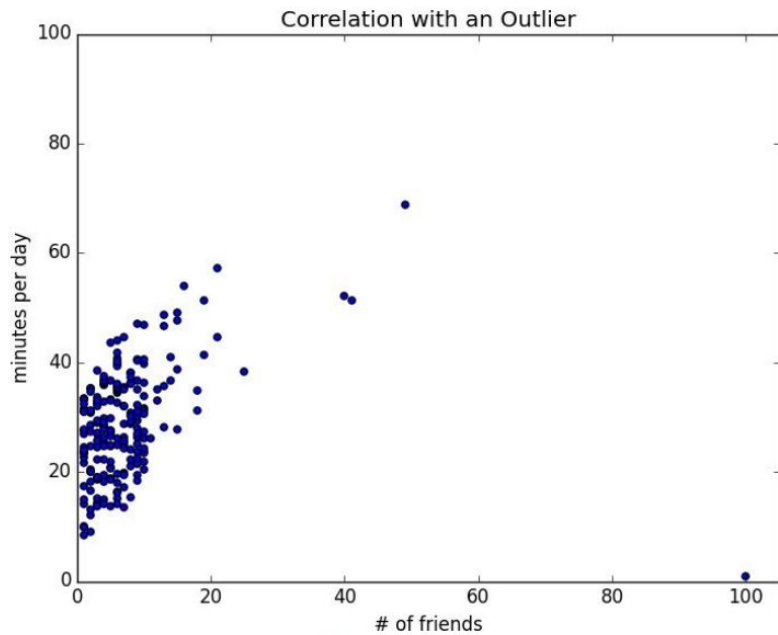


Figure 5-2. Correlation with an outlier

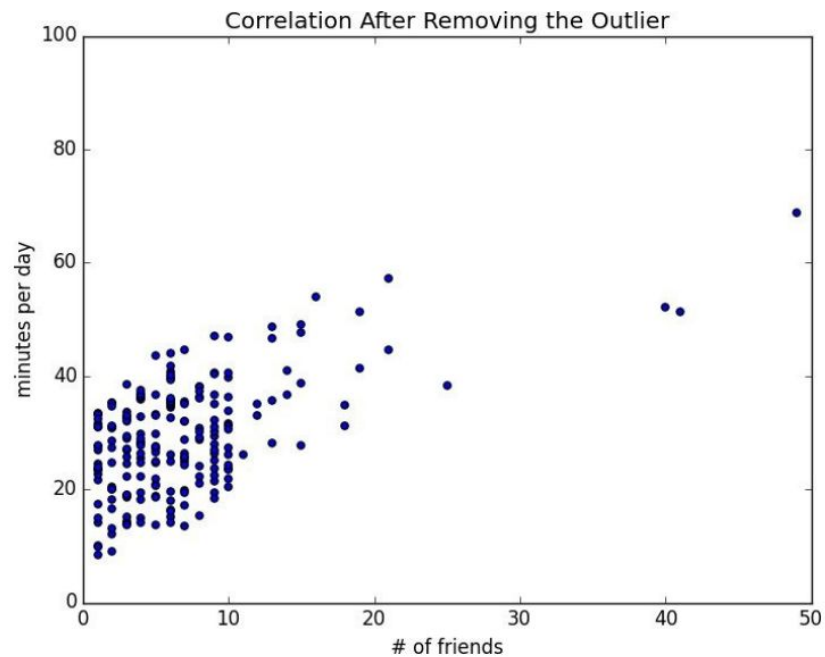


Figure 5-3. Correlation after removing the outlier

# Simpsons Paradox

- ▶ The art of data science is seeing beyond the data — using and developing methods and tools to get an idea of what that hidden reality looks like.
- ▶ The paradox is relatively simple to state, and is often a cause of confusion and misinformation for non-statistically trained audiences:
- ▶ Simpson's Paradox:
- ▶ *A trend or result that is present when data is put into groups that reverses or disappears when the data is separated.*



Flavour	Sample Size	# Liked Flavour
Sinful Strawberry	1000	800
Passionate Peach	1000	750

We can see that 80% of people enjoyed ‘Sinful Strawberry’ whereas only 75% of people enjoyed ‘Passionate Peach’. So ‘Sinful Strawberry’ is more likely to be the preferred flavour.

Flavour	# Men	# Liked Flavour (Men)	# Women	# Liked Flavour (Women)
Sinful Strawberry	900	760	100	40
Passionate Peach	700	600	300	150

This suggests that 84.4% of men and 40% of women liked ‘Sinful Strawberry’ whereas 85.7% of men and 50% of women liked ‘Passionate Peach’.

- ▶ One of the most famous examples of Simpson's paradox is UC Berkley's suspected gender-bias. At the beginning of the academic year in 1973, UC Berkeley's graduate school had admitted roughly 44% of their male applicants and 35% of their female applicants.

	Men		Women	
	# Applied	% Admitted	# Applied	% Admitted
Major A	825	62% admitted	108	82% admitted
Major B	560	63% admitted	25	68% admitted
Major C	325	37% admitted	593	34% admitted
Major D	417	33% admitted	375	35% admitted
Major E	191	28% admitted	393	24% admitted
Major F	373	6% admitted	341	7% admitted
Total:	2,690	44% admitted	1,835	34.5% admitted

# Simpsons Paradox

coast	# of members	avg. # of friends
West Coast	101	8.2
East Coast	103	6.5

coast	degree	# of members	avg. # of friends
West Coast	PhD	35	3.1
East Coast	PhD	70	3.2
West Coast	no PhD	66	10.9
East Coast	no PhD	33	13.4

- ▶ Simpson's paradox can make decision-making hard.
- ▶ We can scrutinize and regroup and resample our data as much as we are able to, but if multiple different conclusions can be drawn from all the different categorizations, then choosing a grouping to draw our conclusions from in order to gain insight and develop strategies is a confusion and difficult problem.
- ▶ We need to know what we are looking for, and to choose the best data-viewpoint giving a fair representation of the truth.

# Probability

- ▶ Our goal is to build a mathematical framework to represent and analyze uncertain phenomena, such as the result of rolling a die, tomorrow's weather, the result of an NBA game, etc.
- ▶ we model the phenomenon of interest as an experiment with several (possibly infinite) mutually exclusive outcomes.
- ▶ Notationally we write  $P(E)$  to mean “the probability of event  $E$ ”.

# Dependence and Independence

- ▶ Two events  $E$  and  $F$  are dependent if knowing something about whether  $E$  happens gives us information about whether  $F$  happens (and vice versa). Otherwise they are independent.
- ▶ If we flip a fair coin twice, knowing whether the first flip is Heads gives us no information about whether the second flip is Heads. These events are independent.
- ▶ On the other hand, knowing whether the first flip is Heads certainly gives us information about whether both flips are Tails. (If the first flip is Heads, then definitely it's not the case that both flips are Tails.) These two events are dependent.

- ▶ Mathematically, we say that two events E and F are independent if the probability that they both happen is the product of the probabilities that each one happens:

$$P(E,F) = P(E).P(F)$$

The probability of “first flip Heads” is  $1/2$ , and the probability of “both flips Tails” is  $1/4$ , but the probability of “first flip Heads and both flips Tails” is 0



## Conditional Probability

When two events  $E$  and  $F$  are independent, then by definition we have:

$$P(E, F) = P(E)P(F)$$

If they are not necessarily independent (and if the probability of  $F$  is not zero), then we define the probability of  $E$  “conditional on  $F$ ” as:

$$P(E \mid F) = P(E, F) / P(F)$$

You should think of this as the probability that  $E$  happens, given that we know that  $F$  happens.

We often rewrite this as:

$$P(E, F) = P(E \mid F)P(F)$$

When  $E$  and  $F$  are independent, you can check that this gives:

$$P(E \mid F) = P(E)$$

which is the mathematical way of expressing that knowing  $F$  occurred gives us no additional information about whether  $E$  occurred.

## Example:

- ▶ One common tricky example involves a family with two (unknown) children.

If we assume that:

1. Each child is equally likely to be a boy or a girl
  2. The gender of the second child is independent of the gender of the first child
- ▶ then the event “no girls” has probability  $1/4$ , the event “one girl, one boy” has probability  $1/2$ , and the event “two girls” has probability  $1/4$ .
  - ▶ Now we can ask what is the probability of the event “both children are girls” (B) conditional on the event “the older child is a girl” (G)?

- ▶ What is the probability of the event both the children are girls(B) conditional on the event atleast one of the child is girl.

# Bayeses theorm

- ▶ Bayes' Theorem states that the conditional probability of an event, based on the occurrence of another event, is equal to the likelihood of the second event given the first event multiplied by the probability of the first event
- ▶ It is a way of reversing conditional probabilities.
- ▶ Ex: we need to know probability of some event E conditional on event F.
- ▶ But we have information about event F conditional on event E.
- ▶ Then we can use bayes theorem.

# Bayes's Theorem

Bayes' theorem may be derived from the definition of [conditional probability](#):

$$P(A|B) = \frac{P(A \cap B)}{P(B)}, \text{ if } P(B) \neq 0,$$

where  $P(A \cap B)$  is the probability of both A and B being true. Similarly,

$$P(B|A) = \frac{P(A \cap B)}{P(A)}, \text{ if } P(A) \neq 0.$$

Solving for  $P(A \cap B)$  and substituting into the above expression for  $P(A|B)$  yields Bayes' theorem:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}, \text{ if } P(B) \neq 0.$$

In a certain city, 20% of the population has a rare genetic condition. There is a diagnostic test available for this condition that correctly identifies 95% of those who have the condition and incorrectly identifies 10% of those who do not have it as positive. If a randomly selected individual takes the test and the result is positive, what is the probability that the individual actually has the rare genetic condition?

To solve this problem using Bayes' Theorem, we define the events:

- A: Individual has the rare genetic condition.
- B: Test result is positive.

Given:

- $P(A) = 20\%$
- $P(\sim A) = 80\%$
- $P(B|A) = 95\%$

Given:

- $P(A) = 20\%$
- $P(\sim A) = 80\%$
- $P(B|A) = 95\%$
- $P(B|\sim A) = 10\%$

We can now calculate the probability that the individual actually has the rare genetic condition given a positive test result by applying Bayes' Theorem:

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B|A) \cdot P(A) + P(B|\sim A) \cdot P(\sim A)}$$

By substituting the given probabilities into the formula, we can determine the probability that an individual has the rare genetic condition given a positive test result.



- ▶ Example: planning for a picnic. And we know that  $p(\text{rain})$  is 0.15. and  $p(\text{cloudy}) = 0.25$ . also  $p(c|r) = 0.80$ . what's  $p(r|c)$ ?

Imagine you are a financial analyst at an investment bank. According to your research of publicly-traded companies, 60% of the companies that increased their share price by more than 5% in the last three years replaced their CEOs during the period. At the same time, only 35% of the companies that did not increase their share price by more than 5% in the same period replaced their CEOs. Knowing that the probability that the stock prices grow by more than 5% is 4%, find the probability that the shares of a company that fires its CEO will increase by more than 5%.

To solve this problem using Bayes' Theorem, we first define the probabilities:

- Let A be the event that a company replaces its CEO.
- Let B be the event that a company's stock price grows by more than 5%.

Given:

- $P(A|B) = 60\%$
- $P(A|B') = 35\%$
- $P(B) = 4\%$

We can now calculate the probability that the shares of a company that replaces its CEO will grow by more than 5% using Bayes' Theorem:

$$P(B|A) = \frac{P(A|B) \cdot P(B)}{P(A|B) \cdot P(B) + P(A|B') \cdot P(B')}$$

By substituting the given values into the formula, we can find the probability that the shares of a company that replaces its CEO will increase by more than 5%.

# Random variable

- ▶ A random variable is a variable whose possible values have an associated probability distribution.
- ▶ A random variable is a numerical function that assigns a real value to each possible outcome of a random experiment or process
- ▶ A very simple random variable equals 1 if a coin flip turns up heads and 0 if the flip turns up tails.
- ▶ A more complicated one might measure is a value picked from  $\text{range}(10)$  where each number is equally likely.
- ▶ The associated distribution gives the probabilities that the variable realizes each of its possible values
- ▶ The coin flip variable equals 0 with probability 0.5 and 1 with probability 0.5. The  $\text{range}(10)$  variable has a distribution that assigns probability 0.1 to each of the numbers from 0 to 9

## contd...

- ▶ The expected value (mean) and variance are important statistical measures that describe the central tendency and spread of a random variable's probability distribution.
- ▶ Random variables are fundamental to probability and statistics, allowing the quantification and analysis of random phenomena, hypothesis testing, and data modeling
- ▶ Two types:
  - ▶ Discrete Random Variable
  - ▶ Continues Random Variable

1. The **expected value** of a random variable is denoted by  $E[X]$ . The expected value can be thought of as the “average” value attained by the random variable; in fact, the expected value of a random variable is also called its **mean**, in which case we use the notation  $\mu_X$ . ( $\mu$  is the Greek letter mu.)
2. The formula for the expected value of a discrete random variable is this:

$$E[X] = \sum_{\text{all possible } x} xP(X = x).$$

In words, the expected value is the sum, over all possible values  $x$ , of  $x$  times its probability  $P(X = x)$ .

3. Example: The expected value of the roll of a die is

$$1\left(\frac{1}{6}\right) + 2\left(\frac{1}{6}\right) + 3\left(\frac{1}{6}\right) + \cdots + 6\left(\frac{1}{6}\right) = 21/6 = 3.5.$$

Notice that the expected value is not one of the possible outcomes: you can't roll a 3.5. However, if you average the outcomes of a large number of rolls, the result approaches 3.5.

- ▶ Expected value for coin flip is:

- ▶  $0 \cdot \frac{1}{2} + 1 \cdot (\frac{1}{2})$

Compute Expected value for range(10)

# Discreate distributions

The rule that assigns specific probabilities to specific values for a discrete random variable is called its **probability mass function** or **pmf**. If  $X$  is a discrete random variable then we denote its pmf by  $P_X$ . For any value  $x$ ,  $P(X = x)$  is the probability of the event that  $X = x$ ; i.e.,

$$P(X = x) = \text{probability that the value of } X \text{ is } x.$$

Example: If  $X$  is the outcome of the roll of a die, then

$$P(X = 1) = P(X = 2) = \cdots = P(X = 6) = 1/6,$$

and  $P(X = x) = 0$  for all other values of  $x$ .

# Continuous Distribution

- ▶ A continuous random variable is one which takes an infinite number of possible values.
- ▶ Ex: the amount of sugar in an orange, the time required to run a mile.
- ▶ Continuous random variables are often used to represent measurements of physical quantities, such as temperature, length, time, and volume, which can have an infinite number of possible values within a given range.
- ▶ The probabilities of ranges of values for a continuous random variable is determined by a **density function**.

▶



# Uniform Distribution

- ▶ Uniform distribution : a type of probability distribution in which all outcomes are equally likely.
- ▶ The uniform distribution is a type of continuous random variable where all values between a minimum and maximum value have the same probability of occurring.
- ▶ The cumulative distribution function (CDF) for a continuous uniform distribution is a function that describes the probability that a random variable is less than or equal to a certain value.

The probabilities of ranges of values of a continuous random variable are determined by a **density** function. The density of  $X$  is denoted by  $f_X$ . The area under a density is always 1. The probability that  $X$  falls between two points  $a$  and  $b$  is the area under  $f_X$  between the points  $a$  and  $b$ . The familiar bell-shaped curve is an example of a density.

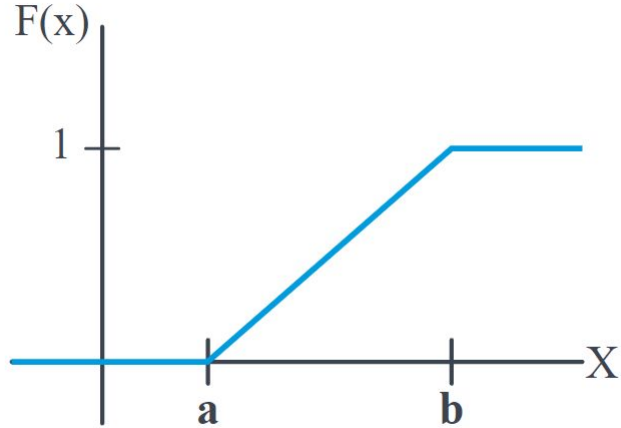
The **cumulative distribution function** or **cdf** gives the probability that a random variable  $X$  takes values less than or equal to a given value  $x$ . Specifically, the cdf of  $X$ , denoted by  $F_X$ , is given by

$$F_X(x) = P(X \leq x).$$

So,  $F_X(x)$  is the area under the density  $f_X$  to the left of  $x$ .

```
def uniform_cdf(x):  
    "returns the probability that a uniform random variable is <= x"  
    if x < 0: return 0      # uniform random is never less than 0  
    elif x < 1: return x    # e.g. P(X <= 0.4) = 0.4  
    else:     return 1      # uniform random is always less than 1
```

- ▶ the uniform distribution puts equal weight on all the numbers between 0 and 1.



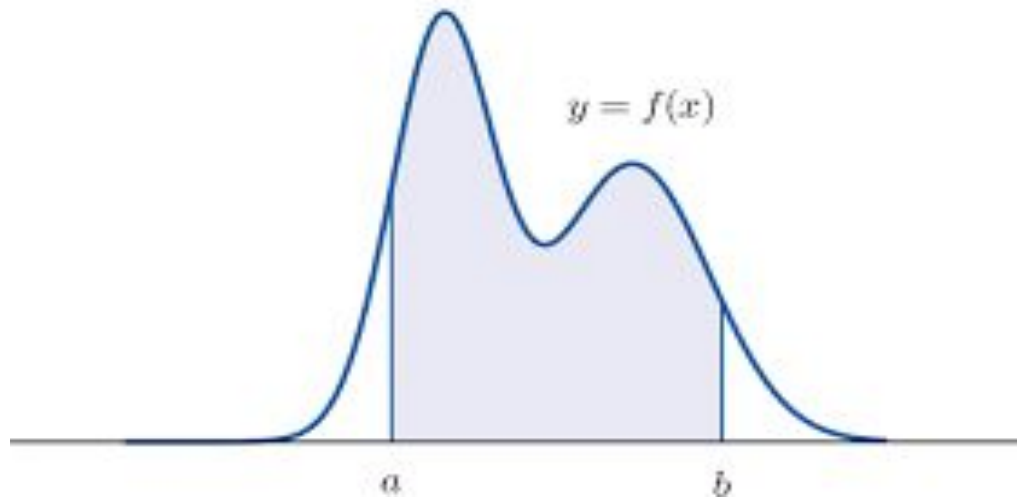
**Cumulative distribution Function of a Uniform Random Variable  $X$**

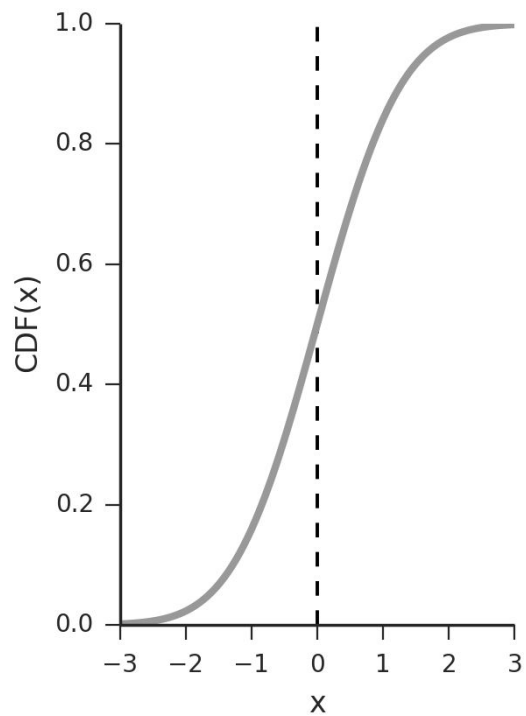
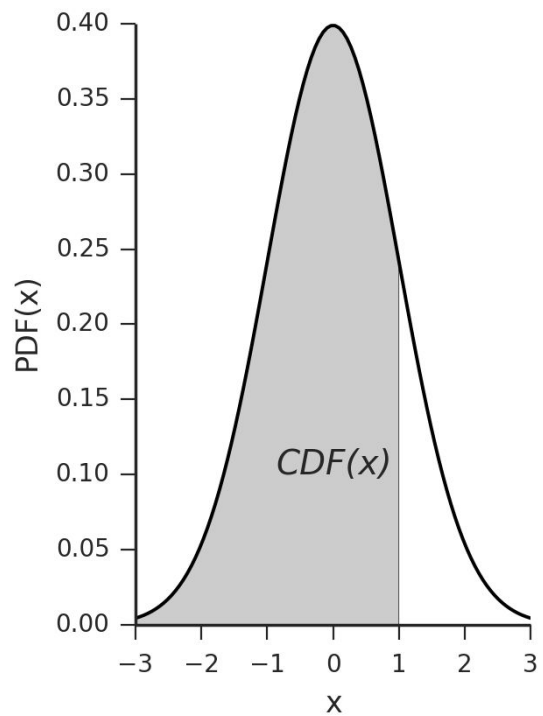
The cumulative distribution function of a uniform random variable  $X$  is:

$$F(x) = \frac{x - a}{b - a}$$

for two constants  $a$  and  $b$  such that  $a < x < b$ . A graph of the c.d.f. looks like this:

$P(a < X < b) = \text{area of shaded region}$





## The Normal Distribution

The normal distribution is the king of distributions. It is the classic bell curve-shaped distribution and is completely determined by two parameters: its mean  $\mu$  (mu) and its standard deviation  $\sigma$  (sigma). The mean indicates where the bell is centered, and the standard deviation how “wide” it is.

It has the distribution function:

$$f(x \mid \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right)$$

- ▶ The two characteristics of the normal distribution are:  
The mean, median, and mode are equal.  
The normal distribution is unimodal and symmetric.

```
def normal_pdf(x, mu=0, sigma=1):  
    sqrt_two_pi = math.sqrt(2 * math.pi)  
    return (math.exp(-(x-mu) ** 2 / 2 / sigma ** 2) / (sqrt_two_pi * sigma))
```

```
xs = [x / 10.0 for x in range(-50, 50)]  
plt.plot(xs, [normal_pdf(x, sigma=1) for x in xs], '-', label='mu=0, sigma=1')  
plt.plot(xs, [normal_pdf(x, sigma=2) for x in xs], '--', label='mu=0, sigma=2')  
plt.plot(xs, [normal_pdf(x, sigma=0.5) for x in xs], ':', label='mu=0, sigma=0.5')  
plt.plot(xs, [normal_pdf(x, mu=-1) for x in xs], '-.', label='mu=-1, sigma=1')  
plt.legend()  
plt.title("Various Normal pdfs")  
plt.show()
```



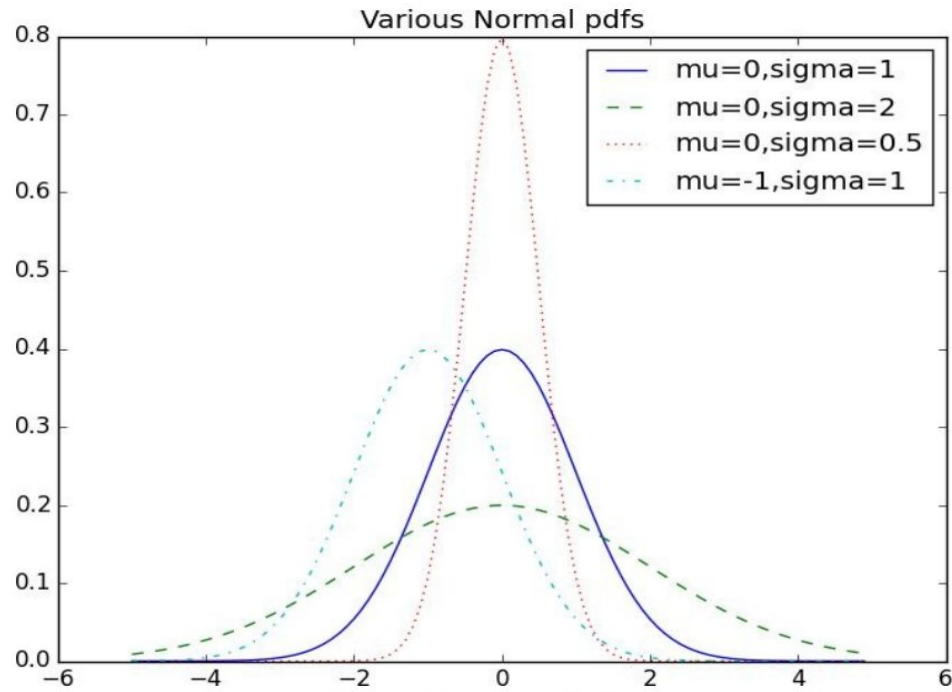


Figure 6-2. Various normal pdfs

- ▶ A standard normal random variable is a normally distributed random variable with mean  $\mu=0$  and standard deviation  $\sigma=1$
- ▶ It will always be denoted by the letter  $Z$

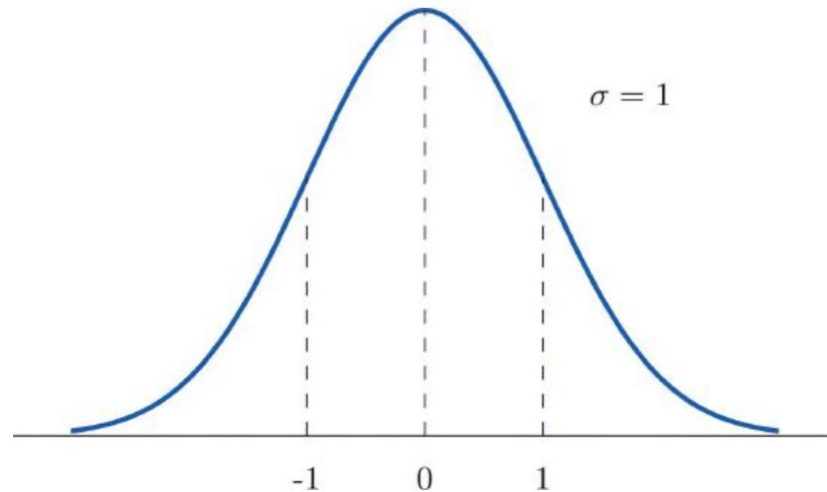


Figure 5.2.1: Density Curve for a Standard Normal Random Variable

When  $\mu = 0$  and  $\sigma = 1$ , it's called the *standard normal distribution*. If  $Z$  is a standard normal random variable, then it turns out that:

$$X = \sigma Z + \mu$$

is also normal but with mean  $\mu$  and standard deviation  $\sigma$ . Conversely, if  $X$  is a normal random variable with mean  $\mu$  and standard deviation  $\sigma$ ,

$$Z = (X - \mu) / \sigma$$

is a standard normal variable.

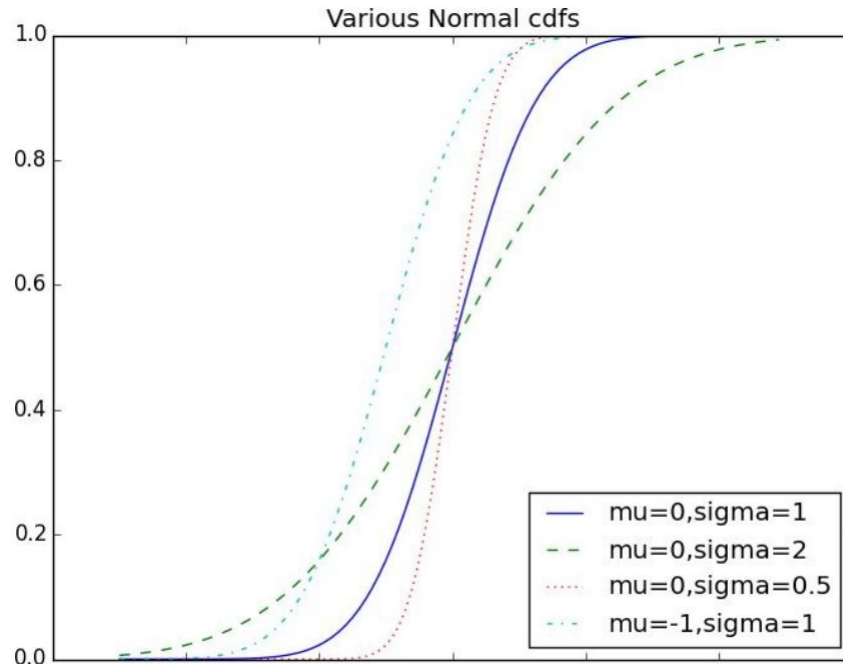
The cumulative distribution function for the normal distribution cannot be written in an “elementary” manner, but we can write it using **Python's `math.erf`**:

```
def normal_cdf(x, mu=0, sigma=1):  
    return (1 + math.erf((x - mu) / math.sqrt(2) / sigma)) / 2
```

The `math.erf()` method returns the error function of a number. This method accepts a value between  $-\infty$  and  $+\infty$ , and returns a value between  $-1$  to  $+1$ .

```
xs = [x / 10.0 for x in range(-50, 50)]
```

```
plt.plot(xs, [normal_cdf(x, sigma=1) for x in xs], '-', label='mu=0, sigma=1')  
plt.plot(xs, [normal_cdf(x, sigma=2) for x in xs], '--', label='mu=0, sigma=2')  
plt.plot(xs, [normal_cdf(x, sigma=0.5) for x in xs], ':', label='mu=0, sigma=0.5')  
plt.plot(xs, [normal_cdf(x, mu=-1) for x in xs], '-.', label='mu=-1, sigma=1')  
plt.legend(loc=4) # bottom right  
plt.title("Various Normal cdfs")  
plt.show()
```



- **The central limit theorem**, says that a random variable defined as the average of a large number of independent and identically distributed random variables is itself approximately normally distributed.

In particular, if  $x_1, \dots, x_n$  are random variables with mean  $\mu$  and standard deviation  $\sigma$ , and if  $n$  is large, then:

$$\frac{1}{n}(x_1 + \dots + x_n)$$

is approximately normally distributed with mean  $\mu$  and standard deviation  $\sigma/\sqrt{n}$ .  
Equivalently (but often more usefully),

$$\frac{(x_1 + \dots + x_n) - \mu n}{\sigma\sqrt{n}}$$

is approximately normally distributed with mean 0 and standard deviation 1.

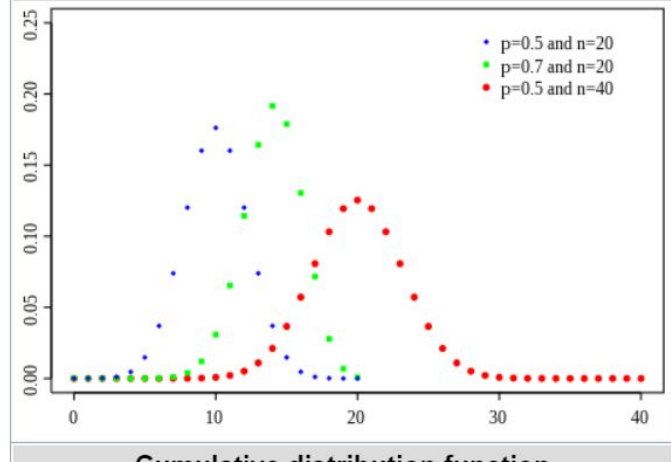
- ▶ The larger the sample size, the more closely the sampling distribution will follow a normal distribution.
- ▶ When the sample size is small, the sampling distribution of the mean is sometimes non-normal. That's because the central limit theorem only holds true when the sample size is “sufficiently large.”
- ▶ By convention, we consider a sample size of 30 to be “sufficiently large.”

# Binomial distribution

- the binomial distribution with parameters  $n$  and  $p$  is the discrete probability distribution of the number of successes in a sequence of  $n$  independent experiments, each asking a yes-no question, and each with its own Boolean-valued outcome: success (with probability  $p$ ) or failure (with probability  $q=1-p$ ).

**Binomial distribution**

**Probability mass function**



# Binomial random variable

- ▶ This is a specific type of discrete random variable. A binomial random variable counts how often a particular event occurs in a fixed number of tries or trials. For a variable to be a binomial random variable, **ALL** of the following conditions must be met:
  - There are a fixed number of trials (a fixed sample size).
  - On each trial, the event of interest either occurs or does not.
  - The probability of occurrence (or not) is the same on each trial.
  - Trials are independent of one another.
- ▶ **Notation:**
  - ▶  $n$  = number of trials (sample size)
  - ▶  $p$  = probability event of interest occurs on any one trial



► **Examples of binomial random variables:**

- Number of correct guesses at 30 true-false questions when you randomly guess all answers
- Number of winning lottery tickets when you buy 10 tickets of the same kind
- Number of left-handers in a randomly selected sample of 100 unrelated people

For the guessing at true questions example above,  $n = 30$  and  $p = .5$  (chance of getting any one question right).

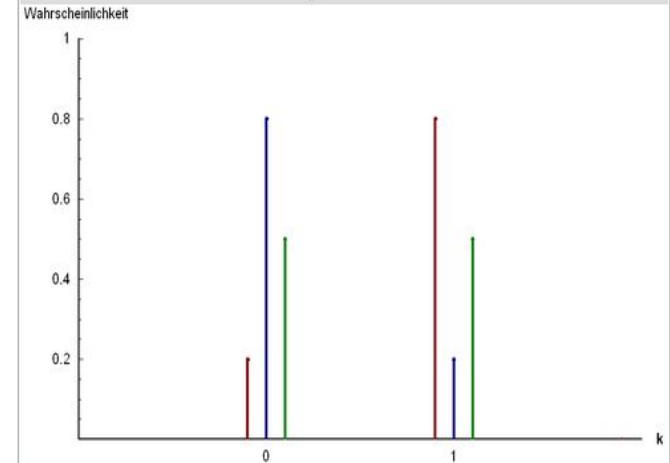
- ▶ A single success/failure experiment is also called a Bernoulli trial or Bernoulli experiment, and a sequence of outcomes is called a Bernoulli process; for a single trial, i.e.,  $n = 1$ ,

# Bernoulli trial

- ▶ Trial with only 2 possible outcomes.
  - ▶ Success (P) probability of success
  - ▶ Failure (1-P) probability of failure.
  - ▶ If X is a random variable then  $X=1$  for success and  $X = 0$  for failure.
  - ▶ Then we can say X is a random variable that follows Bernoulli distribution.
  - ▶ Eg: coin flip.
  - ▶ success=  $\frac{1}{2}$  and failure is  $1 - \frac{1}{2}$

## Bernoulli distribution

### Probability mass function



Three examples of Bernoulli distribution:

- $P(x = 0) = 0.2$  and  $P(x = 1) = 0.8$
- $P(x = 0) = 0.8$  and  $P(x = 1) = 0.2$
- $P(x = 0) = 0.5$  and  $P(x = 1) = 0.5$

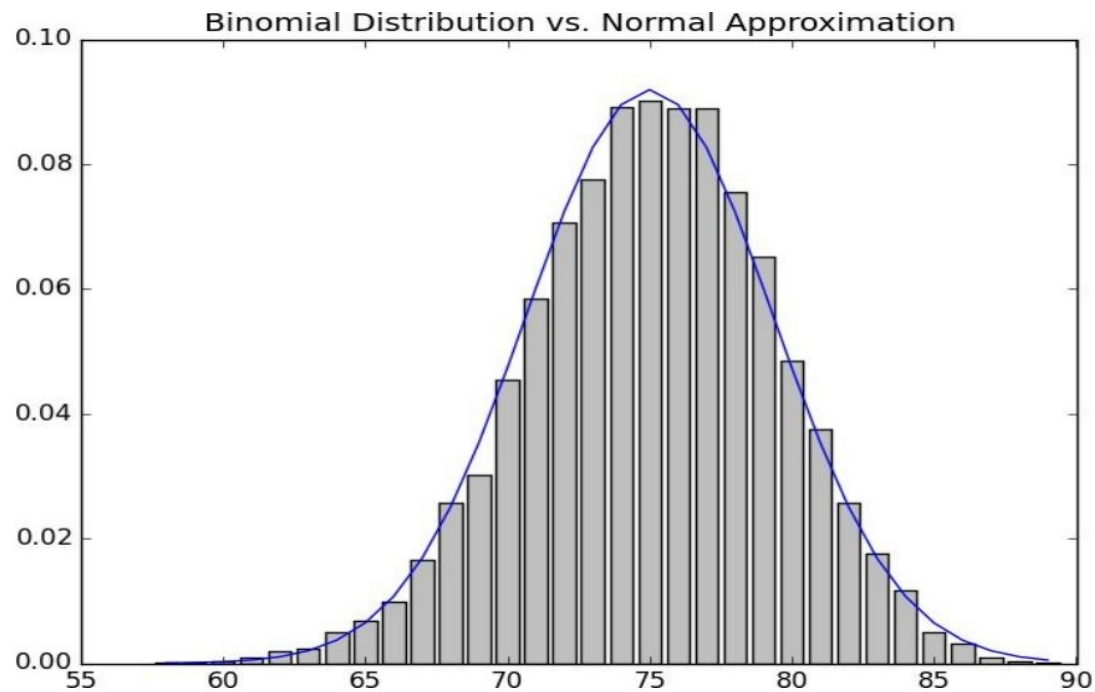


Figure 6-4. The output from `make_hist`