APMA E4990.02: Review of Linear Algebra, Probability and Calculus.

Vectors, Matrices and Solvability
Probability, Expectation, Variance

Outline

- Why do we need to review linear algebra?
- Vectors and Matrices.
- Matrix Multiplication
- Inverses and Transpose
- When do solutions exist, and when are they unique?
- What is a probability? What is Expectation and Variance?
- Conditional probabilities and joint distributions.
- Central Limit Theorem
- Bayesian Statistics

What should you have done by now?

- Installed **Github** and viewed the first assignment on the repo. Clicked the invitation link in the notebook.

- Installed Anaconda and can open Jupyter notebook. If you are having trouble with this, please contact the TA.

- Joined the Slack channel and added your name to the spreadsheet shown in the last lecture.

Why do we need linear algebra?

- All of data science is built around constructing models using a collection of ‘features’ or ‘predictors’ (ie. predict income from GPA). It is almost always the case that **there are several variables involved.** If there is **more than one, we need linear algebra to make sense of any equations.**
- For example, let y be the number of rooms that will be booked at a hotel tomorrow, and X_i be the ith observation of N variables, which could include **location, cost, rating** , etc. We can write a simple linear model as:

```
In [53]: df.head()
```

Out[53]:

	account	const	hotel_rating	location	price_per_night_avg	purchase_velocity_lastweek	rooms_left	sellouts_total
0	72722	1	1.6	9.5	584	60	198	2
1	20627	1	8.2	9.2	503	326	439	8
2	55924	1	8.0	0.9	467	327	240	8
3	14773	1	9.5	9.9	543	102	286	4
4	60469	1	1.8	5.7	144	42	335	2

$$y_i = \sum_{k=1}^{N} \beta_k X_{ik} + c \qquad \text{or} \qquad \mathbf{y} = \beta \cdot \mathbf{X} + \mathbf{c}$$

Advertising Data

$\mathbf{X} =$

	TV	radio	newspaper
1	230.1	37.8	69.2
2	44.5	39.3	45.1
3	17.2	45.9	69.3
4	151.5	41.3	58.5
5	180.8	10.8	58.4

Matrix

$Y =$

sales
22.1
10.4
9.3
18.5
12.9

Vector

$\hat{Z}$

Hat represents an estimator of a statistical quantity Z from the data.

For example:

$$\hat{\mathbb{E}}(Y) := \frac{1}{N}\sum_{i=1}^{N} Y_i$$

$(\mathbf{x}_i, y_i)$ Training sample

$\mathbf{x}_i$ Training of x only

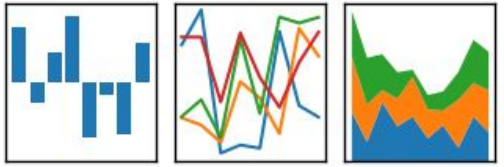
$\mathbf{x}$ Particular fixed value of x

Don’t worry if it seems like a lot at first. We will ease into it over time.

Previewing the data

pandas

$$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$$



```
In [ ]: import pandas as pd

        df = pd.read_csv('sales_data.csv')
```

```
In [53]: df.head()
```

Out[53]:

	account	const	hotel_rating	location	price_per_night_avg	purchase_velocity_lastweek	rooms_left	sellouts_total
0	72722	1	1.6	9.5	584	60	198	2
1	20627	1	8.2	9.2	503	326	439	8
2	55924	1	8.0	0.9	467	327	240	8
3	14773	1	9.5	9.9	543	102	286	4
4	60469	1	1.8	5.7	144	42	335	2

This is done in a jupyter notebook.

Basic Properties of Matrices

$$\begin{aligned} 4x - 2y + 3z &= 4 \\ 5x + 3y - z &= 2 \\ 3x + 2y - 6z &= 8 \end{aligned}$$

Matrices are used to express systems of equations. They can be interpreted as representations of linear operators in a particular basis.

The above can be written as

$$\mathbf{Ax} = \mathbf{b}$$

$$\mathbf{A} = [a_{ij}]$$

$$\mathbf{b} = [4, 2, 8]$$

$$a_{11} = 4, a_{12} = -2, a_{23} == -1, \text{etc}$$

Here, **b** is considered a **vector.**

Matrix Addition - Done by components

$$\begin{aligned} 4x - 2y + 3z &= 4 \\ 5x + 3y - z &= 2 \\ 3x + 2y - 6z &= 8 \end{aligned} \qquad \begin{aligned} 5x - 3y + 4z &= 6 \\ 2x + y - z &= 4 \\ 5x + 7y - 3z &= 1 \end{aligned}$$

$$\mathbf{Ax = b} \qquad \mathbf{Cx = d}$$

$$\mathbf{(A + C)x := Dx = b + d := e} \qquad \begin{aligned} 9x - 5y + 7z &= 10 \\ 7x + 4y - 2z &= 6 \\ 8x - 5y - 9z &= 9 \end{aligned}$$

Matrix Multiplication

$$\begin{aligned} 4x - 2y + 3z &= 4 \\ 5x + 3y - z &= 2 \\ 3x + 2y - 6z &= 8 \end{aligned}$$

$$\begin{aligned} 5x - 3y + 4z &= 6 \\ 2x + y - z &= 4 \\ 5x + 7y - 3z &= 1 \end{aligned}$$

$$\mathbf{Ax} = \mathbf{b}$$

$$\mathbf{Cx} = \mathbf{d}$$

$$\mathbf{AB} := \mathbf{M} = [m_i j] := \sum_{i,j} a_{ik} b_{kj}$$

Matrix Multiplication

$$\begin{bmatrix} 4 & -2 & 3 \\ 5 & 3 & -1 \\ 3 & 2 & -6 \end{bmatrix} \begin{bmatrix} 5 & -3 & 4 \\ 2 & 1 & -1 \\ 5 & 7 & -3 \end{bmatrix}$$

$\mathbf{A}$ $\mathbf{B}$

$$\mathbf{AB} := \mathbf{M} = [m_{ij}] := \sum_k a_{ik} b_{kj}$$

Ith row of A, dotted with j th column of B

$$m_{11} = 4 \cdot 5 - 2 \cdot 2 + 3 \cdot 5$$
$$m_{12} = 4 \cdot (-3) - 2 \cdot 1 + 3 \cdot 7$$

Note: $\mathbf{AB} \neq \mathbf{BA}$

Matrix Multiplication

$$\begin{bmatrix} 4 & -2 & 3 \\ 5 & 3 & -1 \\ 3 & 2 & -6 \end{bmatrix} \begin{bmatrix} 5 & -3 & 4 \\ 2 & 1 & -1 \\ 5 & 7 & -3 \end{bmatrix}$$

$\mathbf{A}$ $\mathbf{B}$

$$\mathbf{AB} := \mathbf{M} = [m_i j] := \sum_{i,j} a_{ik} b_{kj}$$

ith row of A, dotted with j th column of B

$$\mathbf{M} = \begin{bmatrix} 31 & 7 & 9 \\ 26 & -19 & 20 \\ -11 & -49 & 28 \end{bmatrix}$$

Following this, we obtain
(**Exercise:** Verify this if you are rusty)

How does this relate to us?

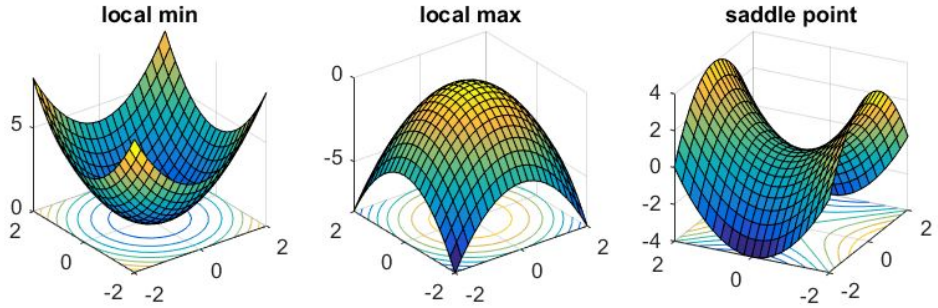
- Matrices are basis representations of linear operators.
- The rows for us will always be observations, and the columns will be the variables themselves.

$$y_i = \sum_{k=1}^{N} \beta_k X_{ik} + c \qquad \text{or} \qquad \mathbf{y} = \beta \cdot \mathbf{X} + \mathbf{c}$$

Vector Calculus

Convex Analysis (Calculus)

- Our goal is often find some convex function such that we learn the rules of our model by its minimization (ie. coefficients). These are referred to as **parametric methods.**
- We can also “fit” data to a probability by **maximizing** the **likelihood** - concave problem.
- This is **not always the case**, with algorithms such as **decision trees** or **neural nets**. However, it covers many cases in supervised learning. These are called **non-parametric methods**



Training Data Notation

$$\mathbf{X} = \begin{pmatrix} x_{11} & \cdots & x_{1k} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nk} \end{pmatrix} \quad \text{and} \quad \mathbf{Y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$$

$$\mathbf{x}_i = \begin{pmatrix} x_{i1} \\ \vdots \\ x_{ik} \end{pmatrix} \qquad \mathbf{x}_i^T = \begin{pmatrix} x_{i1} & x_{i2} \cdots & x_{ik} \end{pmatrix}$$

We will assume that vectors are by default column vectors. When we take the transpose, we end up with a row vector.

Matrix Notation and Dot Products

$$F(\beta) := \frac{1}{N}\sum_{i=1}^{N}(y_i - \beta^T \cdot \mathbf{x}_i)^2$$

$$\mathbf{X} = \begin{pmatrix} x_{11} & \cdots & x_{1k} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nk} \end{pmatrix} \text{ and } \mathbf{Y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$$

It is often far more convenient to rewrite the above sum in terms of matrix algebra and inner products

$$F(\beta) = \frac{1}{N}\|Y - \beta\mathbf{X}\|^2 = \frac{1}{N}\langle Y - \beta^T\mathbf{X}, Y - \beta^T\mathbf{X}\rangle = (Y - \beta^T\mathbf{X})^T(Y - \beta^T\mathbf{X})\,\frac{1}{N}$$

We will assume that our vectors are column vectors and that the transpose makes them row vectors

Normal dot product.

$$\mathbf{a}^T\mathbf{b} = \sum_i a_i b_i$$

Gradients of multivariable functions

$$f : \mathbb{R}^k \to \mathbb{R}$$

$$\nabla f = \begin{pmatrix} \frac{\partial f}{\partial \beta_1} \\ \vdots \\ \frac{\partial f}{\partial \beta_k} \end{pmatrix} \quad \text{where } \beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} \quad F(\beta) := \frac{1}{N} \sum_{i=1}^{N} (y_i - \beta^T \cdot \mathbf{x}_i)^2$$

Recall that the gradient is the vector of partial derivatives.

In this course, we often want to maximize over a collection of parameters for our model.

Gradients of multivariable functions

$$F(\beta) := \frac{1}{N}\sum_{i=1}^{N}(y_i - \beta^T \cdot \mathbf{x}_i)^2 \qquad \frac{\partial f}{\partial \beta_j} = \frac{2}{N}\sum_{i=1}^{N}(y_i - \beta^T \mathbf{x}_i)(-x_{ij})$$

$$F(\beta) = \frac{1}{N}\|Y - \beta\mathbf{X}\|^2 = \frac{1}{N}\langle Y - \beta^T\mathbf{X}, Y - \beta^T\mathbf{X}\rangle = (Y - \beta^T\mathbf{X})^T(Y - \beta^T\mathbf{X})$$

$$\nabla F(\beta) = \frac{1}{N}\langle Y - \beta^T\mathbf{X}, -\mathbf{X}\rangle = \frac{1}{N}(-Y^T\mathbf{X} + \mathbf{X}^T\mathbf{X}\beta)$$

The second form is much easier to read once you're comfortable with matrix notation. More on this in **Homework 0.**

Existence and Uniqueness of Solutions

Inverse, Transpose and Fredholm Alternative

When do systems of equations have solutions?

$$\begin{aligned} 4x - 2y + 3z &= 4 \\ 5x + 3y - z &= 2 \\ 3x + 2y - 6z &= 8 \end{aligned}$$

$$\mathbf{Ax} = \mathbf{b}$$

$$x = \mathbf{A}^{-1}\mathbf{b}$$

$$\begin{bmatrix} 4 & -2 & 3 \\ 5 & 3 & -1 \\ 3 & 2 & -6 \end{bmatrix}$$

$$\mathbf{A}^{-1} = \begin{bmatrix} 16/115 & 6/115 & 7/115 \\ -27/115 & 33/115 & -19/115 \\ -1/115 & 14/115 & -22/115 \end{bmatrix}$$

Solution is computed using by reducing to Reduced Row Echelon form. Since this isn't essential for this course, we are going to omit the details.

When do systems of equations have solutions?

$$\mathbf{Ax} = \mathbf{b}$$

$$x = \mathbf{A}^{-1}\mathbf{b}$$

- Inverses exist only for square matrices (ie. column dimension is the same as the row dimension)

$$\mathbf{A} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \qquad \mathbf{b} = [4, 2]$$

No solution

$$\mathbf{A} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \qquad \mathbf{b} = [4, 0]$$

Infinitely many solutions

$$\mathbf{A} = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix} \qquad \mathbf{b}$$

Exactly one solution

- Review this if you are not familiar. Will be important when we cover **regularization and collinearity.**
- The main point to take home here is to understand that **these three possibilities exist and only these**.

When does this have a unique solution?

$$\nabla F(\beta) = \frac{1}{N}\langle Y - \beta^T \mathbf{X}, -\mathbf{X}\rangle = \frac{1}{N}(-Y^T\mathbf{X} + \mathbf{X}^T\mathbf{X}\beta) = 0$$

Homework 0 will cover vector calculus topics in addition to others.

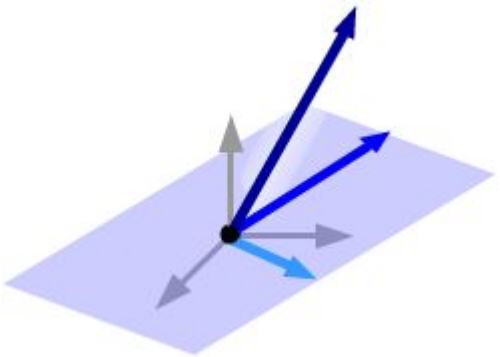
Eigenvalues, Eigenvectors and Linear Dependence.

Important for dimensionality reduction and stability.

Linear Independence

A collection of vectors (or features in our case), are **linearly dependent** if and only if

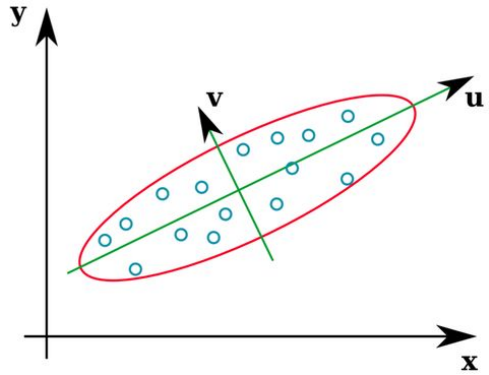
$$\sum_{j=1}^{N} \alpha_j X_j = 0 \text{ for some } \{\alpha_j\} \in \mathbb{R}$$



Otherwise, they are **linearly independent.**

- **Important since linear dependent vectors cause instability in solutions, and take up unnecessary space/resources.**
- **We also often want to find a minimal spanning set of a high dimensional space (dimensionality reduction).**

Eigenvalues and Eigenvectors

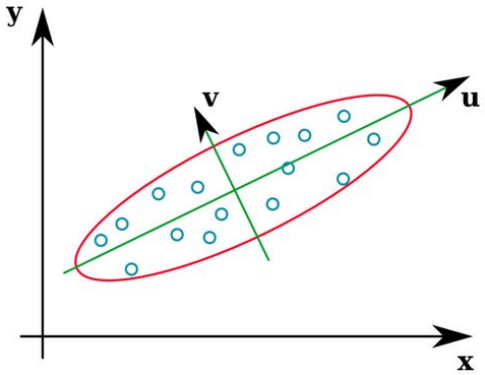


$$Ax = \lambda x \quad \lambda \in \mathbb{R}$$

$$x \in \mathbb{R}^n$$

- If A is invertible, there exists an orthogonal set of eigenvectors x and eigenvalues $\lambda \in \mathbb{R}$
- These are important for dimensionality reduction and for stability of solutions to ordinary least squares (next deck).

Eigenvalues and Eigenvectors



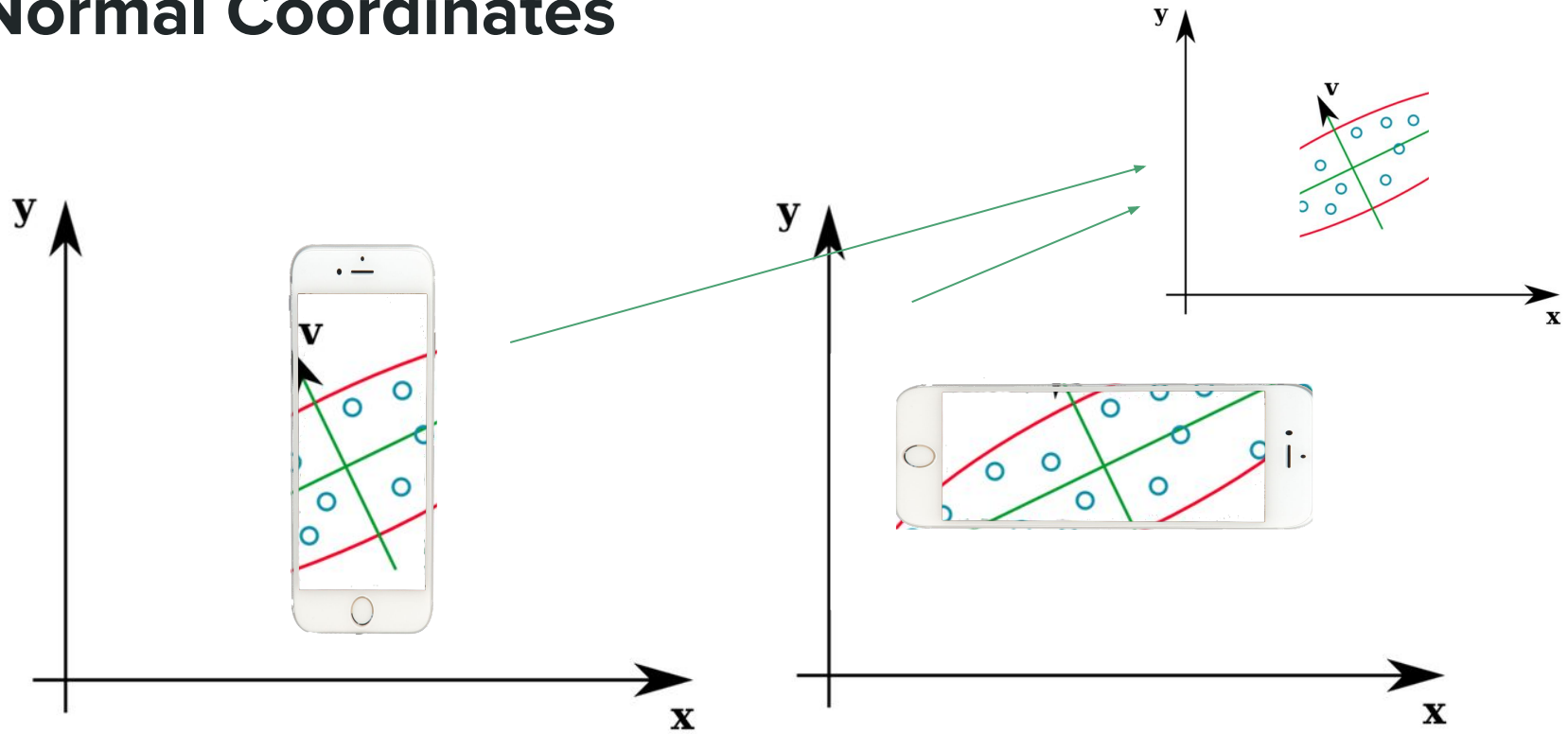
$$X^T X$$

- The matrix here will be common in the course. It is positive semi-definite.

$$y^T X^T X y \geq 0$$

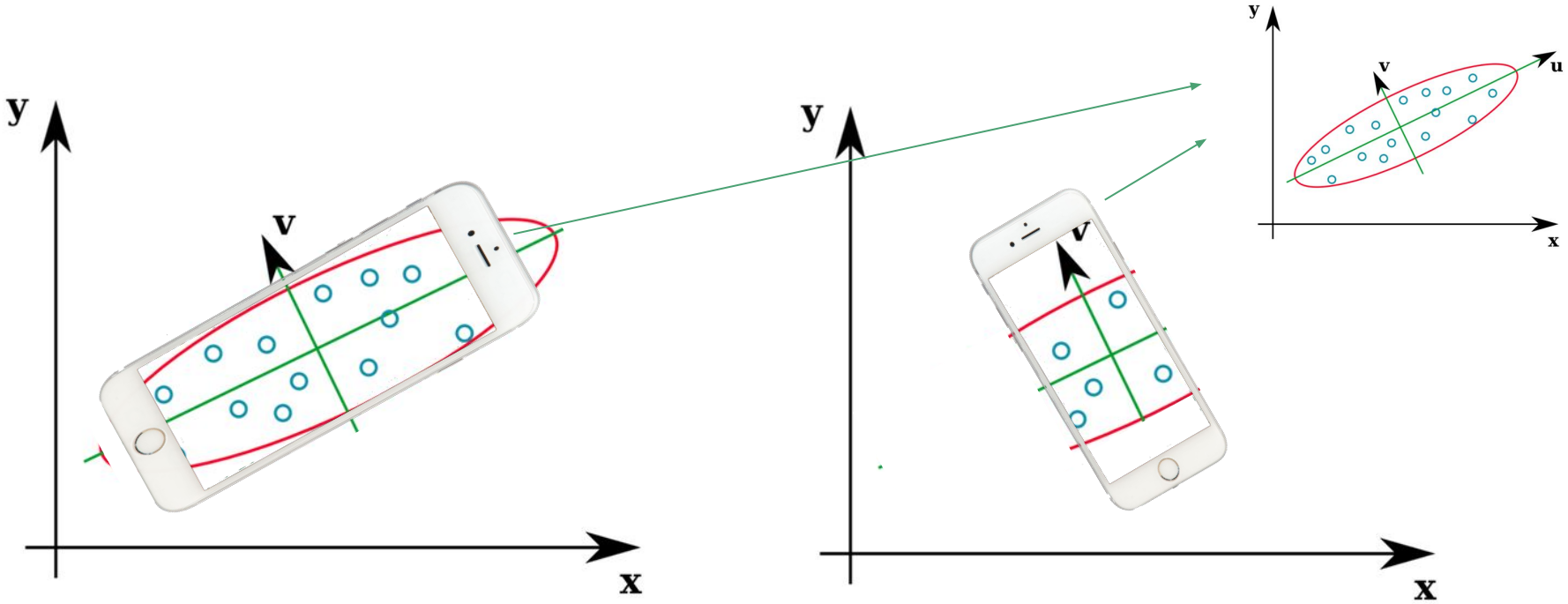
- The matrix measures the dependence between features when X is your data. If your data is mean centered and normalized, it's the same as the correlation matrix.

Normal Coordinates



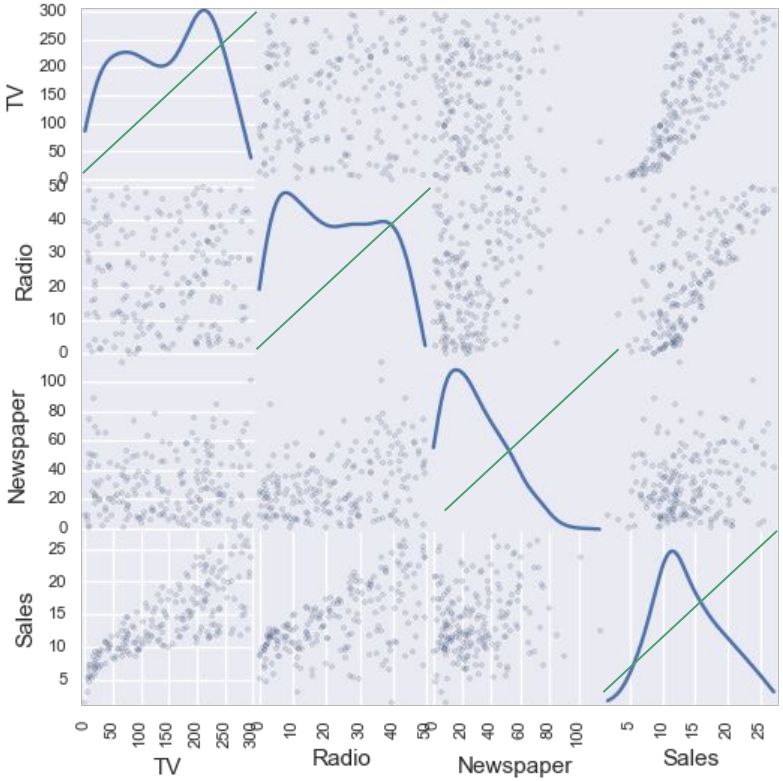
Imagine you can only take a photo of these points in two directions - which ones would you choose?

Eigenvectors



By rotating our camera to “good” angles, we capture more about the distribution of the data! These are the eigenvectors (ie. PCA).

Correlation plots



$$X^T X$$

- Sales related to advertising revenue on TV, Radio and Newspaper
- Content of Homework 1

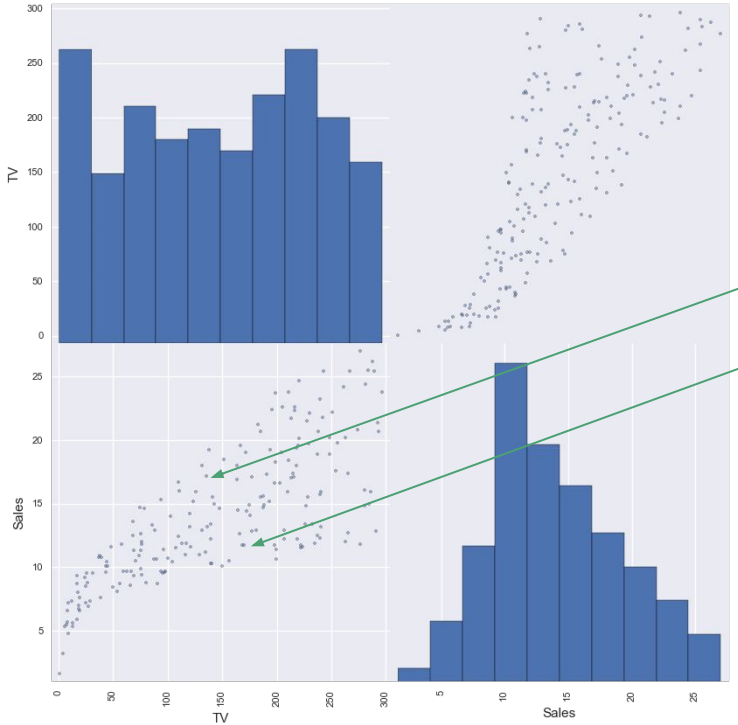
Out[20]:

	TV	Radio	Newspaper	Sales
1	230.1	37.8	69.2	22.1
2	44.5	39.3	45.1	10.4
3	17.2	45.9	69.3	9.3
4	151.5	41.3	58.5	18.5
5	180.8	10.8	58.4	12.9

What are the features?

- **TV:** advertising dollars spent on TV for a single product in a given market (in thousands of dollars)
- **Radio:** advertising dollars spent on Radio
- **Newspaper:** advertising dollars spent on Newspaper

Understanding correlation



$$\mathbb{R}^{2\times n} \times \mathbb{R}^{n\times 2}$$

$$X^T X = \begin{bmatrix} 150 & 160 & \cdots \\ 17 & 12 & \cdots \end{bmatrix} \begin{bmatrix} 150 & 17 \\ 160 & 12 \\ \cdots & \cdots \end{bmatrix}$$

$$\mathbb{R}^{2\times 2}$$

$$[X^T X]_{ij} = \mathbf{x_i} \cdot \mathbf{x}_j$$

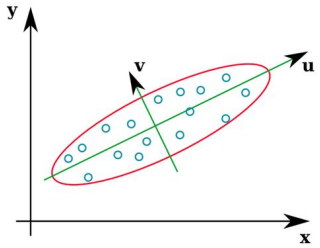
$$i = \text{TV}$$

$$j = \text{Sales}$$

Vectors are in n dimensions (number of data points)

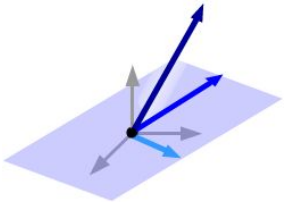
$$[\text{Corr}]_{ij} = \frac{\mathbf{x_i} \cdot \mathbf{x}_j}{\|\mathbf{x_i}\|\|\mathbf{x_j}\|} = \cos(\delta_{ij})$$

MNIST- Cosine Distance



$$\frac{X^T X}{||X||^2}$$

$$\mathbf{x} \cdot \mathbf{y} = \|\mathbf{x}\|\|\mathbf{y}\| \cos(\theta)$$



- The above matrix can be interpreted as the **correlation matrix**, or equivalently here, the **cosine distance**.
- Used to measure the **distance of one feature to another.**
- Recall the MNIST example.

```
In [94]: df = pd.read_csv('/Users/dgoldmal/Downloads/train.csv')

In [95]: df.head()
```

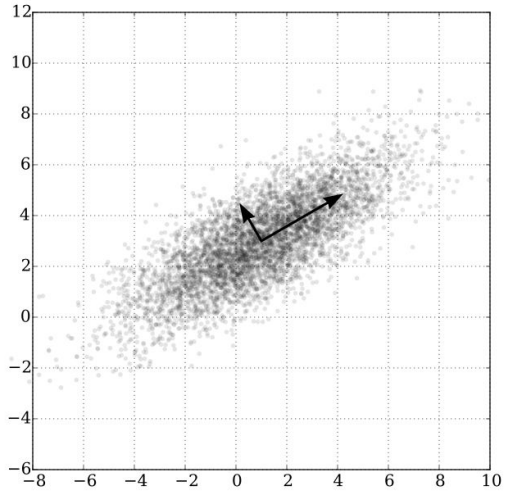
Out[95]:

ixel117	pixel118	pixel119	pixel120	pixel121	pixel122	pixel123	pixel124	pixel125	pixel126	pixel127	pixel128	pixel129	pixel130	pixel131	pixel13
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	188
	0	0	0	0	18	30	137	137	192	86	72	1	0	0	0
	0	0	0	0	0	0	3	141	139	3	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	0	0	0	1	25	130	155	254	254	254	157	30	2	0	0

```
In [81]: import
```

From the MINST Database of Hand–written Digits

Explaining Variance



$$\max_{\mathbf{v}} \frac{\mathbf{v}^T X^T X \mathbf{v}}{\|\mathbf{v}\|^2}$$

$$X^T X \mathbf{v} = \lambda \mathbf{v}$$

- Plot to the left is the first two eigenvectors with the largest eigenvalues. Note, **data is assumed to be mean centered.**
- Can be interpreted as - the two components with the largest ‘variance’ - or in other words, the **most significant directions in the 28x28 dimensional space.**

```
In [94]: df = pd.read_csv('/Users/dgoldmal/Downloads/train.csv')

In [95]: df.head()
```

Out[95]:

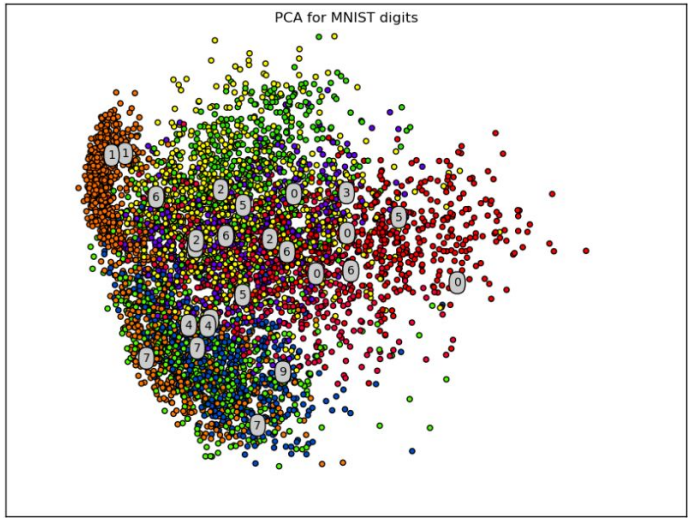
ixel117	pixel118	pixel119	pixel120	pixel121	pixel122	pixel123	pixel124	pixel125	pixel126	pixel127	pixel128	pixel129	pixel130	pixel131	pixel13
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	188
	0	0	0	0	18	30	137	137	192	86	72	1	0	0	0
	0	0	0	0	0	0	3	141	139	3	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	0	0	0	1	25	130	155	254	254	254	157	30	2	0	0

```
In [81]: import
```

Top Eigenvalues - PCA

$$\max_{\mathbf{v}} \frac{\mathbf{v}^T X^T X \mathbf{v}}{\|\mathbf{v}\|^2}$$

$$X^T X \mathbf{v} = \lambda \mathbf{v}$$



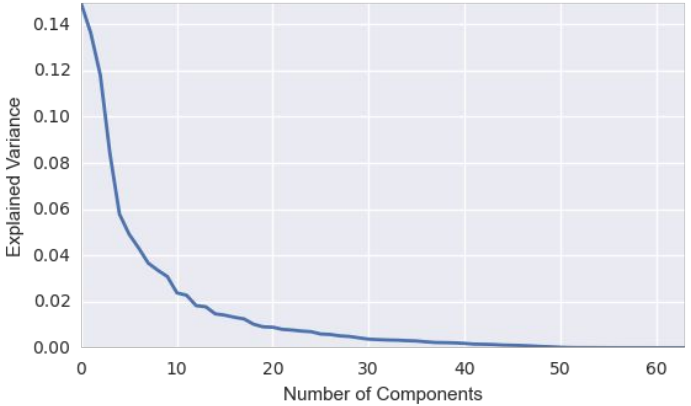
In [94]:
```
df = pd.read_csv('/Users/dgoldma1/Downloads/train.csv')
```

In [95]:
```
df.head()
```

Out[95]:

ixel117	pixel118	pixel119	pixel120	pixel121	pixel122	pixel123	pixel124	pixel125	pixel126	pixel127	pixel128	pixel129	pixel130	pixel131	pixel13
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	188
	0	0	0	0	18	30	137	137	192	86	72	1	0	0	0
	0	0	0	0	0	0	3	141	139	3	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	0	0	0	1	25	130	155	254	254	254	157	30	2	0	0

In [81]:
```
import
```



Review of Probability

Expectation, Variance and Conditional Distributions

Outline for Probability Review

- Definitions of a probability, expectation, variance.
- Basic examples and some problems we will work out.
- Conditional probability distributions
  - Conditional Expectation
  - Linearity of Expectation
  - Law of Total Expectation
- Bayes Theorem
- Central Limit Theorem

Some definitions

- A probability p takes events $\{A_i\}$ as inputs, and outputs a value between 0 and 1 inclusive.

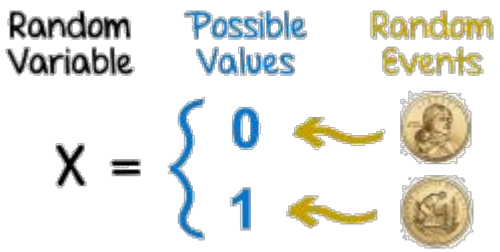
- The closer to 1, the more probable the event. The closer to 0, the less probable.

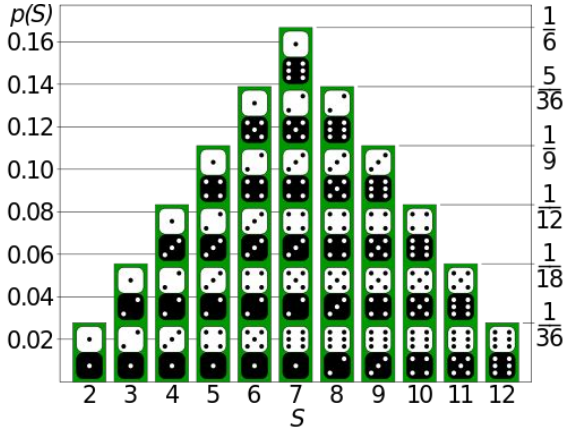
$$\sum_i p(A_i) = 1$$

$$P\left(\bigcup_i A_i\right) = \sum_i p(A_i) \quad \text{when } \{A_i\} \text{ are disjoint}$$

Some definitions

Y and X denote **Random Variables** - they are just the outcome of experiments, such as rolling a die, or how many newspapers will be purchased at Starbucks.



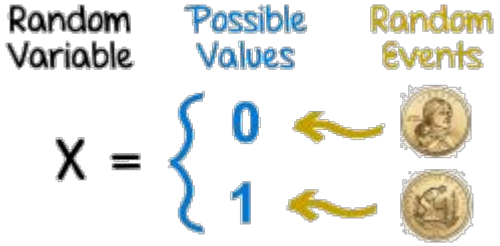


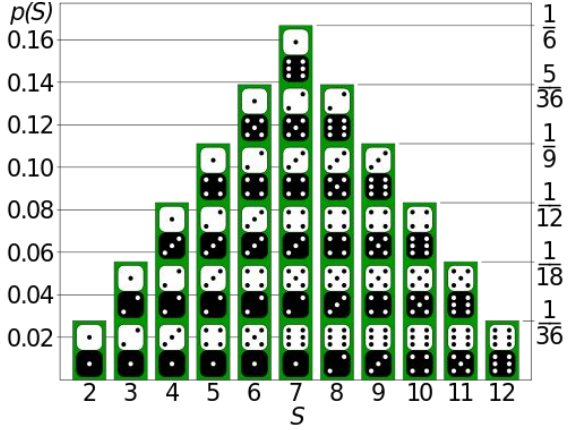
Anything which represents the outcome of an event which is not deterministic

Some definitions

X and Y are said to be **independent**, if

$$P(X = x \text{ and } Y = y) = P(X = x)P(Y = y)$$





More examples later in this deck.

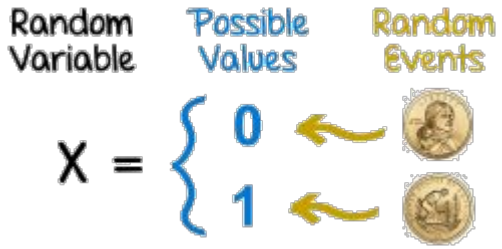
Examples:

- Flipping two coins or dies one after another.
- Do two different users click on an article?

Non-Examples:

- Only flipping the second one if the first is heads.
- Does the same user click on an article today and tomorrow?

Independence



Independent events are absolutely crucial to have for rows in your dataframe!

```
In [53]: df.head()
```

Out[53]:

	account	const	hotel_rating	location	price_per_night_avg	purchase_velocity_lastweek	rooms_left	sellouts_total
0	72722	1	1.6	9.5	584	60	198	2
1	20627	1	8.2	9.2	503	326	439	8
2	55924	1	8.0	0.9	467	327	240	8
3	14773	1	9.5	9.9	543	102	286	4
4	60469	1	1.8	5.7	144	42	335	2

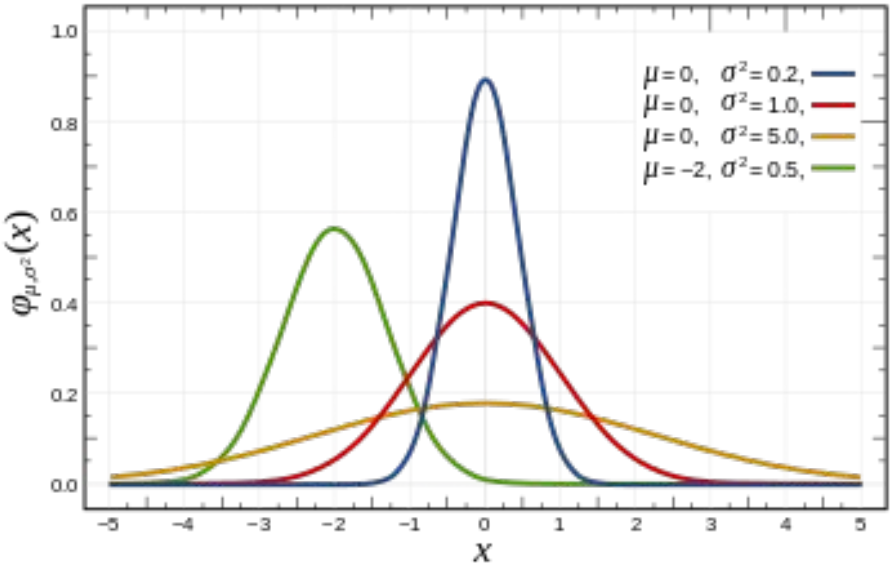
- As an extreme example, imagine **all of your rows were identical.**
- What about if we were predicting rooms left in a hotel, and **our dataframe was sorted? Why is this bad?**

Normal Random Variables

When we write

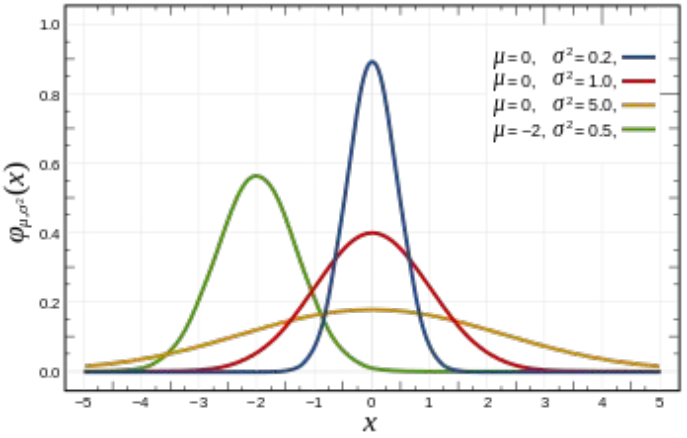
$$X \sim \mathcal{N}(\mu, \sigma^2)$$

We mean that X has mean μ and Variance σ^2 and is distributed normally:



$$p(X|\mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}$$

Examples of Normal Distributions



Examples:

- Age, height, IQ of the population.
- Number of clicks that an article gets.
- The number of heads in a large number of coin tosses.
- Any kind of counting data for which you have a **sufficient number of samples -** this is related to the **Central Limit Theorem**, which will be covered in this class.

$$p(X|\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}$$

- This is why we will often assume the error in a linear regression is normally distributed - **this will be covered in more detail later.**

Properties of summing random variables

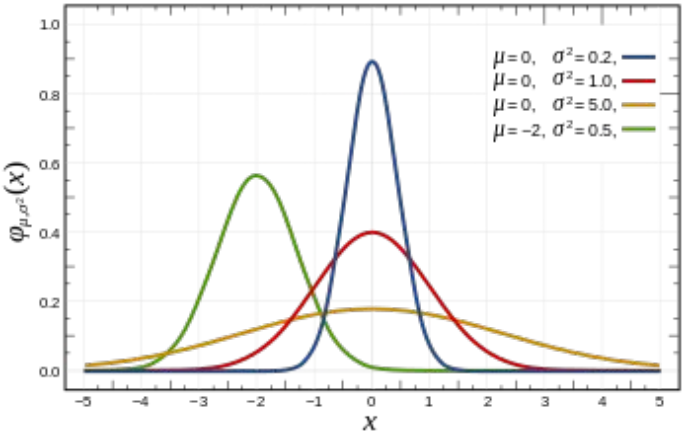
$$X \sim \mathcal{N}(\mu_X, \sigma_X^2)$$

$$Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$$

$$Z = X + Y$$

If X and Y are **independent,** then:

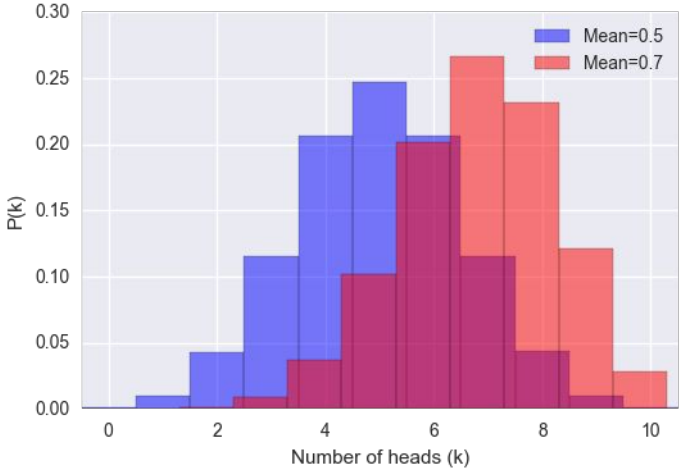
$$Z \sim \mathcal{N}(\mu_X + \mu_Y, \sigma_X^2 + {}_Y^2)$$

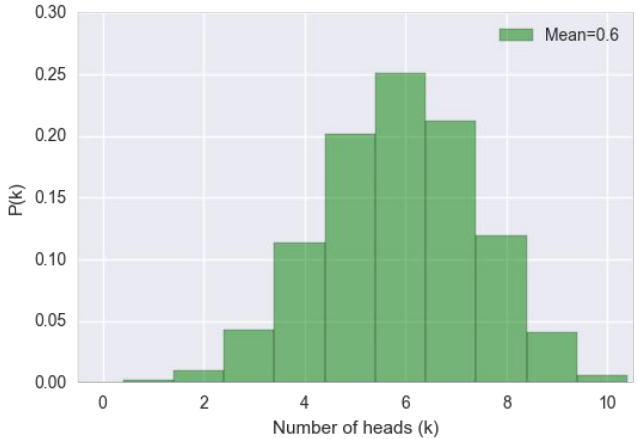


Intuition of Reason

- The number of heads when flipping a coin many times is (approximately) normally distributed.
- If you flip **two coins** N times, it's the same as flipping one coin with the average mean of the two coins 2N times (ie. another coin!), which is also normally distributed.

Sum of Two Normal Random Variables





$$Z \sim \mathcal{N}(\mu_X + \mu_Y, \sigma_X^2 + {}^2_Y)$$

https://github.com/Columbia-Intro-Data-Science/APMAE4990-/blob/master/notebooks/Example%20of%20summing%20two%20normal%20random%20variables.ipynb

Expectation and Variance

- The **expected value** of a random variable Y is the sum over all outcomes x the probability of that outcome - it's the same as the mean, but has a different interpretation.

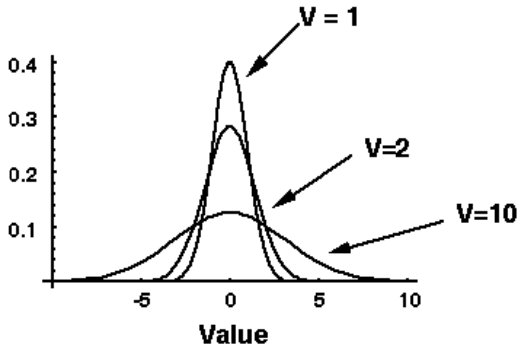
$$\mathbb{E}(Y) = \sum_{i=1}^{N} y_i p(y_i)$$

- The **variance** of a random variable is the L_2 distance to the mean with respect to the probability distribution. It's the same as the standard deviation when calculating on a list of numbers.

$$\mathrm{Var}(Y) = \sum_{i=1}^{N} (y_i - \mathbb{E}(y))^2 p(y_i)$$

Special case for uniformly distributed data:

$$p(y_i) = \frac{1}{N}$$



Example - Coin Flipping

$$p(Y = k) = \begin{cases} p & \text{if } k \text{ is } 1 \\ 1 - p & \text{if } k \text{ is } 0 \end{cases}$$

What is the mean and variance of the random variable Y?

This is known as the Bernoulli distribution, and is very important in machine learning and for A/B tests.

Example - Coin Flipping

$$p(Y = k) = \begin{cases} p & \text{if } k \text{ is } 1 \\ 1 - p & \text{if } k \text{ is } 0 \end{cases}$$

$$\mathbb{E}(Y) = p \cdot (1) + (1 - p) \cdot 0 = p$$

$$\text{Var}(Y) = (1 - p)^2 p + (0 - p)^2 (1 - p) = p - 2p^2 + p^2 = p(1 - p)$$

This is known as the Bernoulli distribution, and is very important in machine learning and for A/B tests.

Example - Dice Rolling

$$p(X = k) = \frac{1}{6}$$

What is the mean and variance of the random variable X?

Example - Dice Rolling

$$p(X = k) = \frac{1}{6}$$

$$\mathbb{E}(X) = \sum_{k=1}^{6} \frac{k}{6} = 3.5$$

$$\mathrm{Var}(X) = \frac{1}{6} \sum_{k=1}^{6} (k - 3.5)^2 = \frac{17.5}{6}$$

Joint Distributions & Conditional Probability

Suppose the discrete random variables X and Y have supports S_x and S_y. Then we require P to satisfy

$$\sum_{x \in S_x, y \in S_y} P(X = x, Y = y) = 1$$

$$P(X|Y = y) = \frac{P(X \text{ and } Y = y)}{P(y)}$$

	$Y = -1$	$Y = 0$	$Y = 1$
$X = 0$	0	$\frac{1}{3}$	0
$X = 1$	$\frac{1}{3}$	0	$\frac{1}{3}$

What is the conditional probability when Y=-1?

Are X and Y independent?

Can be read as - probability of X = x and Y = y is X conditioned on Y=y times the probability that Y=y.

Joint Distributions & Conditional Probability

Suppose the discrete random variables X and Y have supports S_x and S_y. Then we require P to satisfy

$$\sum_{x \in S_x, y \in S_y} P(X = x, Y = y) = 1$$

$$P(X|Y = y) = \frac{P(X \text{ and } Y = y)}{P(y)}$$

	$Y = -1$	$Y = 0$	$Y = 1$
$X = 0$	0	$\frac{1}{3}$	0
$X = 1$	$\frac{1}{3}$	0	$\frac{1}{3}$

$$P(X = 0|Y = -1) = \frac{0}{0 + \frac{1}{3}} = 0$$

$$P(X = 1|Y = -1) = \frac{1/3}{0 + \frac{1}{3}} = 1$$

Can be read as - probability of X = x and Y = y is X conditioned on Y=y times the probability that Y=y.

Conditional Expectation

	$Y=-1$	$Y=0$	$Y=1$
$X=0$	0	$\frac{1}{3}$	0
$X=1$	$\frac{1}{3}$	0	$\frac{1}{3}$

$$\mathbb{E}(X|Y=y)=\sum_x xP(X=x|Y=y)$$

$$P(X=0|Y=-1)=\frac{0}{0+\frac{1}{3}}=0$$

$$P(X=1|Y=-1)=\frac{1/3}{0+\frac{1}{3}}=1$$

$$\mathbb{E}(X|Y=-1)=0\cdot 0+1\cdot 1=1$$

Why is this intuitive?

Conditional Variance

	$Y=-1$	$Y=0$	$Y=1$
$X=0$	0	$\frac{1}{3}$	0
$X=1$	$\frac{1}{3}$	0	$\frac{1}{3}$

$$\mathrm{Var}(X|Y=y) = \sum_x (x - \mathbb{E}(X|Y=y))^2 p(X=x|Y=y)$$

$$P(X=0|Y=-1) = \frac{0}{0+\frac{1}{3}} = 0$$

$$P(X=1|Y=-1) = \frac{1/3}{0+\frac{1}{3}} = 1$$

$$\mathbb{E}(X|Y=-1) = 0 \cdot 0 + 1 \cdot 1 = 1$$

Why is the result below obvious as well?

$$\mathrm{Var}(X|Y=-1) = (0-1)^2 \cdot 0 + (1-1)^2 \cdot 1 = 0$$

Properties of Expectations

Linearity of Expectation

$$\mathbb{E}(cX + bY) = c\mathbb{E}(X) + b\mathbb{E}(Y)$$

Law of Total Expectation

$$\mathbb{E}(X) = \sum_i \mathbb{E}(X|A_i)P(A_i)$$

Knowing these rules well also covers most **job interview questions about probability.**

Matching Algorithm

Problem: Can we match users who use multiple devices who visit The New Yorker?

Let’s assume that we have
 If we build a model, can we do better than random? (common comparison)

Matching Algorithm

What is the expected number of correct guesses if we guessed randomly?

```
xids = pd.read_csv('xdevrecon.csv')
```

```
In [2]: xids
```

Out[2]:

	xid	match	score	sde	sos	sbr	lat	long	lci	lco
0	33b4073a-369e-4752-9311-cc17d4248911	9366ff38-4aad-4faf-85aa-53999128613a	1	desktop	Mac OS X	Safari	41.4184	-72.9073	Hamden	United States
1	fe3ea415-a5fa-4ec2-a576-fd741da9ec50	f78c194b-67a2-4f6a-8553-caed9f529bef	1	mobile	iOS	Mobile Safari	38.8822	-77.1411	Arlington	United States
2	fe3ea415-a5fa-4ec2-a576-fd741da9ec50	715bb439-d4a5-4b3b-ba5a-83d49577e04a	1	mobile	iOS	Mobile Safari	38.8822	-77.1411	Arlington	United States
3	d30c366d-0665-42f9-915c-06569b6e5789	988eb87c-a07e-4dd3-988d-f3f843acd986	1	desktop	Windows 7	Chrome	25.7651	-80.3603	Miami	United States
4	d30c366d-0665-42f9-915c-06569b6e5789	366095ad-95df-44a3-a405-9f77cfc38f38	1	desktop	Windows 7	Chrome	25.7651	-80.3603	Miami	United States
5	d30c366d-0665-42f9-915c-06569b6e5789	0e568f75-4e72-4a27-8211-9ca0e84b2e4c	1	desktop	Windows 7	Chrome	25.7651	-80.3603	Miami	United States

Matching Algorithm

Often problems with expectation require a clever definition of a random variable:

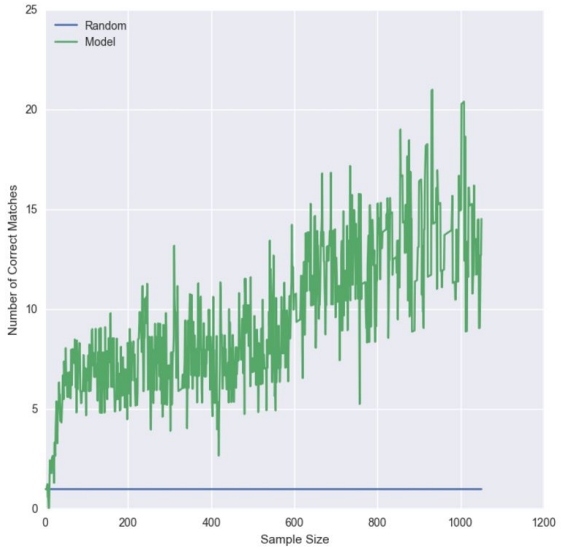
$$X_{ij} = \begin{cases} 1 & \text{if } i \text{ matches } j \\ 0 & \text{otherwise} \end{cases}$$

Arrange users j randomly and let's compute the expected number of matches.

Matching Algorithm

Often problems with expectation require a clever definition of a random variable:

$$X_{ij} = \begin{cases} 1 & \text{if } i \text{ matches } j \\ 0 & \text{otherwise} \end{cases}$$



Arrange users j randomly and let's compute the expected number of matches.

$$\mathbb{E}(\sum_i X_{ij}) = \sum_i \mathbb{E}(X_{ij}) = \sum_i \frac{1}{n} = n \cdot n = 1.$$

Used linearity of expectation - prove this.

Example for law of total expectation

What's the expected number of tosses of an unfair coin until you get a heads?

Hint: Condition on the first step

Common interview question!

Example for law of total expectation

What's the expected number of tosses of an unfair coin until you get a heads?

Let X denote the number of flips until we see a heads. We could compute:

$$\mathbb{E}(X) = \sum_{n=1}^{\infty} p(1-p)^{n-1} n$$

Example for law of total expectation

$$\mathbb{E}(X) = p\mathbb{E}(X|\text{ heads on first flip}) + (1-p)\mathbb{E}(X|\text{tails on first flip})$$

$$= p \cdot 1 + (1-p)(\mathbb{E}(X) + 1)$$

$$\mathbb{E}(X) = \frac{1}{p}$$

This is an extremely common problem on job interviews! And you need to know the trick.

Bayes Theorem

$$P(A \mid B) = \frac{P(B \mid A)\, P(A)}{P(B)},$$

Bayes' theorem is stated mathematically as the following equation:[2]

where *A* and *B* are events and P(B) ≠ 0.

- *P*(*A*) and *P*(*B*) are the probabilities of observing *A* and *B* without regard to each other.
- *P*(*A* | *B*), a conditional probability, is the probability of observing event *A* given that *B* is true.
- P(A) is known as the *prior - our initial assumption on how the data is distributed.*

P(*B* | *A*) is the probability of observing event *B* given that *A* is true.

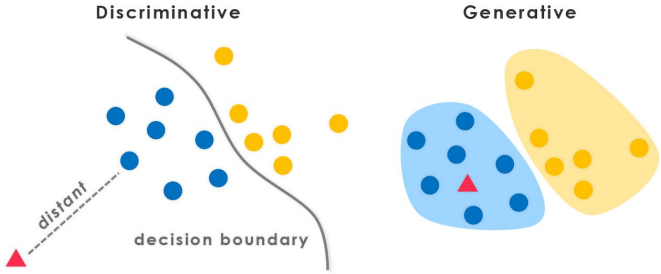
Bayesian interpretation

For proposition *A* and evidence *B*,

- *P* (*A*), the *prior*, is the initial degree of belief in *A*.
- *P* (*A* | *B*), the “posterior,” is the degree of belief having accounted for *B*.
- the quotient *P*(*B* |*A*)/*P*(*B*) represents the ***support*** *B* provides for *A*.

Bayes Theorem - Why do we use it?

$$P(A \mid B) = \frac{P(B \mid A)\, P(A)}{P(B)},$$



- We usually want to know P(Y|X) in discriminative models. This tells us the probability of an outcome Y given information about the population X.
- But what if we want to know more about the distribution of X given outcomes?
- For instance, can we identify if someone is male or female based on weight, height and shoe size?

Who stole the cookie?

Suppose there are two full bowls of cookies. Bowl #1 has 10 chocolate chip and 30 plain cookies, while bowl #2 has 20 of each. Our friend Fred picks a bowl at random, and then picks a cookie at random. We may assume there is no reason to believe Fred treats one bowl differently from another, likewise for the cookies. The cookie turns out to be a plain one. How probable is it that Fred picked it out of bowl #1?

Who stole the cookie?

Suppose there are two full bowls of cookies. Bowl #1 has 10 chocolate chip and 30 plain cookies, while bowl #2 has 20 of each. Our friend Fred picks a bowl at random, and then picks a cookie at random. We may assume there is no reason to believe Fred treats one bowl differently from another, likewise for the cookies. The cookie turns out to be a plain one. How probable is it that Fred picked it out of bowl #1?

$$P(E \mid H_1) = 30/40 = 0.75$$

$$P(E \mid H_2) = 20/40 = 0.5.$$

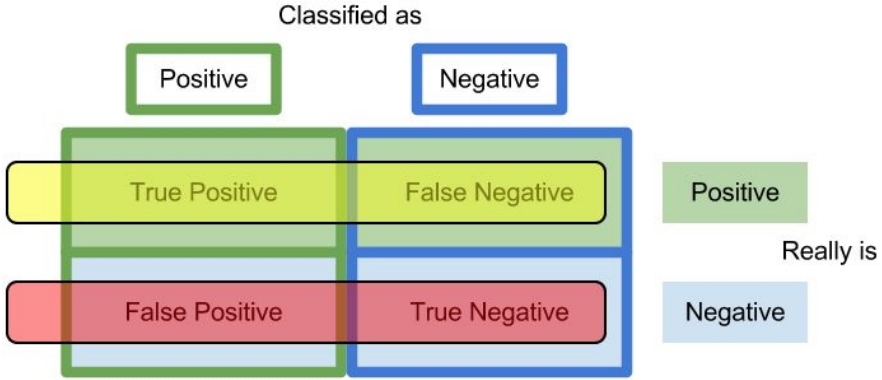
$$P(H_1 \mid E) = \frac{P(E \mid H_1)\,P(H_1)}{P(E \mid H_1)\,P(H_1) \;+\; P(E \mid H_2)\,P(H_2)}$$

$$= \frac{0.75 \times 0.5}{0.75 \times 0.5 + 0.5 \times 0.5}$$

$$= 0.6$$

Drug Test

Suppose a drug test is 99% sensitive and 99% specific. That is, the test will produce 99% true positive results for drug users and 99% true negative results for non-drug users. Suppose that 0.5% of people are users of the drug. If a randomly selected individual tests positive, what is the probability that he is a user?

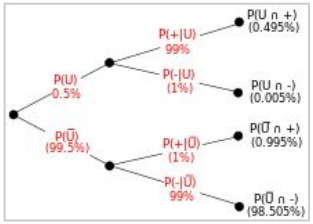


Computing the conditional probability

$$P(\text{User} \mid +) = \frac{P(+ \mid \text{User})P(\text{User})}{P(+ \mid \text{User})P(\text{User}) + P(+ \mid \text{Non-user})P(\text{Non-user})}$$

$$= \frac{0.99 \times 0.005}{0.99 \times 0.005 + 0.01 \times 0.995}$$

$$\approx 33.2\%$$



Predicting bugs

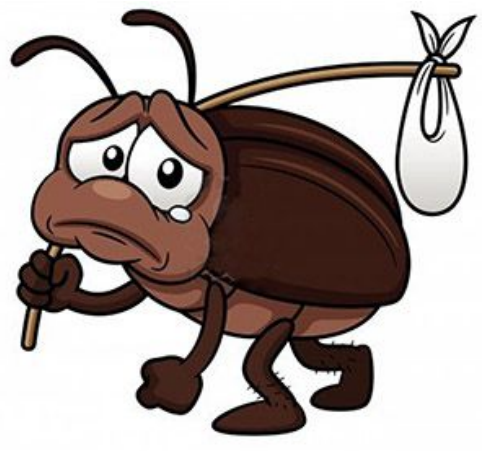
An entomologist spots what might be a rare subspecies of beetle, due to the pattern on its back. In the rare subspecies, 98% have the pattern, or $P(\text{Pattern} \mid \text{Rare}) = 98\%$. In the common subspecies, 5% have the pattern. The rare subspecies accounts for only 0.1% of the population. How likely is the beetle having the pattern to be rare, or what is $P(\text{Rare} \mid \text{Pattern})$?

Computing the conditional probability

$$P(\text{Rare} \mid \text{Pattern}) = \frac{P(\text{Pattern} \mid \text{Rare})P(\text{Rare})}{P(\text{Pattern} \mid \text{Rare})P(\text{Rare}) + P(\text{Pattern} \mid \text{Common})P(\text{Common})}$$

$$= \frac{0.98 \times 0.001}{0.98 \times 0.001 + 0.05 \times 0.999}$$

$$\approx 1.9\%$$



Exercise: Sad Lamps

Factory	% of total production	Probability of defective lamps
A	$0.35 = P(A)$	$0.015 = P(D \mid A)$
B	$0.35 = P(B)$	$0.010 = P(D \mid B)$
C	$0.30 = P(C)$	$0.020 = P(D \mid C)$

Question: Given that a lamp is defective, what is the probability that it was of type A?

Bayesian vs Frequentist

- Draw your conclusions from sample data by emphasizing the frequency or proportion of the data.

- Believe probability expresses a degree of belief in an event, which can change as new information is gathered, rather than a fixed value based upon frequency or propensity.

Finding the bias of a coin

Problem: If I flip a coin N times and I get k heads, what's the the bias of the coin?

$$p(Y = k) = \begin{cases} p & \text{if } k \text{ is } 1 \\ 1 - p & \text{if } k \text{ is } 0 \end{cases}$$

The main problem machine learning tries to solve is inference. We know our method is a good one if **we can recover parameters when we sample data from known distributions.**

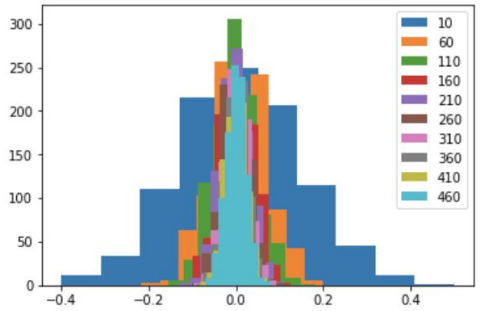
Inferring coin bias as a frequentist

$$\hat{p} := \frac{1}{N}\sum_{i=1}^{N} X_i$$

```
In [284]: n = 100
          p = 0.5

          for n in range(10,500,50):
              sum_xi = np.random.binomial(n, p, 1000)
              sum_xi_over_n = np.array([float(i)/n for i in sum_xi])
              plt.hist((sum_xi_over_n - p),label=str(n))
          plt.legend()

Out[284]: <matplotlib.legend.Legend at 0x1a304e4590>
```



Flip a fair coin 100 times. What are the possible outcomes and how are they distributed?

$$\mathbb{E}(\hat{p}) = p$$

Mean

$$\mathbb{E}(\hat{p} - \mathbb{E}(\hat{p}))^2 = \frac{p(1-p)}{n}$$

Variance

Homework 0

Thus our estimator has mean p and variance which shrinks to zero. This is known as the **Law of Large Numbers.**

Inferring coin bias as a frequentist

$$\hat{p} := \frac{1}{N}\sum_{i=1}^{N} X_i$$

$$\mathbb{E}(\hat{p}) = p$$

$$\mathbb{E}(\hat{p} - \mathbb{E}(\hat{p}))^2 = \frac{p(1-p)}{n}$$

$$Z_n := \frac{\hat{p} - p}{\sqrt{p(1-p)/n}}$$

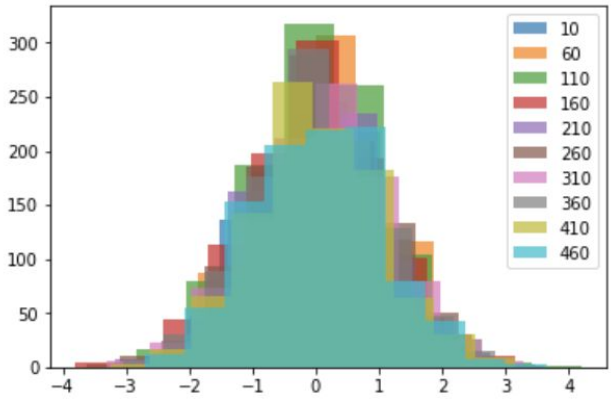
This distribution has mean zero and unit variance and can allow us to infer bounds on the range of p.

```
In [293]: n = 100
          p = 0.5

          for n in range(10,500,50):
              s = np.random.binomial(n, p, 1000)
              plt.hist((s - n*p)/(np.sqrt(n)*np.sqrt(p*(1-p))),alpha=0.7,label=str(n))

          plt.legend()

Out[293]: <matplotlib.legend.Legend at 0x1a31537250>
```



Inferring coin bias as a frequentist

$$\hat{p} := \frac{1}{N}\sum_{i=1}^{N} X_i$$

$$\mathbb{E}(\hat{p}) = p$$

$$\mathbb{E}(\hat{p} - \mathbb{E}(\hat{p}))^2 = \frac{p(1-p)}{n}$$

$$Z_n := \frac{\hat{p} - p}{\sqrt{\hat{p}(1-\hat{p})/n}}$$

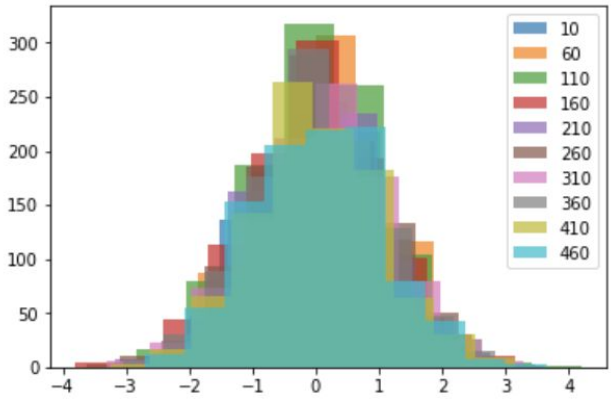
But we don't know p! So we have to replace the variance with our estimate of it.

```
In [293]: n = 100
          p = 0.5

          for n in range(10,500,50):
              s = np.random.binomial(n, p, 1000)
              plt.hist((s - n*p)/(np.sqrt(n)*np.sqrt(p*(1-p))),alpha=0.7,label=str(n))

          plt.legend()

Out[293]: <matplotlib.legend.Legend at 0x1a31537250>
```



Central Limit Theorem

$$Z_n := \frac{\hat{p} - p}{\sqrt{\hat{p}(1-\hat{p})/n}}$$

$$Z_n \sim \mathcal{N}(0,1)$$

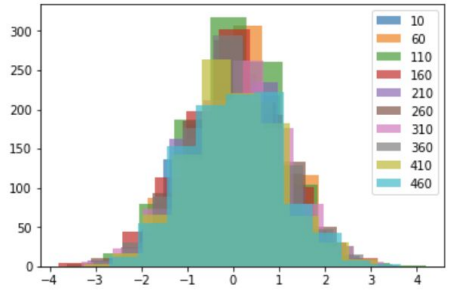
For large n, if the samples are independent and drawn from the same distribution, then our estimator has an approximately normal distribution.

```
In [293]: n = 100
          p = 0.5

          for n in range(10,500,50):
              s = np.random.binomial(n, p, 1000)
              plt.hist((s - n*p)/(np.sqrt(n)*np.sqrt(p*(1-p))),alpha=0.7,label=str(n))

          plt.legend()

Out[293]: <matplotlib.legend.Legend at 0x1a31537250>
```



Therefore the true p lies in the interval

$$\hat{p} \pm 1.96\sqrt{\frac{\hat{p}\,(1-\hat{p})}{n}}$$

Central Limit Theorem

Let $\{X_1, \ldots, X_n\}$ be a random sample of size n—that is, a sequence of independent and identically distributed (i.i.d.) random variables drawn from a distribution of expected value given by μ and finite variance given by σ^2. Suppose we are interested in the sample average

$$S_n := \frac{X_1 + \cdots + X_n}{n}$$

Lindeberg–Lévy CLT. Suppose $\{X_1, X_2, \ldots\}$ is a sequence of i.i.d. random variables with $\mathrm{E}[X_i] = \mu$ and $\mathrm{Var}[X_i] = \sigma^2 < \infty$. Then as n approaches infinity, the random variables $\sqrt{n}(S_n - \mu)$ converge in distribution to a normal $N(0, \sigma^2)$:[3]

$$\sqrt{n}\left(S_n - \mu\right) \xrightarrow{d} N\left(0, \sigma^2\right). \longrightarrow \Phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

Example

An electrical component is guaranteed by its suppliers to have 2% defective components. To check a shipment, you test a random sample of 500 components. If the supplier’s claim is true, what is the probability of finding 15 or more defective components?

Example

An electrical component is guaranteed by its suppliers to have 2% defective components. To check a shipment, you test a random sample of 500 components. If the supplier's claim is true, what is the probability of finding 15 or more defective components?

$$f(Y \geq 15) = \sum_{k=15}^{500} \binom{n}{k} 0.02^k (1 - 0.02)^{500-k}$$

Binomial distribution

Solution

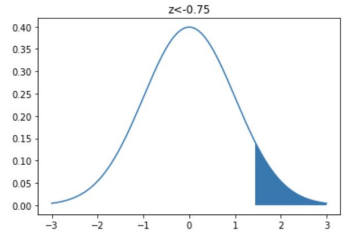
$$f(Y \geq 15) = \sum_{k=15}^{500} \binom{n}{k} 0.02^k (1-0.02)^{500-k}$$

$$z := \frac{Y - np}{\sqrt{p(1-p)n}} = \frac{15 - 500(0.02)}{\sqrt{0.02(1-0.02)500}} = 1.44$$

$$f(Y \geq 15) \sim 1 - \Phi(1.44) = 1 - 0.9251 = 0.0749$$

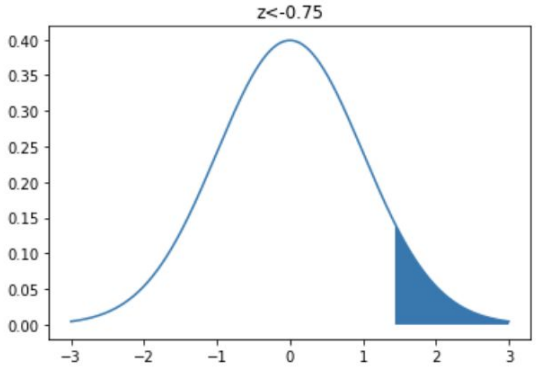
$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-z^2/2} dx$$

```
In [439]: import numpy as np
          import matplotlib.pyplot as plt
          from scipy.stats import norm
          def draw_z_score(x, cond, mu, sigma, title):
              y = norm.pdf(x, mu, sigma)
              z = x[cond]
              plt.plot(x, y)
              plt.fill_between(z, 0, norm.pdf(z, mu, sigma))
              plt.title(title)
              plt.show()
          x =np.arange(-3,3,0.001)
          z0 = 1.44
          draw_z_score(x, x>z0, 0, 1, 'z<-0.75')
```



Solution

```
In [439]: import numpy as np
          import matplotlib.pyplot as plt
          from scipy.stats import norm
          def draw_z_score(x, cond, mu, sigma, title):
              y = norm.pdf(x, mu, sigma)
              z = x[cond]
              plt.plot(x, y)
              plt.fill_between(z, 0, norm.pdf(z, mu, sigma))
              plt.title(title)
              plt.show()
          x =np.arange(-3,3,0.001)
          z0 = 1.44
          draw_z_score(x, x>z0, 0, 1, 'z<-0.75')
```



$$f(Y \geq 15) \sim 1 - \Phi(1.44) = 1 - 0.9251 = 0.0749$$

“Proof” of central limit theorem

$$Z_t = \sum_{i=1}^{N} X_i$$

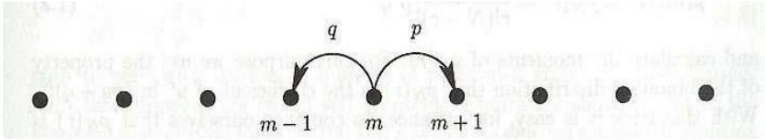
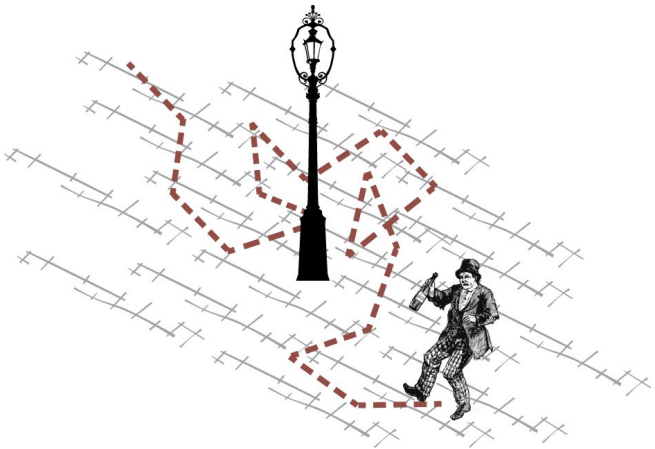


Figure 2.1: *A random walker on a 1-dimensional lattice of sites that are a fixed distance Δx apart. The walker jumps to the right with probability p and to the left with probability $q = 1 - p$.*

$$P(X_i = +\Delta x) = \frac{1}{2} = P(X_i = -\Delta x)$$

Let’s reinterpret these sums now as distance from the origin. If we flip the coin and get heads, we go right. If we get tails, we go left!

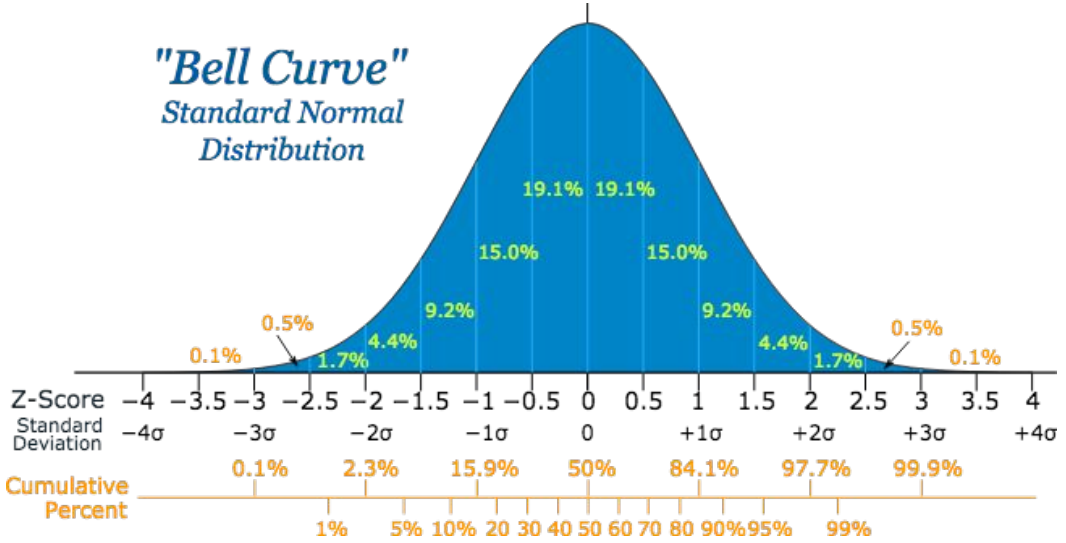


“Proof” of central limit theorem

$$Z_t = \sum_{i=1}^{N} X_i$$

$$P(X_i = +\Delta x) = \frac{1}{2} = P(X_i = -\Delta x)$$

Now Z tells us the distance from the origin we’ve travelled after N steps, of size Δx



We are less likely to end up in the tails than we are back at the origin. Can we find an expression for this process?

“Proof” of central limit theorem

$$Z_t = \sum_{i=1}^{N} X_i \qquad P(X_i = +\Delta x) = \frac{1}{2} = P(X_i = -\Delta x)$$

$$p(x,t) = p(x-\Delta x, t-\Delta t)\frac{1}{2} + p(x+\Delta x, t-\Delta t)\frac{1}{2}$$

$$p(x,t) - p(x,t-\Delta t) =$$

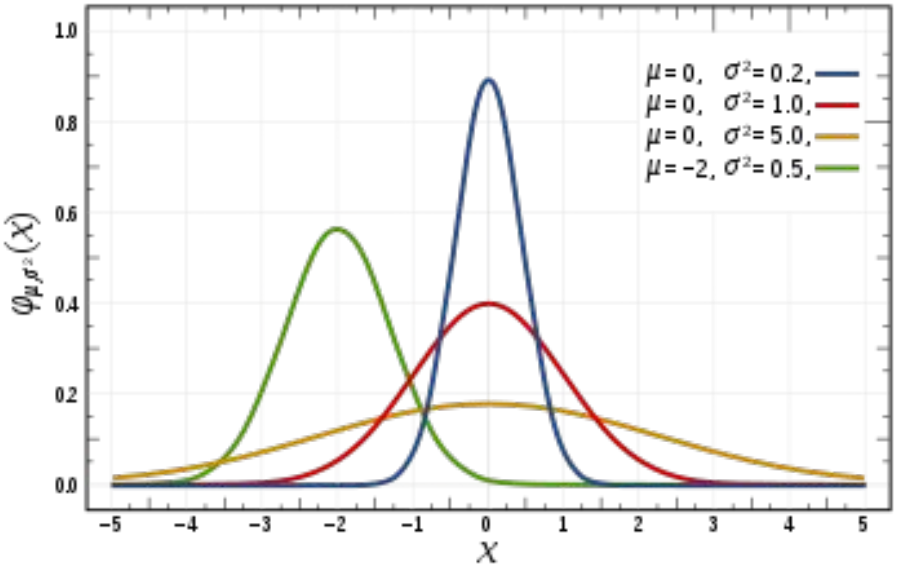
$$(p(x-\Delta x, t-\Delta t) - p(x,t-\Delta t))\frac{1}{2} + (p(x+\Delta x, t-\Delta t) - p(x,t-\Delta t))\frac{1}{2}$$

The heat equation!!

$$\Rightarrow \frac{\partial p(x,t)}{\partial t}\Delta t \sim \frac{\partial^2 p(x,t)}{\partial x^2}\Delta x^2 \qquad \Rightarrow \frac{\partial p(x,t)}{\partial t} = \alpha \frac{\partial^2 p(x,t)}{\partial x^2}$$

Solution is a normal distribution!

$$p(x,t) = \frac{1}{\sqrt{4\pi\alpha t}} \exp\left(-\frac{x^2}{4\alpha t}\right).$$



Replace t with n up to scaling and we have convergence to the normal distribution. This is the central limit theorem.

Inferring coin bias as a Bayesian

For us we have:

- B is an integer space representing **k, the number of successes, N trials.**
- A is our distribution on p, the bias of the coin.
- P(A) is our initial prior belief. In our case we will assume P(A) = f(p) = 1 (uniform).

$$P(A \mid B) = \frac{P(B \mid A)\, P(A)}{P(B)},$$

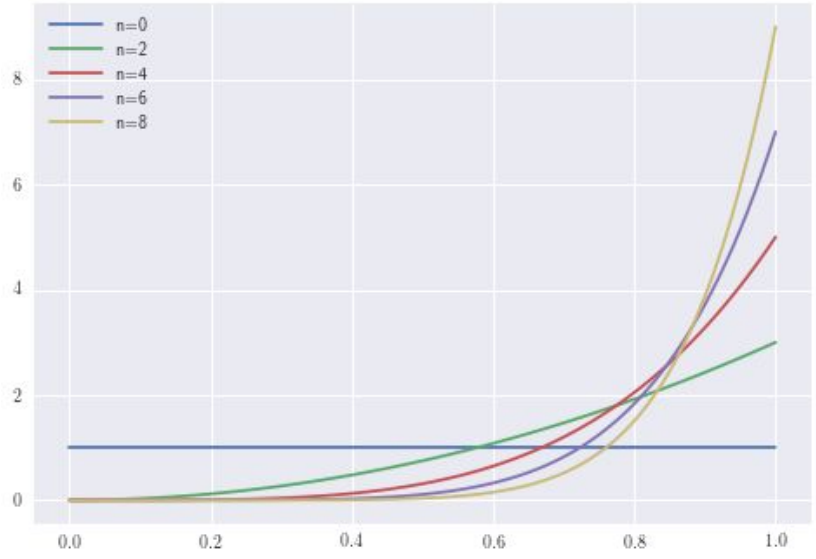
$$f(k|p,n) = \binom{n}{k} p^n (1-p)^{n-k}$$

Homework 0 will cover this topic.

$$f(p|k,n) = \frac{f(k|p,n)f(p)}{\int_0^1 f(k|p,n)f(p)dp}$$

Notice that we now have a distribution over possible coin biases.

Inferring coin bias as a Bayesian



$$P(A \mid B) = \frac{P(B \mid A)\, P(A)}{P(B)},$$

$$f(k|p,n) = \binom{n}{k} p^n (1-p)^{n-k}$$

$$f(p|k,n) = \frac{f(k|p,n) f(p)}{\int_0^1 f(k|p,n) f(p) dp}$$

An important distinction

- In the **Frequentist Method**, you assume a ground truth exists that you are approximating with your estimations. You use the central limit theorem to infer confidence in your estimates from the data, knowing their distribution is approximately normal. There are no a priori assumptions on the distribution you are sampling data from though!

- In the **Bayesian Method**, you instead start with a prior distribution for the parameter you wish to estimate. Thus rather than inferring confidence bounds on your samples from CLT, you update your distribution whenever new data comes in. All estimates of quantities are based on that distribution alone.

Probability Review

Please review Discrete Probability Distributions if you feel rusty.

https://github.com/Columbia-Intro-Data-Science/APMAE4990-/blob/master/pdfs/Discrete-Probabilities.pdf

Please review Conditional Probability and Random Variables if you feel rusty.

https://github.com/Columbia-Intro-Data-Science/APMAE4990-/blob/master/pdfs/Conditional-Probability.pdf

Excellent source for job interview questions **(Chapter 3)** and mastering probability basics.

https://github.com/Columbia-Intro-Data-Science/APMAE4990-/blob/master/pdfs/%5BMark%20Joshi%5DQuant%20Job%20Interview%20Questions%20And%20Answers%20(1).pdf