22.01.2019

Statistical Methods in AI (CSE/ECE 471)

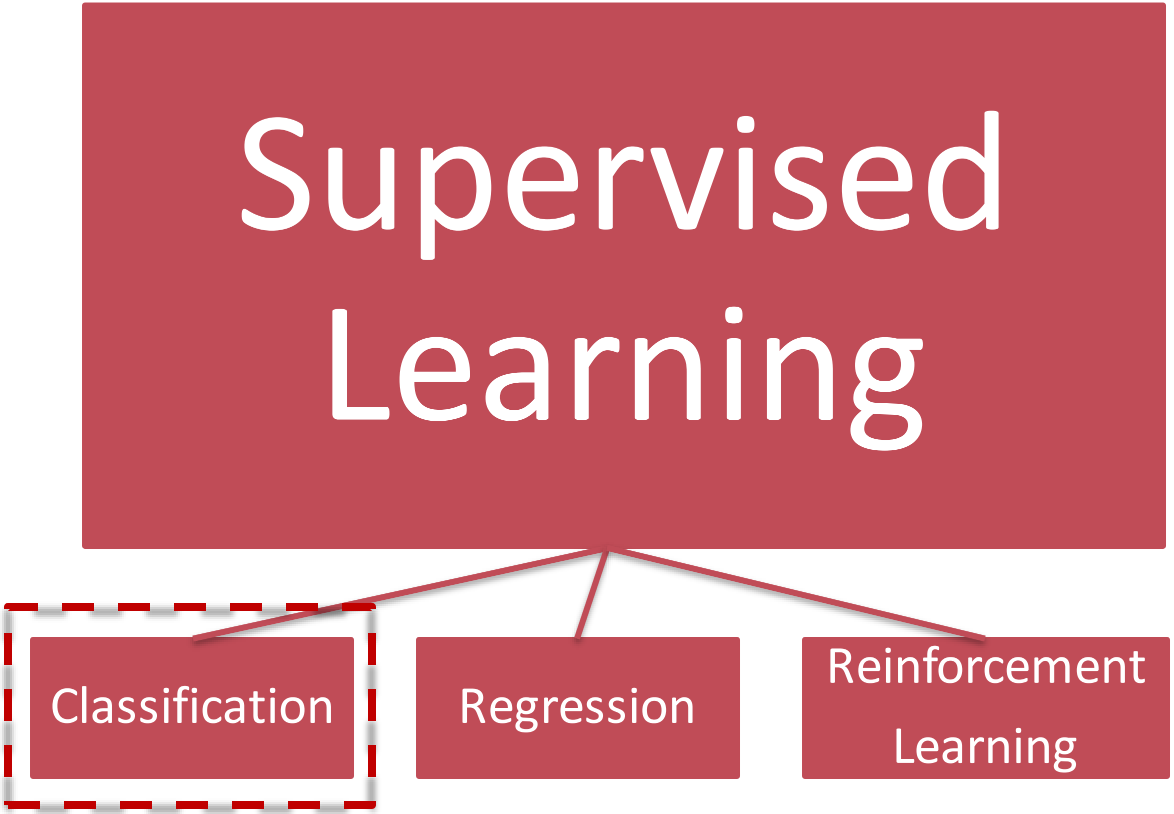
Lecture-6: Naïve Bayes classifier, Linear Regression

Ravi Kiran

Center for Visual Information Technology (CVIT), IIIT Hyderabad

Announcements

- A1 due today, 11.59 pm
- A2 will be posted. Due Feb 2, 11.59 pm


Supervised Learning
Classification
Regression
Reinforcement Learning

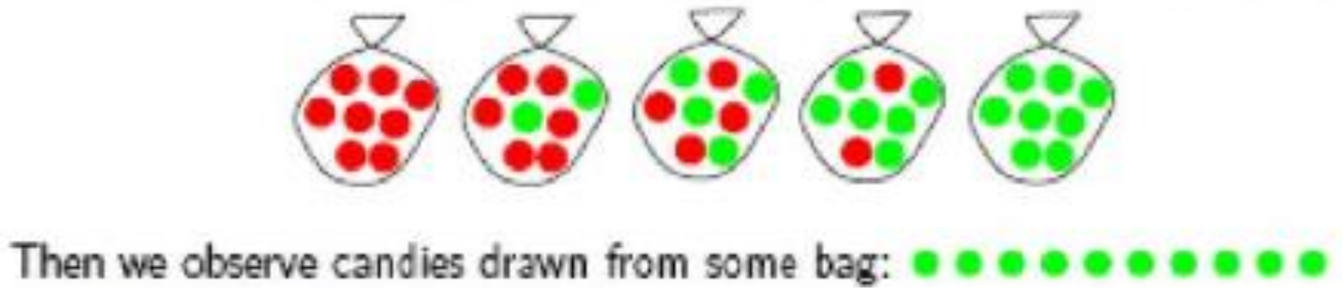
So far

- Decision Tree classifier
- K-NN classifier


PROBABILITY = EVENT / OUTCOMES
wikiHow to Calculate Probability

Data – a probability-based perspective

- The basis for Statistical Learning Theory



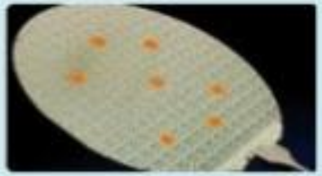
  - Domain described by random variables (r.v.)
    - X = {apple, grape}
    - $b_i \in [1,5]$
  - **Data = Instantiation of some or all r.v.'s in the domain**

Uncertainty arises from many sources

Process Uncertainty

Processes contain "randomness"

Uncertain travel times



Semiconductor yield

Data Uncertainty

Data input is uncertain

Intended Spelling | Text Entry | Actual Spelling

GPS Uncertainty

Testimony

{Paris Airport}

Ambiguity

Contaminated?

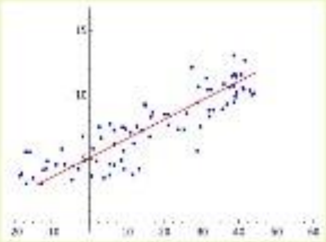
Rumors

{John Smith, Dallas}
{John Smith, Kansas}

Conflicting Data

Model Uncertainty

All modeling is approximate



Fitting a curve to data

Forecasting a hurricane (www.noaa.gov)

Data: a probabilistic perspective

	DBAName	AKAName	Address	City	State	Zip
t1	John Veliotis Sr.	Johnnyo's	3465 S Morgan ST	***Chicago***	IL	***60608***
t2	John Veliotis Sr.	Johnnyo's	3465 S Morgan ST	Chicago	IL	***60609***
t3	John Veliotis Sr.	Johnnyo's	3465 S Morgan ST	Chicago	IL	***60609***
t4	***Johnnyo's***	Johnnyo's	3465 S Morgan ST	***Cicago***	IL	60608

Conflicts

Does not obey data distribution

Conflict

Output

Proposed Cleaned Dataset

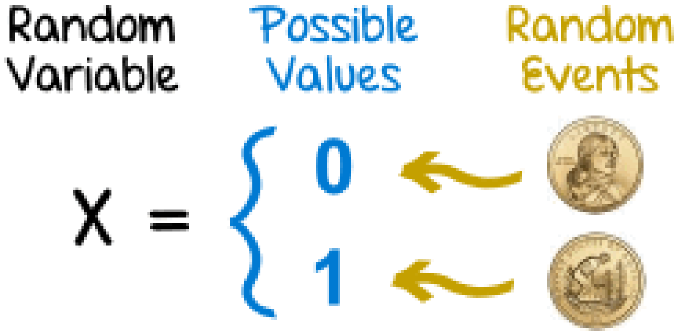
	DBAName	Address	City	State	Zip
t1	John Veliotis Sr.	3465 S Morgan ST	Chicago	IL	60608
t2	John Veliotis Sr.	3465 S Morgan ST	Chicago	IL	**60608**
t3	John Veliotis Sr.	3465 S Morgan ST	Chicago	IL	**60608**
t4	**John Veliotis Sr.**	3465 S Morgan ST	**Chicago**	IL	60608

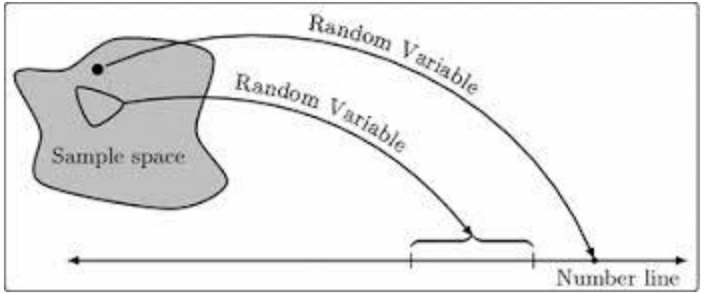
Marginal Distribution of Cell Assignments

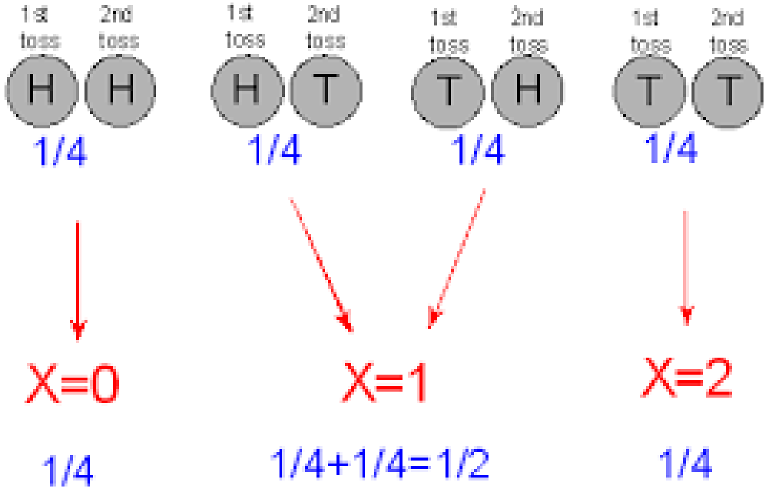
Cell	Possible Values	Probability
t2.Zip	60608	0.84
	60609	0.16
t4.City	Chicago	0.95
	Cicago	0.05
t4.DBAName	John Veliotis Sr.	0.99
	Johnnyo's	0.01

Random Variables

R.V. = A numerical value from a random experiment







Random variables

- A **discrete random variable** can assume a countable number of values.
  - Number of steps to the top of the Eiffel Tower*

Random variables

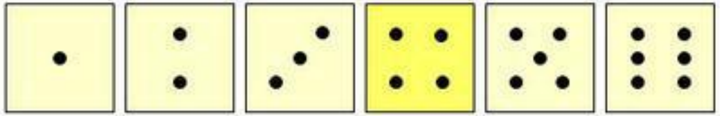
- A **discrete random variable** can assume a countable number of values.
  - Number of steps to the top of the Eiffel Tower*
- A **continuous random variable** can assume any value along a given interval of a number line.
  - The time a tourist stays at the top once s/he gets there

*Believe it or not, the answer ranges from 1,652 to 1,789. See Great Buildings

Discrete Random Variables

- Can only take on a countable number of values

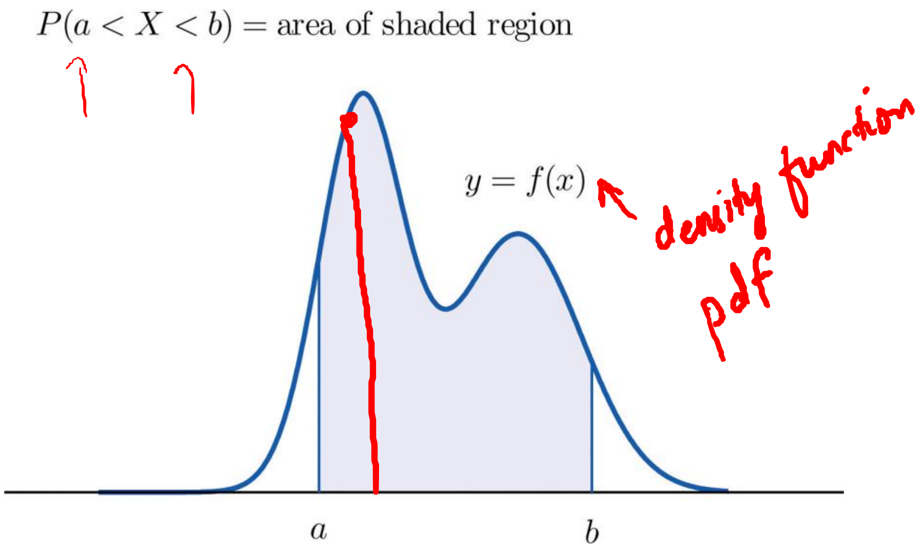
Examples:



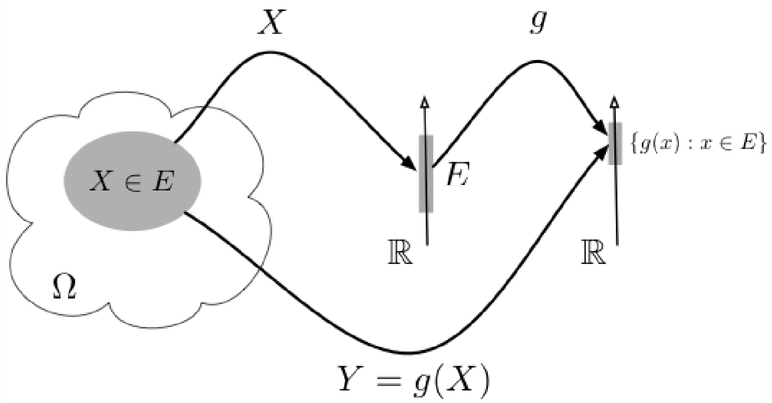
- **Roll a die twice**
  Let X be the number of times 4 comes up
  (then X could be 0, 1, or 2 times)

- **Toss a coin 5 times.**
  Let X be the number of heads
  (then X = 0, 1, 2, 3, 4, or 5)

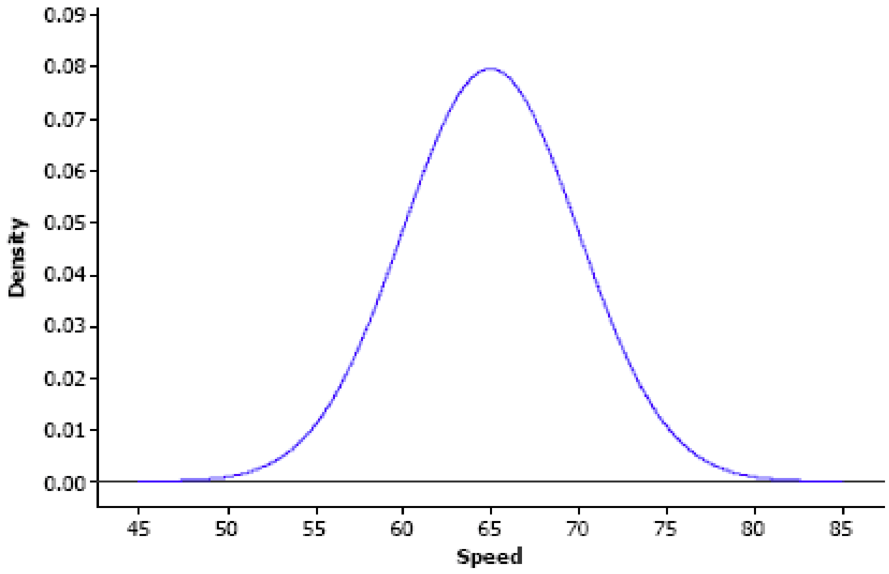
Continuous random variable

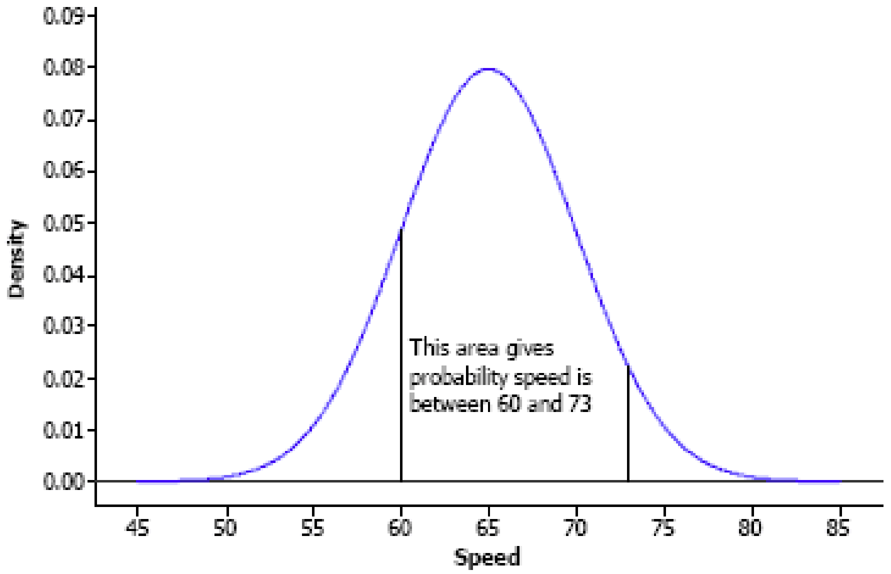


$P(Y = 0.2) = 0$

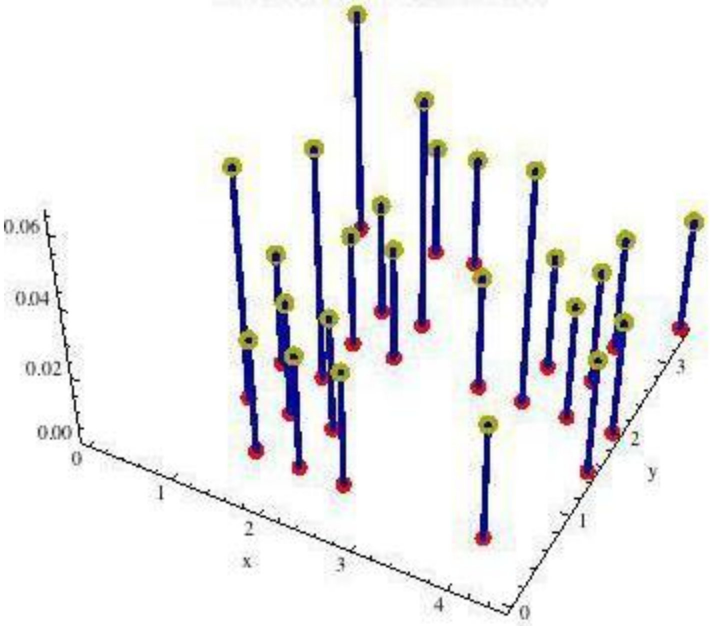


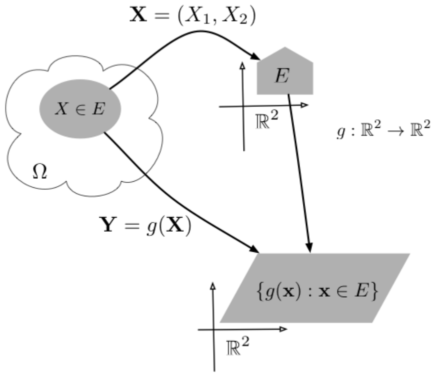
Continuous random variable

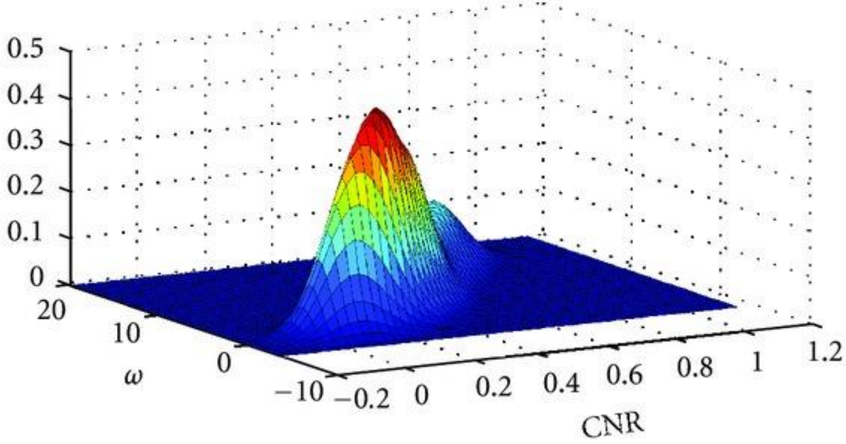




Random vectors







Data → r.v.

Relative frequency

Relative frequency is the same as **experimental probability**.
We use relative frequency to predict probabilities from experimental data.

The experiment

This spinner was spun 40 times and the results recorded in this table:

Colour	Frequency
Blue	20
Yellow	10
Red	5
Green	5

Relative frequency

$$\frac{\text{frequency of event}}{\text{total number of trials}}$$

Event means **one possible outcome**; here, one colour on the spinner.

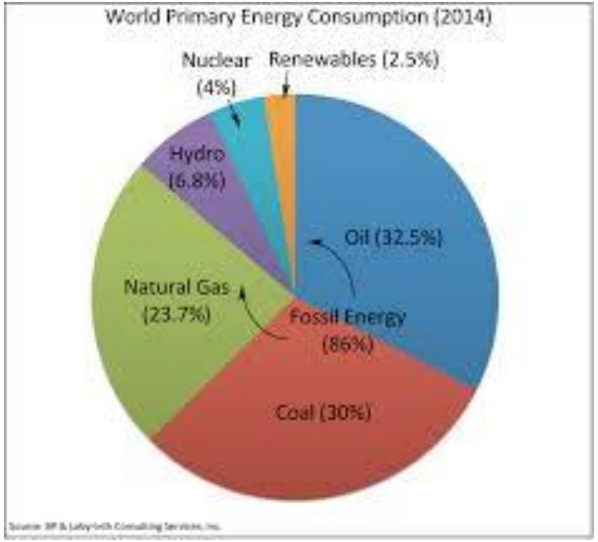
There were 20 blues recorded...

...out of 40 spins.

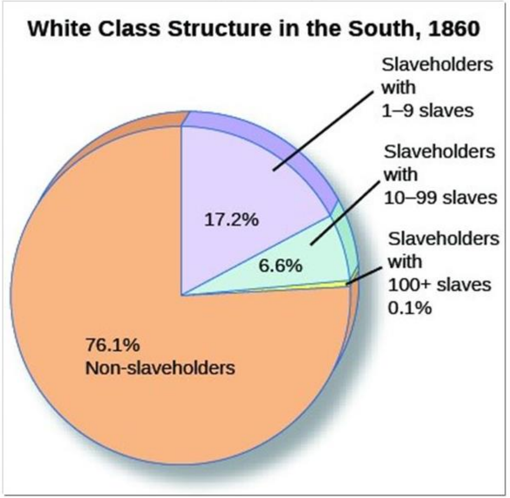
$$P(\text{blue}) = \frac{20}{40}$$

Simplify: $P(\text{blue}) = \frac{20}{40} = \frac{2}{4} = \frac{1}{2}$

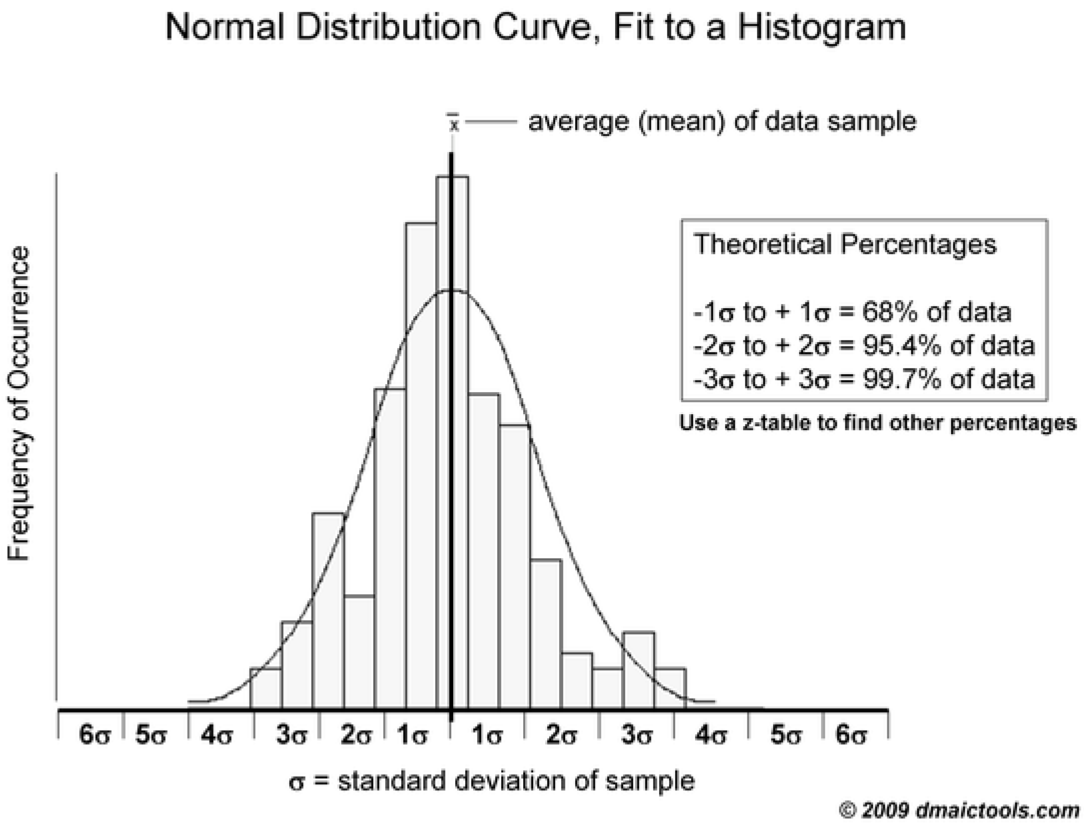
Discrete Prior distributions



Slave-Owning Population (1860)



Data → r.v.

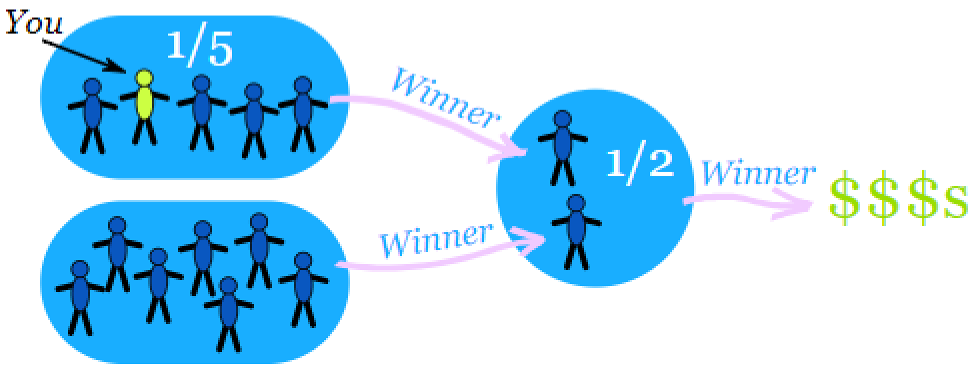


Normal Distribution Curve, Fit to a Histogram
x̄ average (mean) of data sample
Frequency of Occurrence
Theoretical Percentages
-1σ to + 1σ = 68% of data
-2σ to + 2σ = 95.4% of data
-3σ to + 3σ = 99.7% of data
Use a z-table to find other percentages
6σ 5σ 4σ 3σ 2σ 1σ 1σ 2σ 3σ 4σ 5σ 6σ
σ = standard deviation of sample
© 2009 dmaictools.com

Independent Events

Imagine there are two groups:

- A member of each group gets randomly chosen for the winners circle,
- **then** one of those gets randomly chosen to get the big money prize:



What is your chance of winnning the big prize?

Independent vs. Dependent Events

Using the bag of marbles on the left, what is the probability of pulling a black marble two times in a row? $P(black, black)$

When you put 1st marble back in
(*Independent Events*)

$$\frac{2}{10} * \frac{2}{10}$$

$$\frac{1}{5} * \frac{1}{5} \boxed{= \frac{1}{25}}$$

When you KEEP 1st marble
(*Dependent Events*)

$$\frac{2}{10} * \frac{1}{9}$$

$$\frac{1}{5} * \frac{1}{9}$$

Independent Events

The outcome of one event does not affect the outcome of the other.

If A and B are independent events then the probability of both occurring is

$$P(A \text{ and } B) = P(A) \times P(B)$$

Dependent Events

The outcome of one event affects the outcome of the other.

If A and B are dependent events then the probability of both occurring is

$$P(A \text{ and } B) = P(A) \times P(B | A)$$

Probability of B given A

Independent vs. Dependent Events

Using the bag of marbles on the left, what is the probability of pulling a black marble two times in a row? $P(black, black)$

When you put 1st marble back in (*Independent Events*)

$$\frac{2}{10} * \frac{2}{10}$$

$$\frac{1}{5} * \frac{1}{5} = \frac{1}{25}$$

When you KEEP 1st marble (*Dependent Events*)

$$\frac{2}{10} * \frac{1}{9}$$

$$\frac{1}{5} * \frac{1}{9}$$

$$P(A \text{ and } B) = P(A) \times P(B)$$

$$P(A \text{ and } B) = P(A) \times P(B|A)$$

Probability of B given A

Marginal Probabilities

$$\begin{cases} x = 1 & \text{(Rains)} \\ x = 0 & \text{(Doesn't rain)} \end{cases}$$

$$Pr(x = 1) = 0.6$$
$$Pr(x = 0) = 0.4$$

$$\begin{cases} y = 1 & \text{(Have umbrella)} \\ y = 0 & \text{(Don't have umbrella)} \end{cases}$$

$$Pr(y = 1) = 0.3$$
$$Pr(y = 0) = 0.7$$

Joint Probability

$$\begin{cases} x = 1 & \text{(Rains)} \\ x = 0 & \text{(Doesn't rain)} \end{cases}$$

$$\begin{cases} y = 1 & \text{(Have umbrella)} \\ y = 0 & \text{(Don't have umbrella)} \end{cases}$$

$Pr(x = 1) = 0.6$
$Pr(x = 0) = 0.4$
$Pr(y = 1) = 0.3$
$Pr(y = 0) = 0.7$

$$Pr(x = 0) = \sum_{y=0}^{1} Pr(x = 0, y)$$
$$= Pr(x = 0, y = 0) + Pr(x = 0, y = 1)$$
$$= 0.28 + 0.12 = 0.4$$

<u>Case 1 : Rains but you have an umbrella</u>

$$Pr(x = 1, y = 1) = Pr(x = 1) \times Pr(y = 1)$$
$$= 0.6 \times 0.3$$
$$= 0.18$$

<u>Case 2 : Rains but you DON'T have an umbrella</u>

$$Pr(x = 1, y = 0) = Pr(x = 1) \times Pr(y = 0)$$
$$= 0.6 \times 0.7$$
$$= 0.42$$

Conditional Probability

Given

$x = 1$ (Rains)

$\downarrow$

What's the Probability of

$y = 1$ (Bring umbrella)

$$\begin{cases} x = 1 & \text{(Rains)} \\ x = 0 & \text{(Doesn't rain)} \end{cases}$$

$$\begin{cases} y = 1 & \text{(Have umbrella)} \\ y = 0 & \text{(Don't have umbrella)} \end{cases}$$

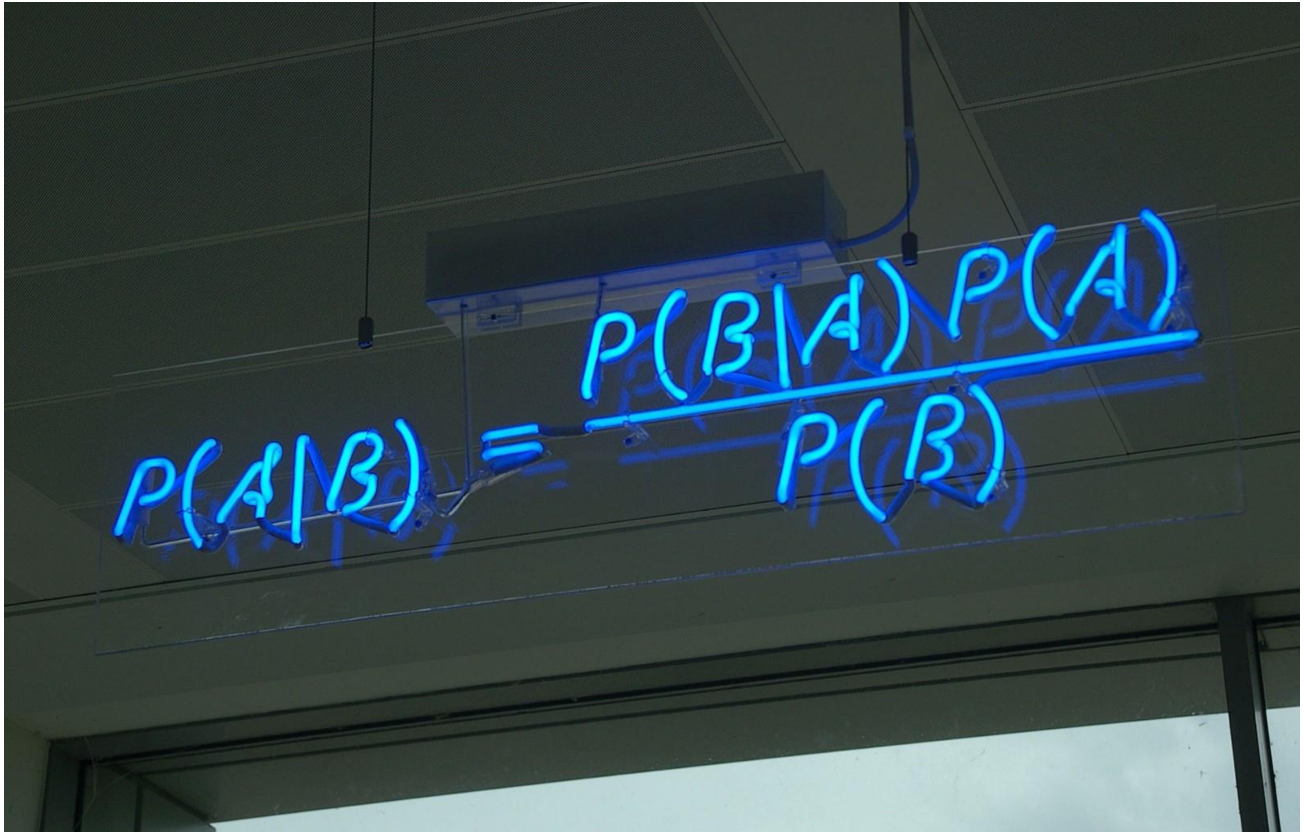
$Pr(x = 1) = 0.6$

$Pr(x = 0) = 0.4$

$Pr(y = 1) = 0.3$

$Pr(y = 0) = 0.7$

Bayes' Rule



- A disease occurs in 0.5% of population
- A diagnostic test gives a positive result
  - in 99% of people that have the disease
  - in 5% of people that do not have the disease (false positive)

A random person from the street is found to be positive on this test. What is the probability that they have the disease?

A: 0-30%

B: 30-60%

C: 60-90%

- A disease occurs in 0.5% of population
- A diagnostic test gives a positive result
  - in 99% of people that have the disease $P(B|A)$
  - in 5% of people that do not have the disease (false positive) $P(B|\sim A)$

A = disease

B = positive test result

P(A) = 0.005 probability of having disease

P(~A) = 1 – 0.005 = 0.995 probability of not having disease

P(B) = 0.005 * 0.99 (people with disease) + 0.995 * 0.05 (people without disease) = 0.0547 (slightly more than 5% of *all* tests are positive)

conditional probabilities

P(B|A) = 0.99 probability of pos result **given** you have disease

P(~B|A) = 1 – 0.99 = 0.01 probability of neg result **given** you have disease

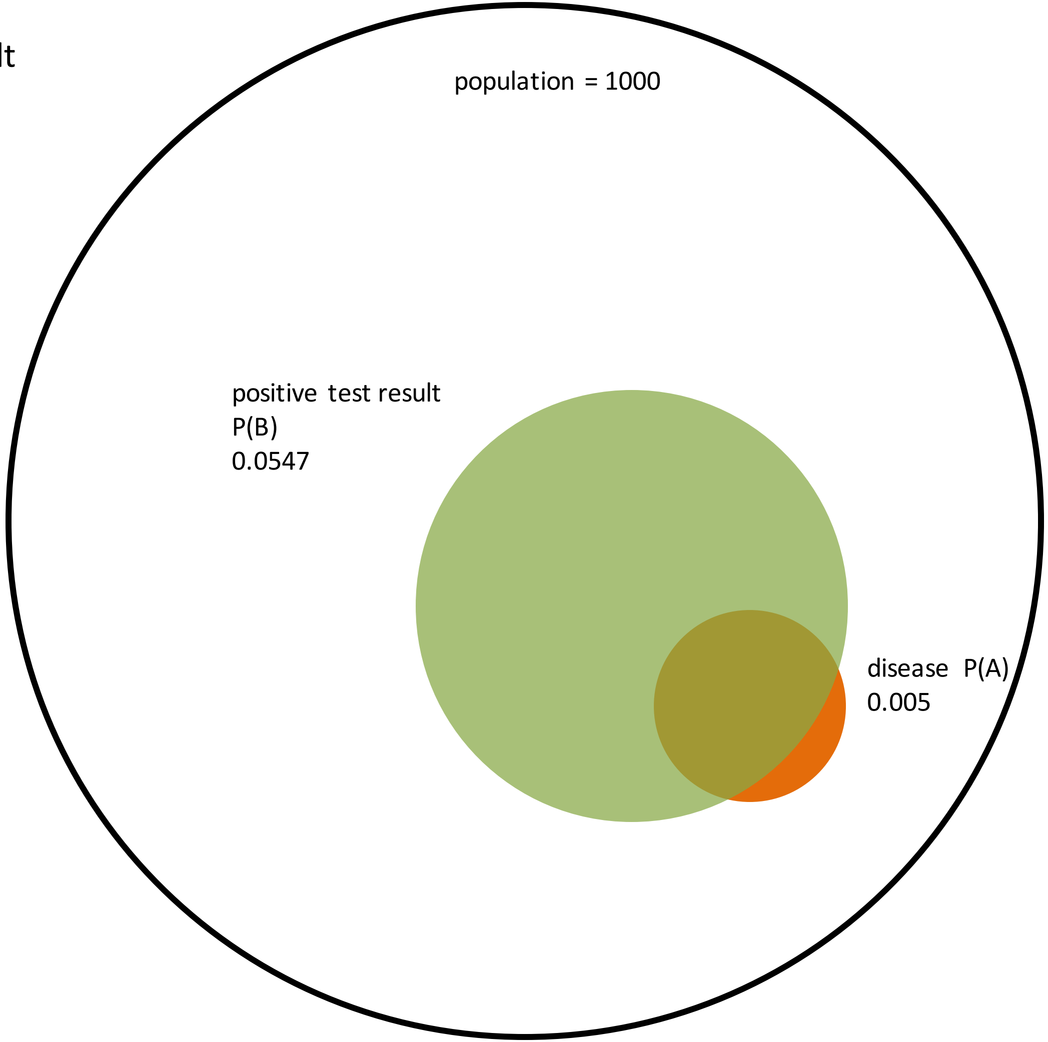
P(B|~A) = 0.05 probability of pos result **given** you do not have disease

P(~B|~A) = 1 – 0.05 = 0.95 probability of neg result **given** you do not have disease

P(A|B) is probability of disease *given* the test is positive (which is what we're interested in)

Very different from P(B|A): probability of positive test results given you have the disease.

A = disease
B = positive test result



A = disease

B = positive test result

P(A,B) is the *joint probability*, or the probability that both events occur. P(A,B) is the same as P(B,A).

But we already *know* that the test was positive, so we have to take that into account.

Of all the people already in the green circle, how many fall into the P(A,B) part? That's the probability we want to know!

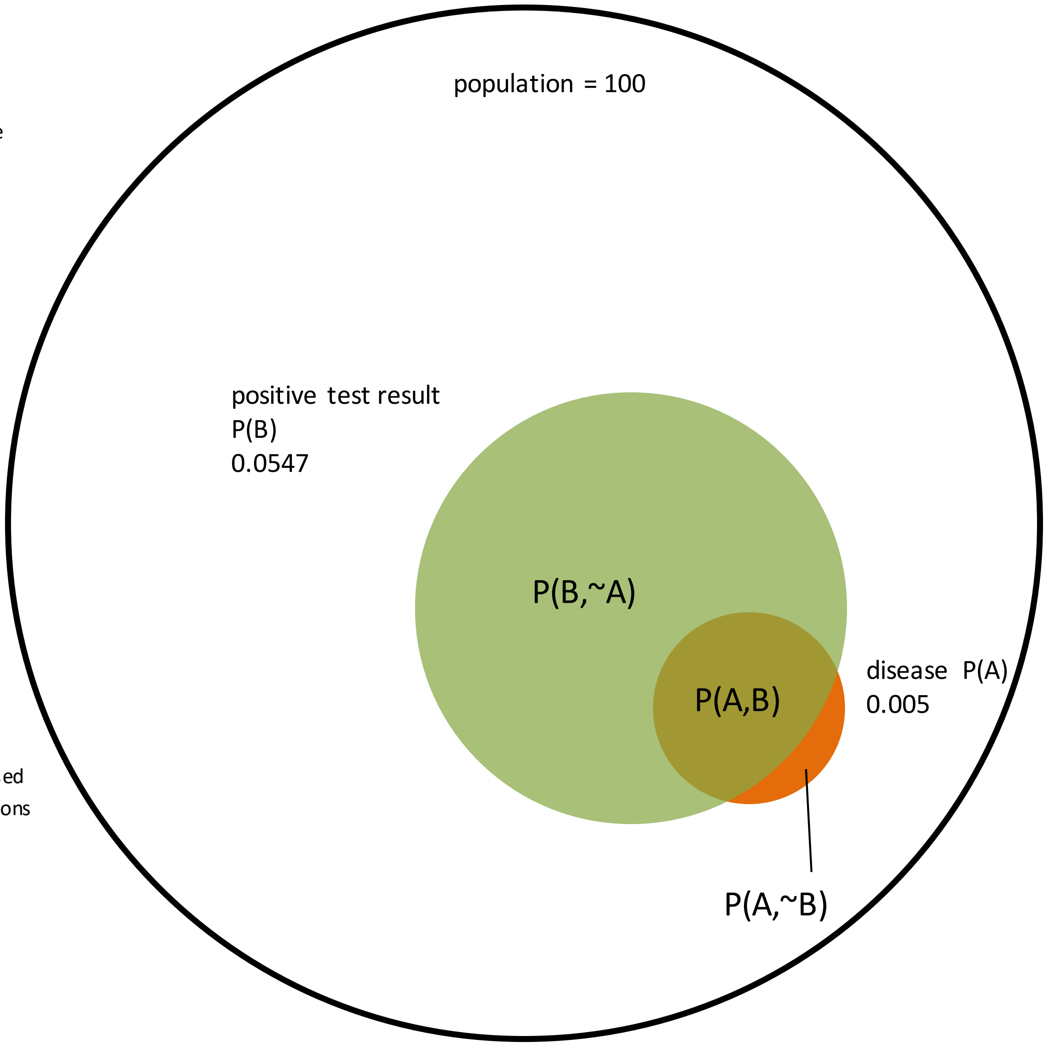
That is:
P(A|B) = P(A,B) / P(B)

You can write down same thing for the inverse:
P(B|A) = P(A,B) / P(A)

The *joint probability* can be expressed in two ways by rewriting the equations
P(A,B) = P(A|B) * P(B)
P(A,B) = P(B|A) * P(A)

Equating the two gives
P(A|B) * P(B) = P(B|A) * P(A)

P(A|B) = P(B|A) * P(A) / P(B)



A = disease
B = positive test result

P(A) = 0.005 probability of having disease
P(B|A) = 0.99 probability of pos result **given** you have disease
P(B) = 0.005 * 0.99 (people with disease) + 0.995 * 0.05 (people without disease) = 0.0547

Bayes' Theorem
P(A|B) = P(B|A) * P(A) / P(B)

P(A|B) = 0.99 * 0.005 / 0.0547
= 0.09

So a positive test result increases your probability of having the disease to 'only' 9%, simply because the disease is very rare (relative to the false positive rate).

P(A) is called the **prior**: before we have any information, we estimate the chance of having the disease 0.5%
P(B|A) is called the **likelihood**: probability of the data (pos test result) given an underlying cause (disease)
P(B) is the **marginal probability of the data:** the probability of observing this particular outcome, taken over all possible values of A (disease and no disease)
P(A|B) is the **posterior probability**: it is a combination of what you thought before obtaining the data, and the new information the data provided (combination of **prior** and **likelihood**)

Let's do another one…

It rains on 20% of days.
When it rains, it was forecasted 80% of the time
When it doesn't rain, it was erroneously forecasted 10% of the time.

The weatherman forecasts rain. What's the probability of it actually raining?

A = forecast rain
B = it rains

What information is given in the story?

$P(B) = 0.2$ (**prior**)
$P(A|B) = 0.8$ (**likelihood**)
$P(A|\sim B) = 0.1$

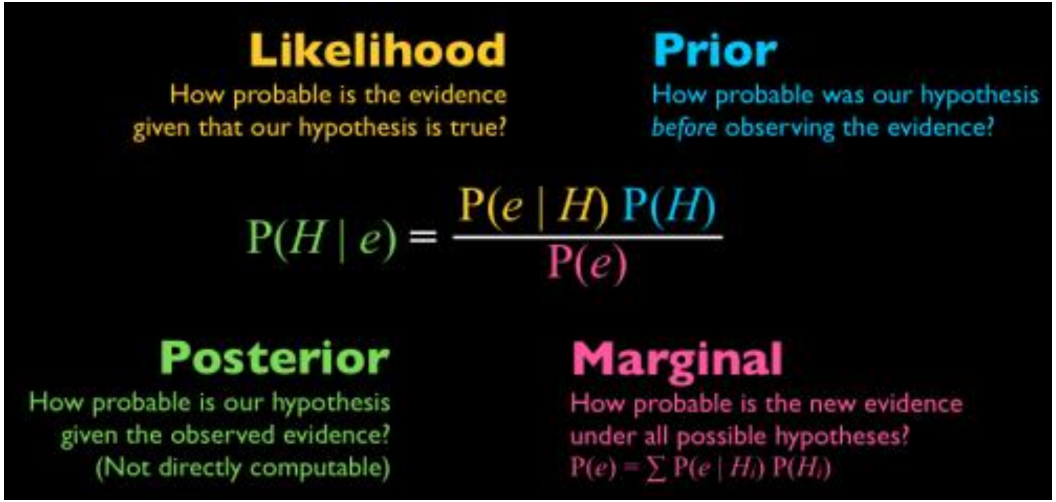
$P(B|A) = P(A|B) * P(B) / P(A)$

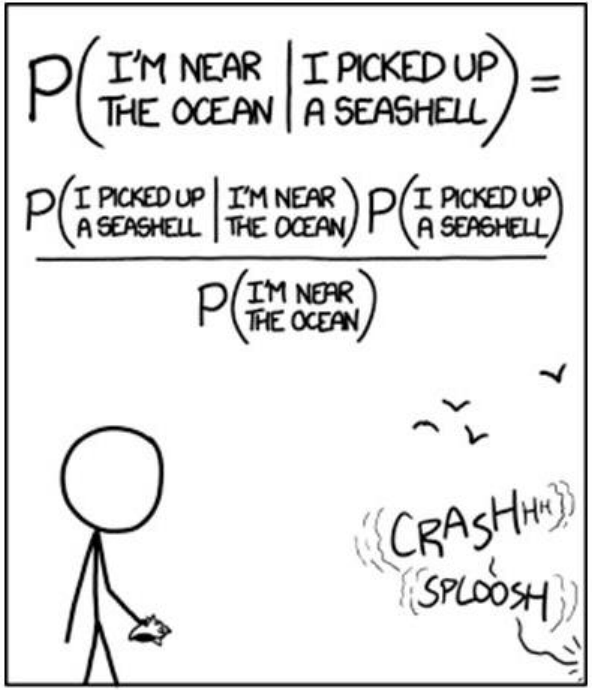
What is P(A), probability of rain forecast? Calculate over all possible values of B (**marginal probability**)
$P(A|B) * P(B) + P(A|\sim B) * P(\sim B) = 0.8 * 0.2 + 0.1 * 0.8 = 0.24$

$P(B|A) = 0.8 * 0.2 / 0.24$
$= 0.67$

So before you knew anything you thought P(rain) was 0.2. Now that you heard the weather forecast, you adjust your expectation upwards P(rain|forecast) = 0.67

Bayes Theorem





Naïve Bayes Classification

Material borrowed from
Jonathan Huang and
I. H. Witten's and E. Frank's "Data Mining"
and Jeremy Wyatt and others

Things We'd Like to Do

- Spam Classification
  - Given an email, predict whether it is spam or not

- Medical Diagnosis
  - Given a list of symptoms, predict whether a patient has disease X or not

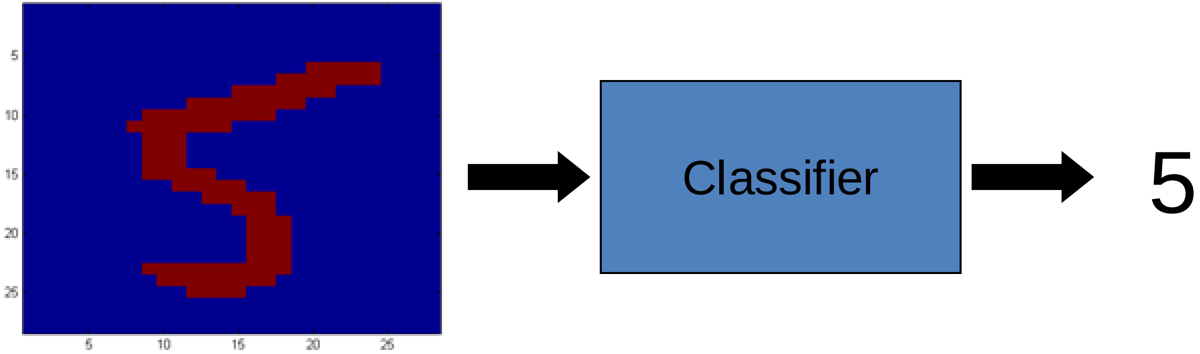
- Weather
  - Based on temperature, humidity, etc… predict if it will rain tomorrow

Bayesian Classification

- Problem statement:
  - Given features $X_1,X_2,\ldots,X_n$
  - Predict a label Y

Another Application

- **Digit Recognition**



- $X_1,\ldots,X_n \in \{0,1\}$ (Black vs. White pixels)
- $Y \in \{5,6\}$ (predict whether a digit is a 5 or a 6)

0 1

The Bayes Classifier

- A good strategy is to predict:

$$\arg\max_{Y} P(Y|X_1, \ldots, X_n)$$

  - (for example: what is the probability that the image represents a 5 given its pixels?)

- How do we compute that?

The Bayes Classifier

- Use Bayes Rule!

$$P(Y|X_1,\ldots,X_n) = \frac{\overbrace{P(X_1,\ldots,X_n|Y)}^{\text{Likelihood}}\overbrace{P(Y)}^{\text{Prior}}}{\underbrace{P(X_1,\ldots,X_n)}_{\text{Normalization Constant}}}$$

- Why did this help? Well, we think that we might be able to specify how features are “generated” by the class label

The Bayes Classifier

- Let's expand this for our digit recognition task:

P(x₁|5) P(x₂|5) . . . . . .

$$P(Y=5|X_1,\ldots,X_n) = \frac{P(X_1,\ldots,X_n|Y=5)P(Y=5)}{P(X_1,\ldots,X_n|Y=5)P(Y=5)+P(X_1,\ldots,X_n|Y=6)P(Y=6)}$$
$$P(Y=6|X_1,\ldots,X_n) = \frac{P(X_1,\ldots,X_n|Y=6)P(Y=6)}{P(X_1,\ldots,X_n|Y=5)P(Y=5)+P(X_1,\ldots,X_n|Y=6)P(Y=6)}$$

- To classify, we'll simply compute these two probabilities and predict based on which one is greater

Model Parameters

- For the Bayes classifier, we need to “learn” two functions, the likelihood and the prior

- How many parameters are required to specify the prior for our digit recognition example?

Model Parameters

- How many parameters are required to specify the likelihood?
  - (Supposing that each image is 30x30 pixels)

$2 \cdot 2^n$

?

$^n P(X_1, X_2 \ldots X_n | Y)$ $Y=5$

X_1	X_2	…	X_n	$P(X_1, X_2 \ldots | Y)$

$0_5 \to 30$

$\to 70$

$\to 80$

$1_5 \to 100$

Model Parameters

- The problem with explicitly modeling $P(X_1,\ldots,X_n|Y)$ is that there are usually way too many parameters:
  - We'll run out of space
  - We'll run out of time
  - And we'll need tons of training data (which is usually not available)

The Naïve Bayes Model

- The *Naïve Bayes Assumption*: Assume that all features are independent **given the class label Y**
- Equationally speaking:

$$P(X_1, \ldots, X_n | Y) = \prod_{i=1}^{n} P(X_i | Y)$$

- (We will discuss the validity of this assumption later)

$X_i =$

$P(X_1 = 1 | Y)$

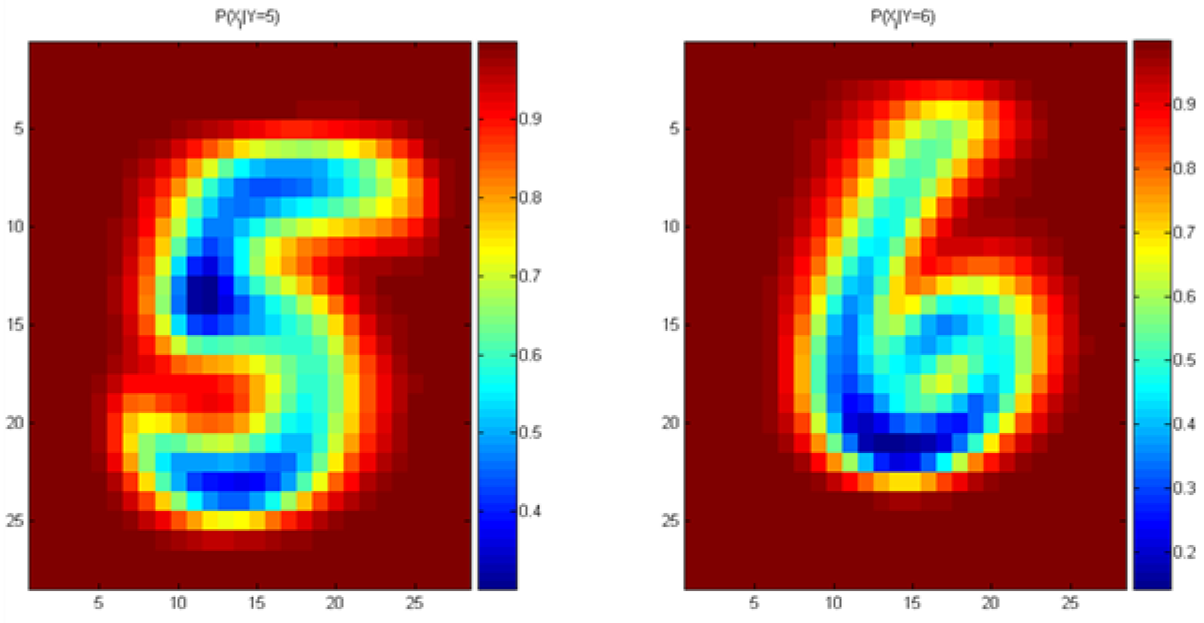
$P(X_2 = 1 | Y)$

$\vdots$

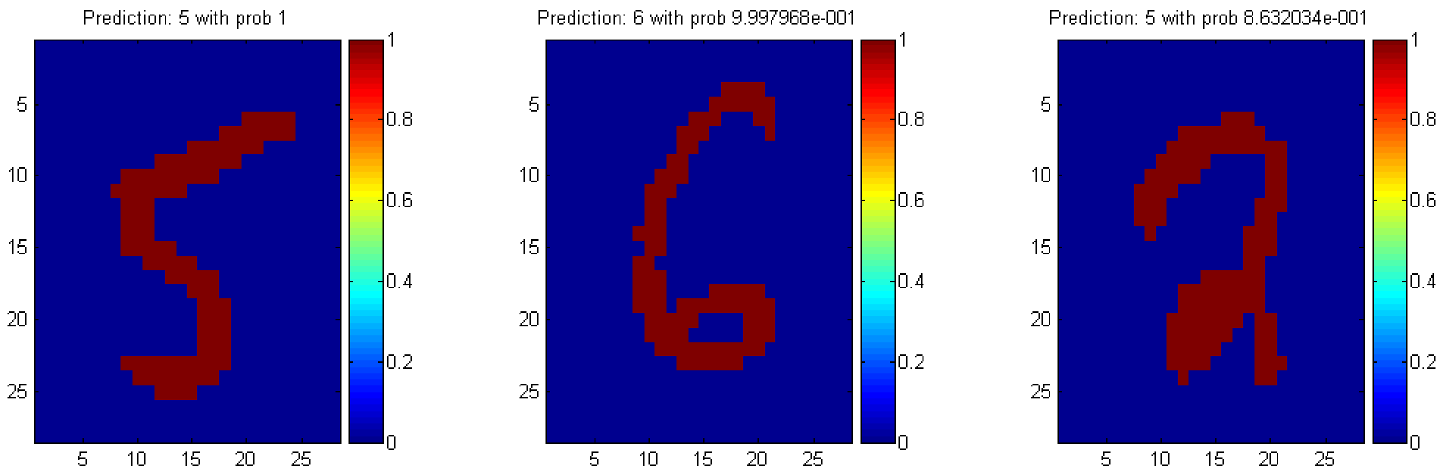
$P(X_n = 1 | Y)$

Naïve Bayes Training

- For binary digits, training amounts to averaging all of the training fives together and all of the training sixes together.



Naïve Bayes Classification



Numerical Attribute Values

- Assume normal distributions for numerical attributes.

The numeric weather data with summary statistics

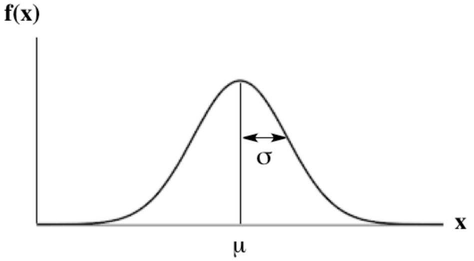
outlook			temperature			humidity			windy			play	
	yes	no		yes	no		yes	no		yes	no	yes	no
sunny	2	3		83	85		86	85	false	6	2	9	5
overcast	4	0		70	80		96	90	true	3	3		
rainy	3	2		68	65		80	70					
				64	72		65	95					
				69	71		70	91					
				75			80						
				75			70						
				72			90						
				81			75						
sunny	2/9	3/5	mean	73	74.6	mean	79.1	86.2	false	6/9	2/5	9/14	5/14
overcast	4/9	0/5	std dev	6.2	7.9	std dev	10.2	9.7	true	3/9	3/5		
rainy	3/9	2/5											

- Let x_1, x_2, …, x_n be the values of a numerical attribute in the training data set.

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

$$\sigma = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \mu)^2$$

$$f(w) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(w-\mu)^2}{\sigma^2}}$$



$$f(x) = \frac{1}{\boxed{\sigma}\sqrt{2\pi}} e^{-(x-\boxed{\mu})^2/2\sigma^2}$$

$e = 2.71828$

Strictly speaking, this is not a probability

- For examples,

$$f(\text{temperature}=66 \mid \text{Yes}) = \frac{1}{\sqrt{2\pi}(6.2)} e^{-\frac{(66-73)^2}{2(6.2)^2}} = 0.0340$$

- Likelihood of Yes = $$\frac{2}{9} \times 0.0340 \times 0.0221 \times \frac{3}{9} \times \frac{9}{14} = 0.000036$$

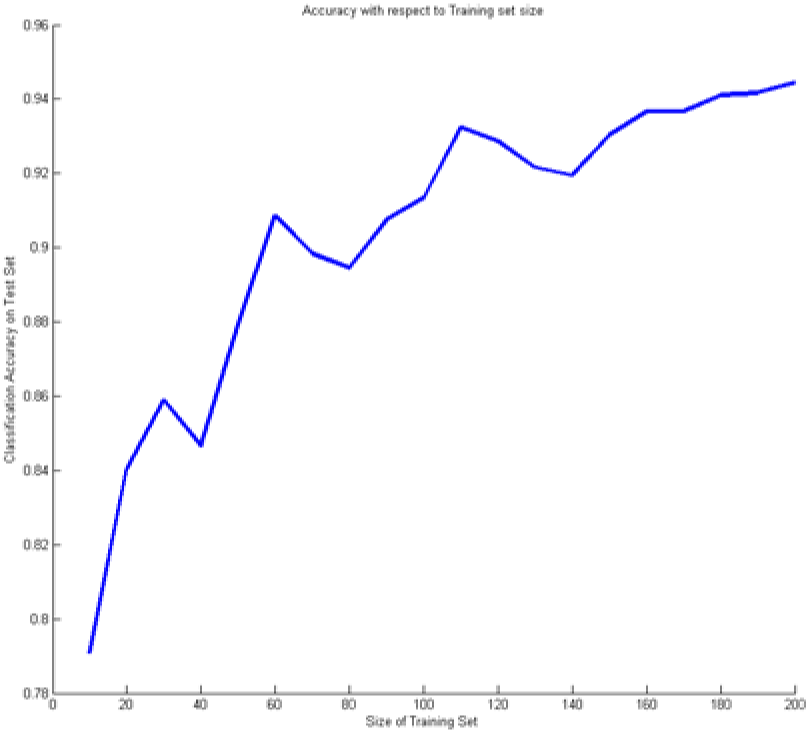
- Likelihood of No = $$\frac{3}{5} \times 0.0291 \times 0.038 \times \frac{3}{5} \times \frac{5}{14} = 0.000136$$

Outputting Probabilities

- What’s nice about Naïve Bayes (and generative models in general) is that it returns probabilities
  - **These probabilities can tell us how confident the algorithm is**

Performance on a Test Set

- Naïve Bayes is often a good choice if you don't have much training data!



Naïve Bayes Assumption

- Recall the Naïve Bayes assumption:
  - that all features are independent **given the class label Y**

- Does this hold for the digit recognition problem?

- Actually, the Naïve Bayes assumption is almost never true

- Still… Naïve Bayes often performs surprisingly well even when its assumptions do not hold

Numerical Stability

- It is often the case that machine learning algorithms need to work with very small numbers
  - Imagine computing the probability of 2000 independent coin flips
  - MATLAB/scikit-learn thinks that $(.5)^{2000}=0$

Underflow Prevention

- Multiplying lots of probabilities

➔ floating-point underflow.

- Recall: $log(xy) = log(x) + log(y)$,

➔ better to sum logs of probabilities rather than multiplying probabilities.

Underflow Prevention

- Class with highest final un-normalized log probability score is still the most probable.

$$c_{NB} = \underset{c_j \in C}{\operatorname{argmax}} \log P(c_j) + \sum_{i \in positions} \log P(x_i \mid c_j)$$

Numerical Stability

- Instead of comparing $P(Y=5|X_1,\ldots,X_n)$ with $P(Y=6|X_1,\ldots,X_n)$,
  - Compare their logarithms

$$
\begin{aligned}
\log\left(P(Y|X_1,\ldots,X_n)\right) &= \log\left(\frac{P(X_1,\ldots,X_n|Y)\cdot P(Y)}{P(X_1,\ldots,X_n)}\right) \\
&= \text{constant} + \log\left(\prod_{i=1}^{n} P(X_i|Y)\right) + \log P(Y) \\
&= \text{constant} + \sum_{i=1}^{n} \log P(X_i|Y) + \log P(Y)
\end{aligned}
$$

Recovering the Probabilities

- What if we want the probabilities though??
- Suppose that for some constant K, we have:

$$\log P(Y = 5|X_1, \dots, X_n) + K$$

  - And

$$\log P(Y = 6|X_1, \dots, X_n) + K$$

- How would we recover the original probabilities?

Recovering the Probabilities

- Given: $\alpha_i = \log a_i + K$
- Then for any constant C:

$$
\begin{aligned}
\frac{a_i}{\sum_i a_i} &= \frac{e^{\alpha_i}}{\sum_i e^{\alpha_i}} \\
&= \frac{e^C \cdot e^{\alpha_i}}{\sum_i e^C \cdot e^{\alpha_i}} \\
&= \frac{e^{\alpha_i + C}}{\sum_i e^{\alpha_i + C}}
\end{aligned}
$$

- One suggestion: set C such that the greatest α_i is shifted to zero:

$$C = -\max_i \{\alpha_i\}$$

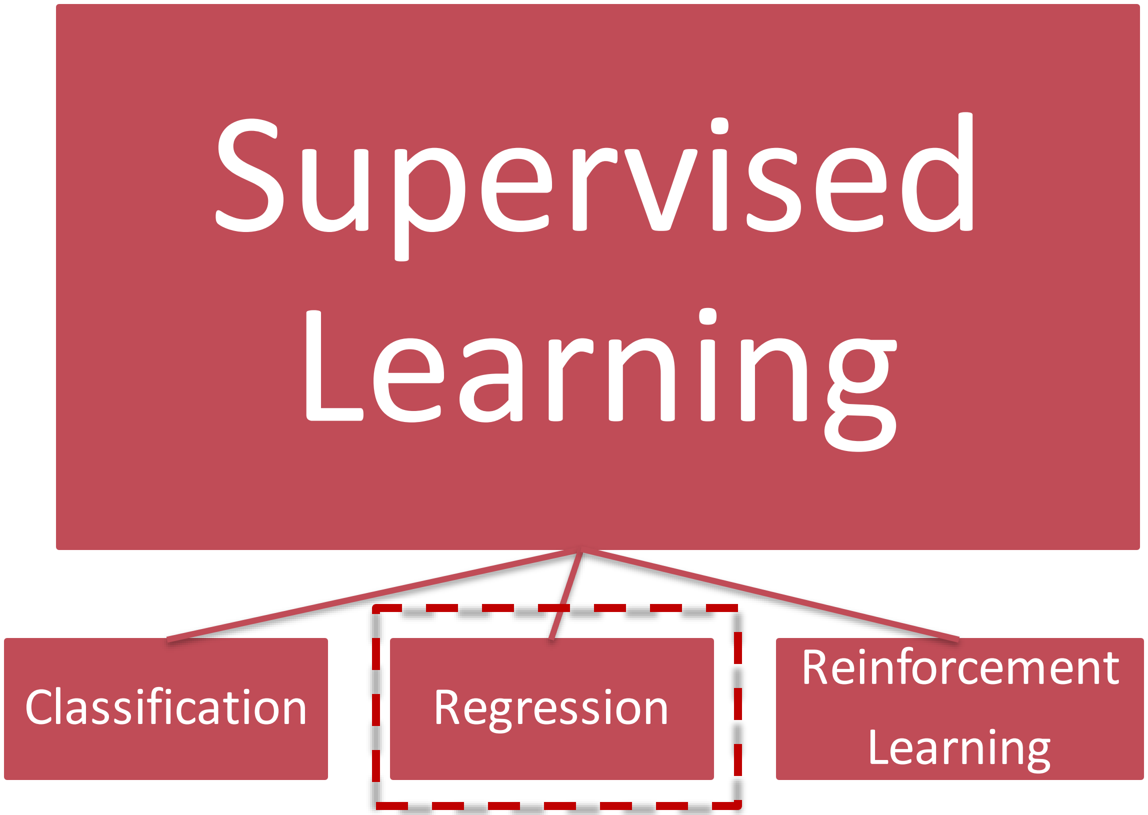
See https://stats.stackexchange.com/questions/105602/example-of-how-the-log-sum-exp-trick-works-in-naive-bayes?noredirect=1&lq=1

Recap

- We defined a *Bayes classifier* but saw that it's intractable to compute $P(X_1,\ldots,X_n|Y)$
- We then used the *Naïve Bayes assumption* – that everything is independent given the class label Y

Conclusions

- Naïve Bayes is:
  - Really easy to implement and often works well
  - Often a good first thing to try
  - Commonly used as a “punching bag” for smarter algorithms

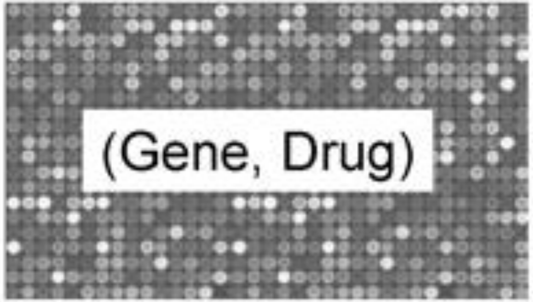

Supervised Learning
Classification
Regression
Reinforcement Learning

ML::Tasks → Predictive → Regression

Feature Space $\mathcal{X}$

Label Space $\mathcal{Y}$

Share Price
"$ 24.577"



Expression level
"6.88"

Continuous Labels

Regression model

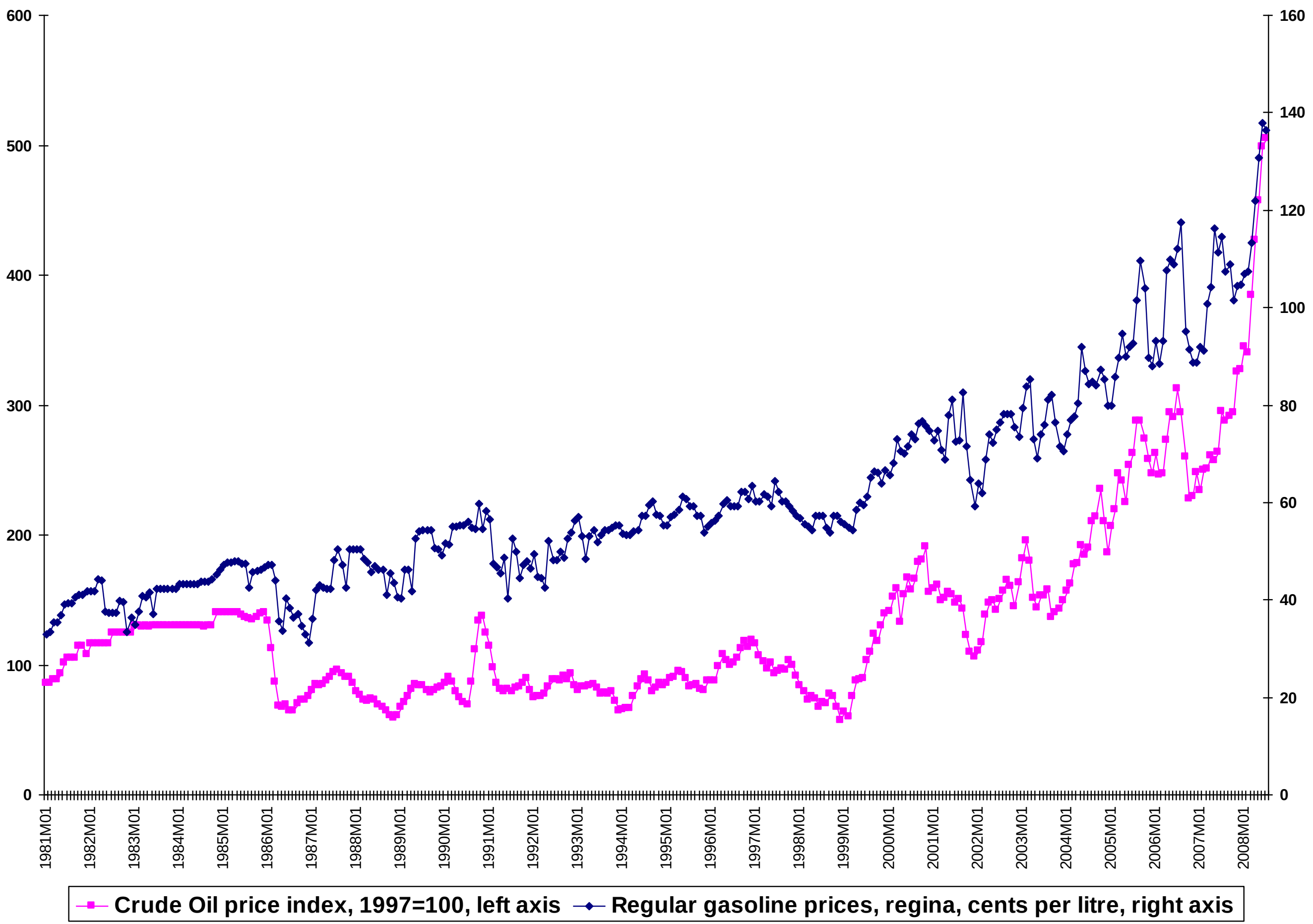
- Relation between variables where changes in some variables may "explain" changes in other variables.
- Regression model
  - Explanatory variables: **independent** variables
  - Variables to be explained : **dependent** variables
- estimates the <u>nature</u> of the relationship between the independent and dependent variables.
  - Change in dependent variables that results from changes in independent variables, i.e. <u>size</u> of the relationship
  - <u>Strength</u> of the relationship
  - <u>Statistical significance</u> of the relationship

Examples

- Independent variable: Price of crude oil
- Dependent variable: Retail price of petrol

- Independent variables: hours of work, education, occupation, sex, age, years of experience etc.
- Dependent variable: Employment income

- Price of a product and quantity produced or sold:
  - Quantity sold affected by price. Dependent variable is quantity of product sold – independent variable is price.
  - Price affected by quantity offered for sale. Dependent variable is price – independent variable is quantity sold.



Source: CANSIM Ⅱ Database (Vector v1576530 and v735048 respectively)

Bivariate and multivariate models

Bivariate or simple regression model

(Education) x $\longrightarrow$ y (Income)

Multivariate or multiple regression model

(Education) x_1

(Sex) x_2

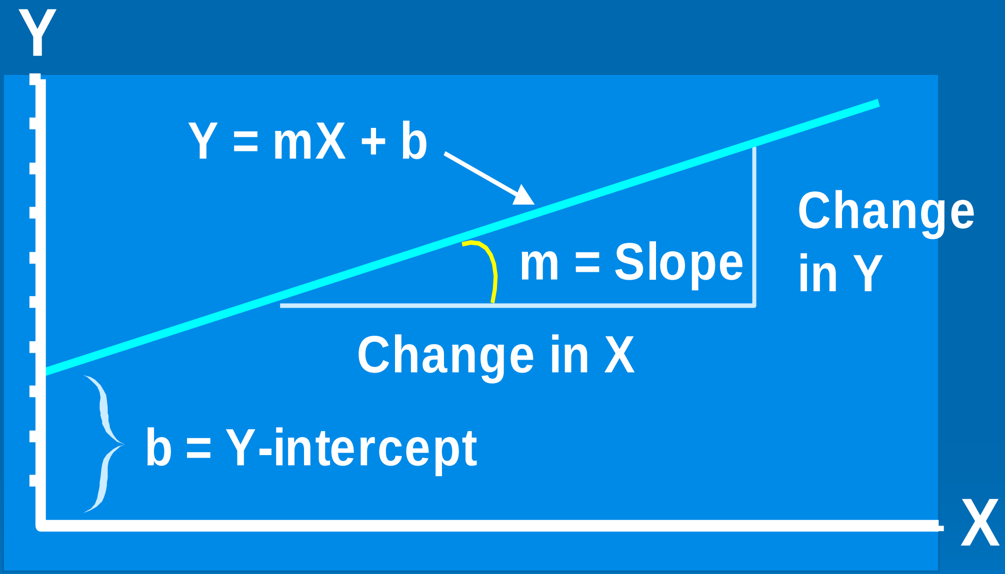
(Experience) x_3 $\longrightarrow$ y (Income)

(Age) x_4

Model with simultaneous relationship

Price of wheat $\longleftrightarrow$ Quantity of wheat produced

Linear Equations



Linear Regression Model

- 1. Relationship Between Variables Is a Linear Function

Population Y-Intercept

Population Slope

Random Error

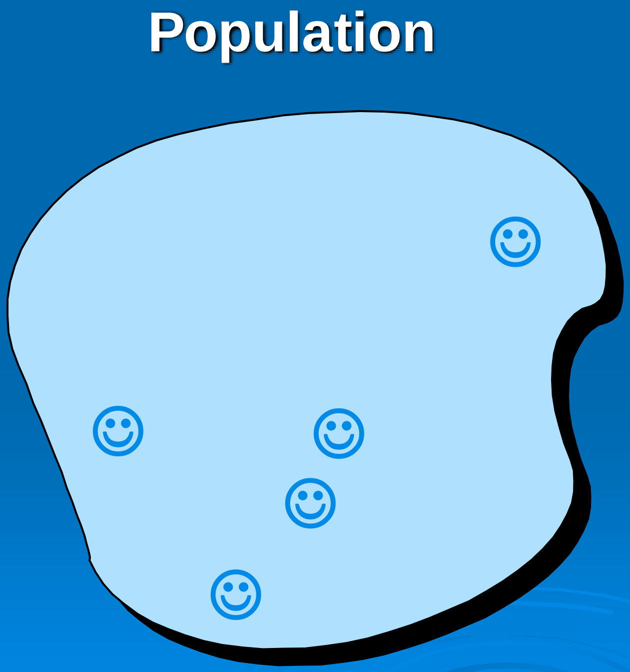
$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Dependent (Response) Variable (e.g. Salary)

Independent (Explanatory) Variable (e.g. Yrs of experience)

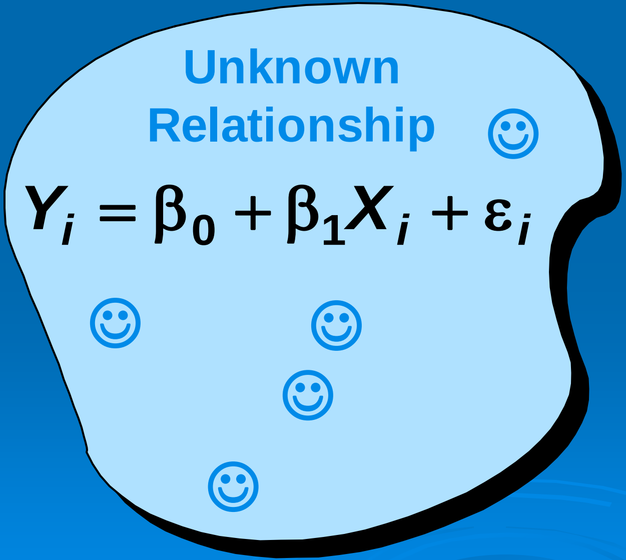
Population & Sample Regression Models

Population & Sample Regression Models

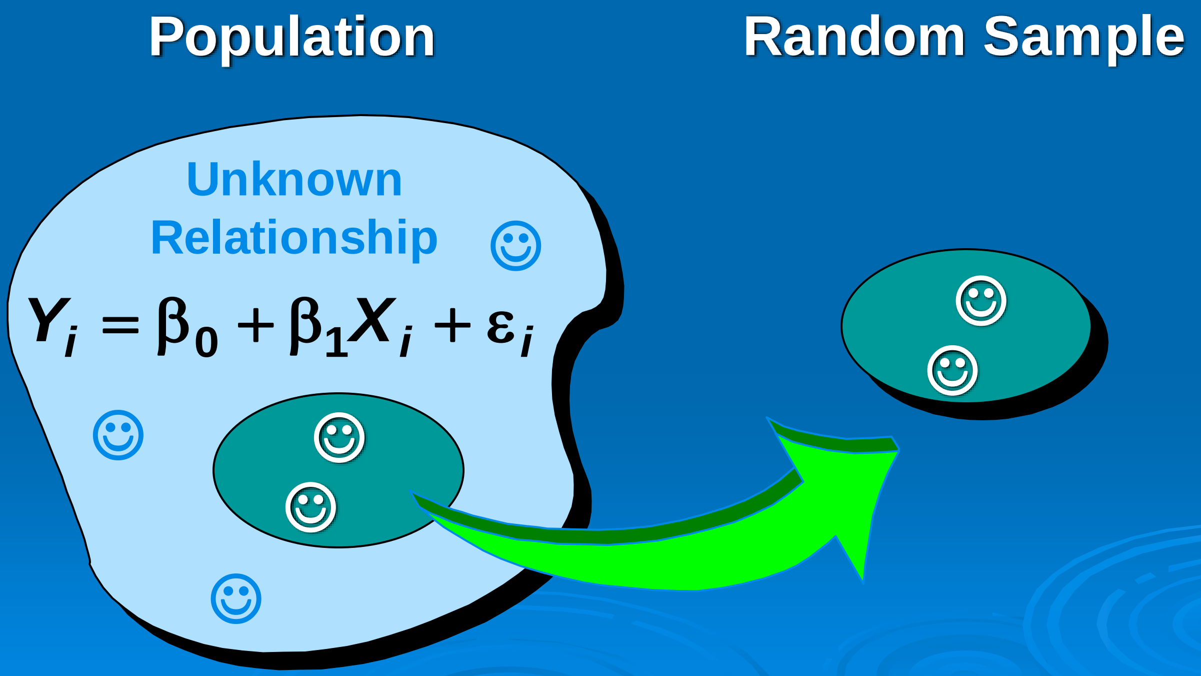


Population & Sample Regression Models

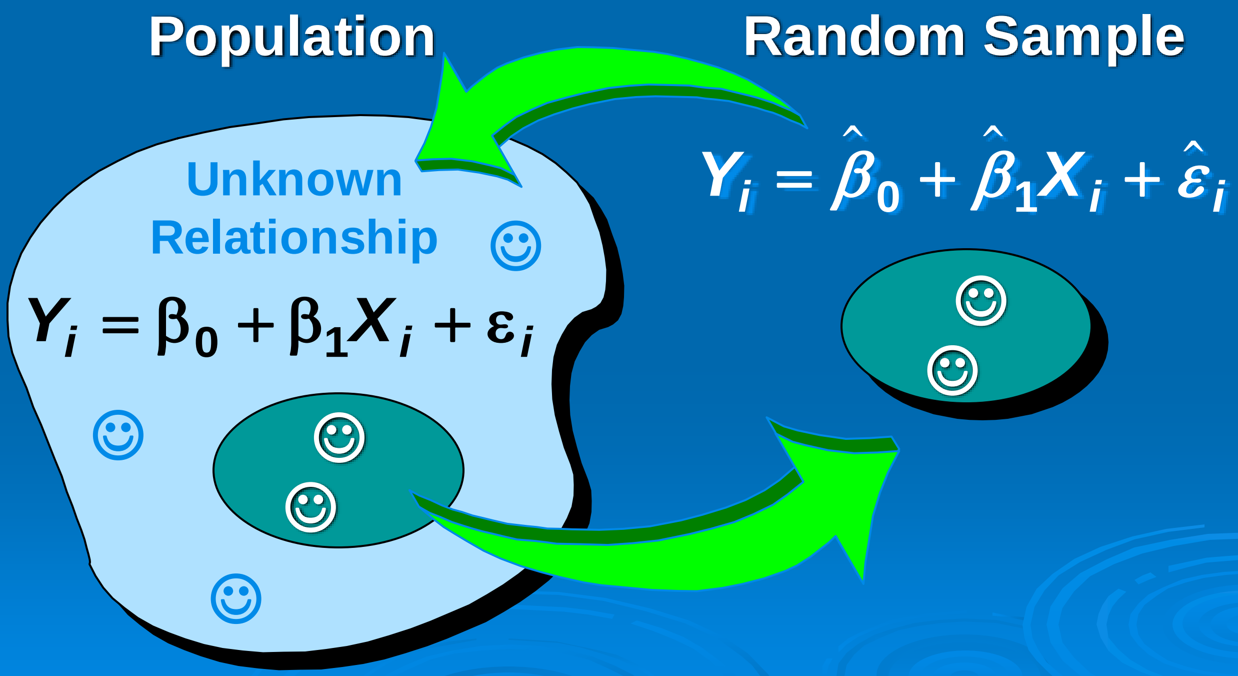
Population



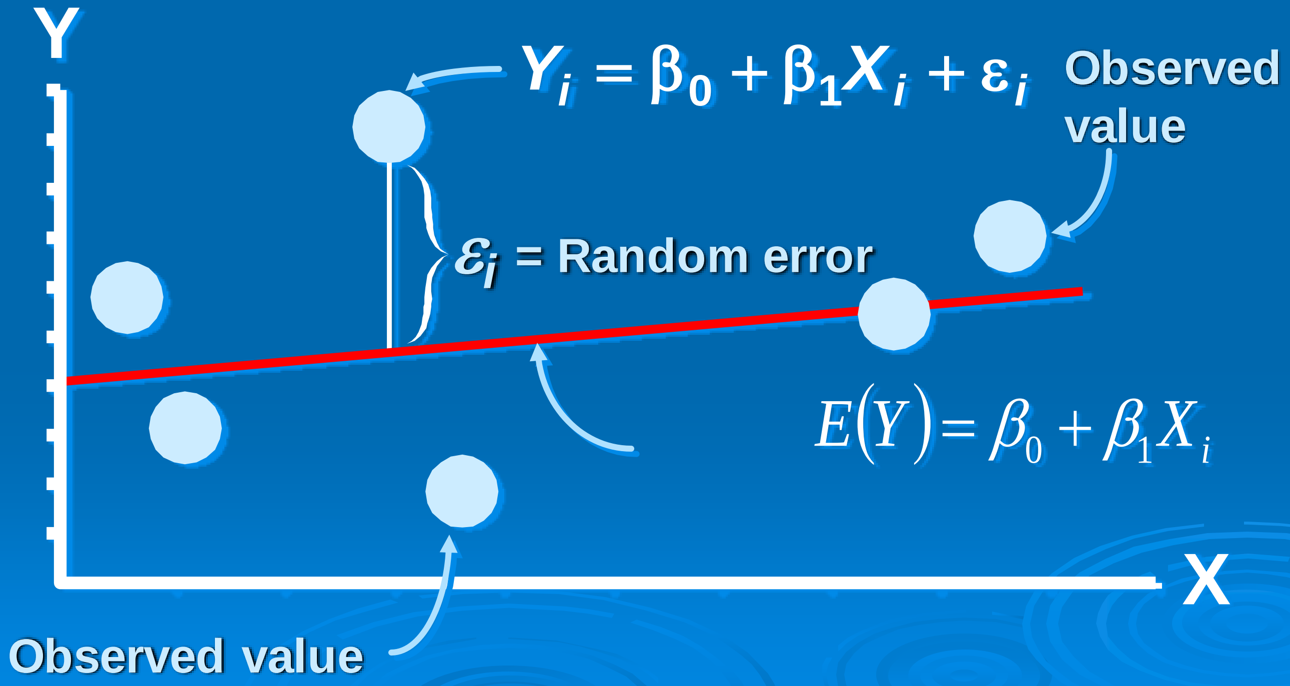
Population & Sample Regression Models



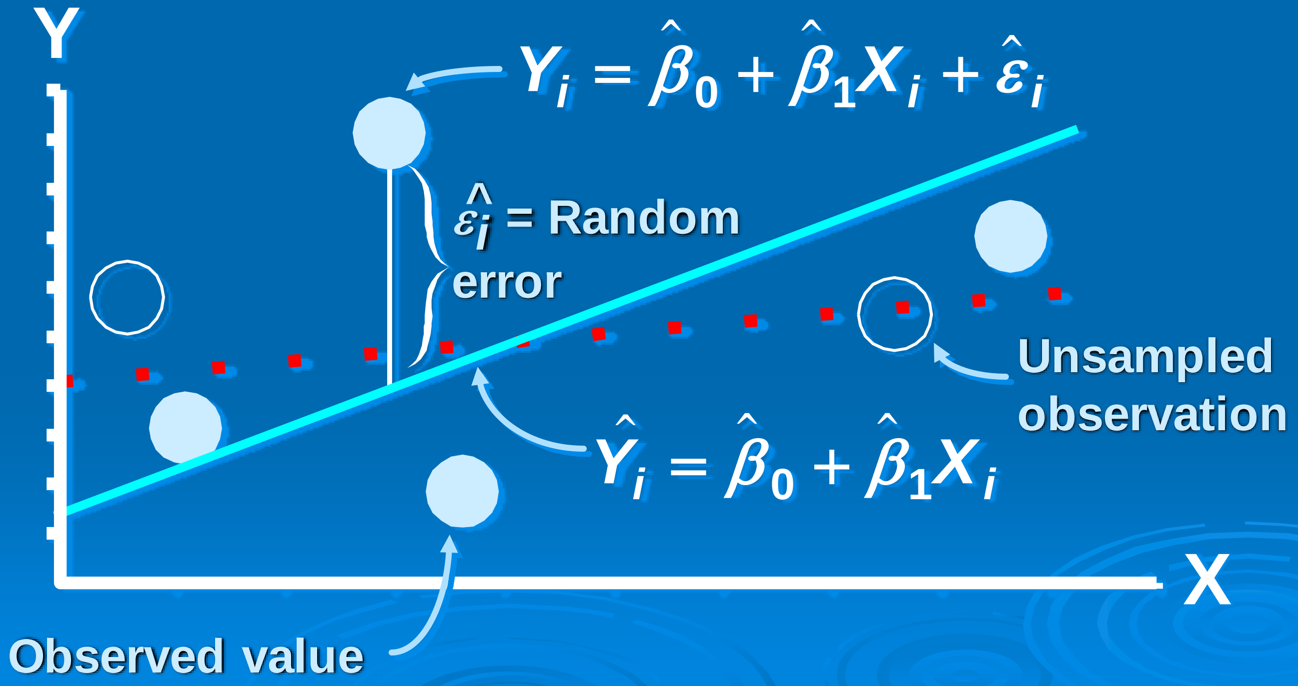
Population & Sample Regression Models



Population Linear Regression Model



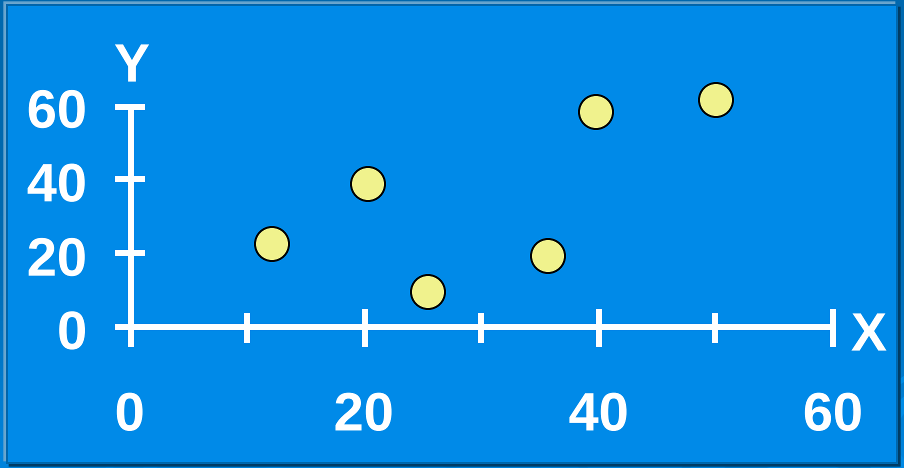
Sample Linear Regression Model



Estimating Parameters: Least Squares Method

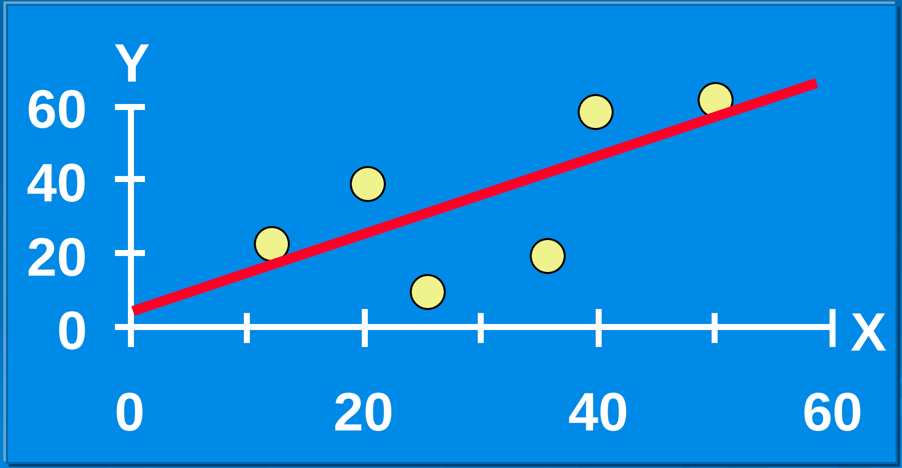
Scatter plot

- 1. Plot of All (X_i, Y_i) Pairs
- 2. Suggests How Well Model Will Fit



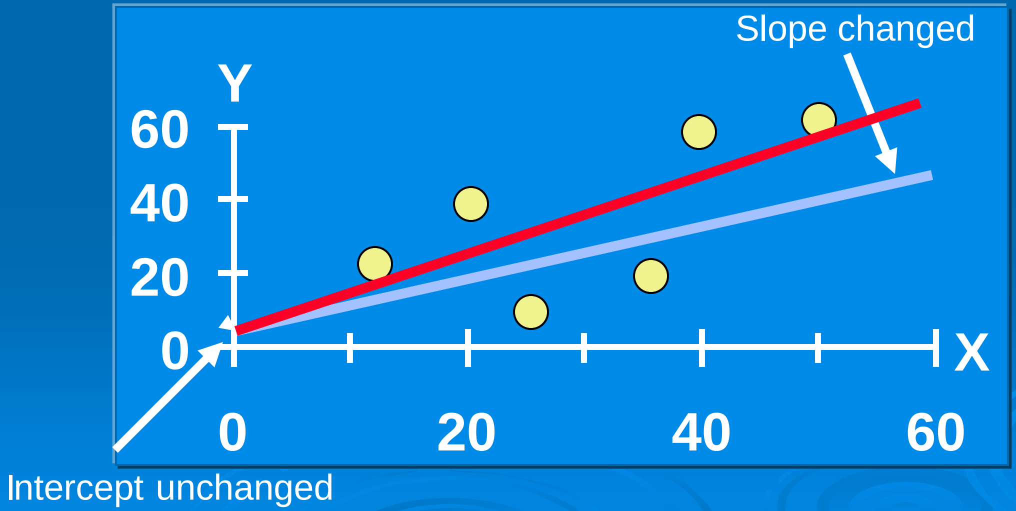
Thinking Challenge

How would you draw a line through the points? How do you determine which line 'fits best'?



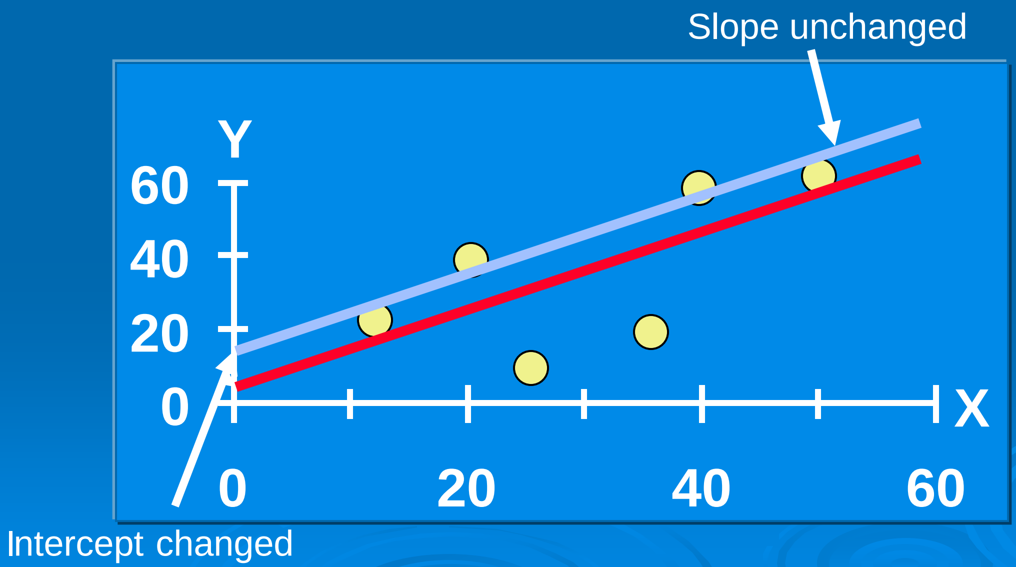
Thinking Challenge

How would you draw a line through the points? How do you determine which line 'fits best'?



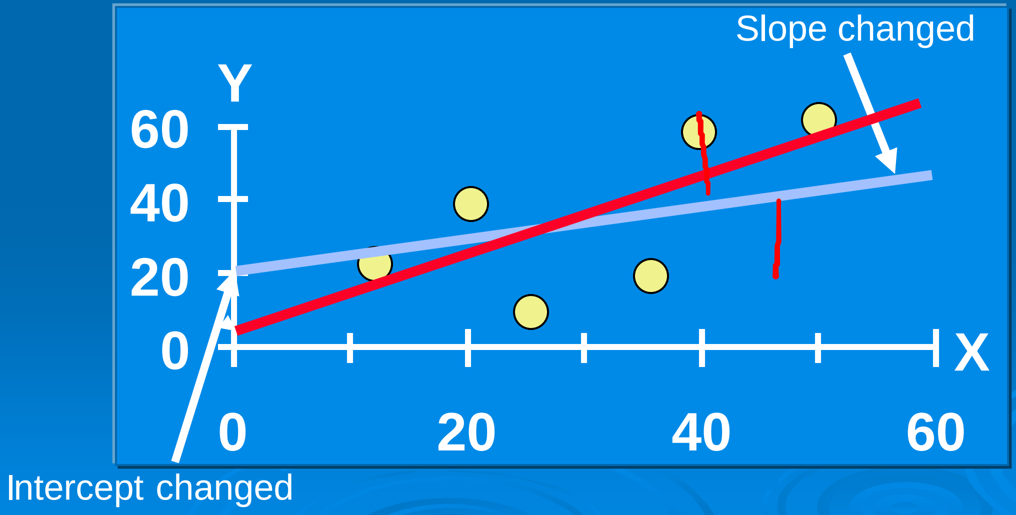
Thinking Challenge

How would you draw a line through the points? How do you determine which line 'fits best'?



Thinking Challenge

How would you draw a line through the points? How do you determine which line 'fits best'?



Least Squares

- 1. ‘Best Fit’ Means Difference Between Actual Y Values & Predicted Y Values Are a Minimum. *But* Positive Differences Off-Set Negative ones

Least Squares

- 1. 'Best Fit' Means Difference Between Actual Y Values & Predicted Y Values is a Minimum. *But* Positive Differences Off-Set Negative ones. So square errors!

$$\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2 = \sum_{i=1}^{n}\hat{\varepsilon}_i^2$$

Least Squares

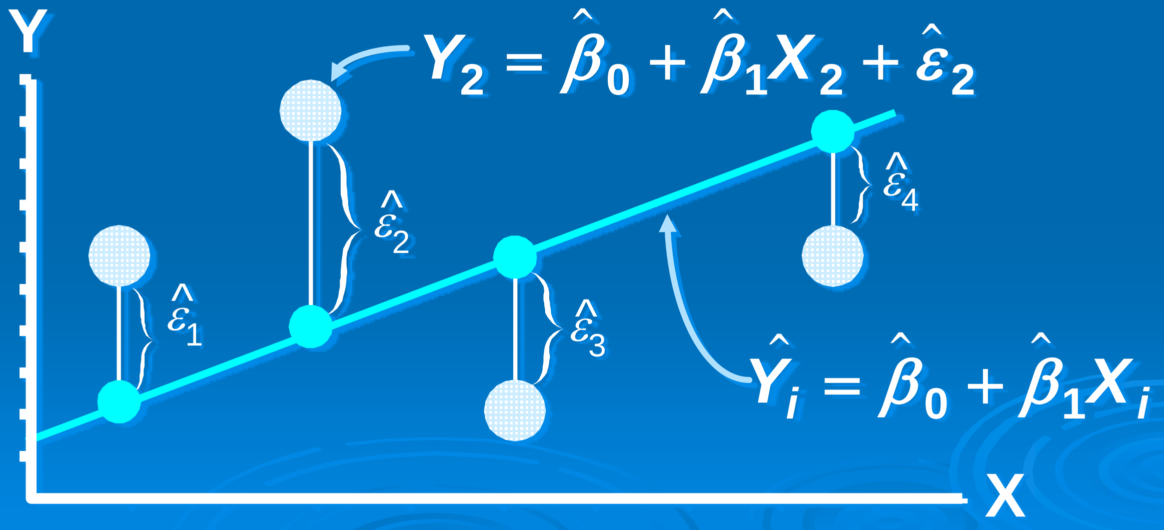
- 1. 'Best Fit' Means Difference Between Actual Y Values & Predicted Y Values Are a Minimum. *But* Positive Differences Off-Set Negative. So square errors!

$$\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2 = \sum_{i=1}^{n}\hat{\varepsilon}_i^2$$

- 2. LS Minimizes the Sum of the Squared Differences (errors) (SSE)

Least Squares Graphically

LS minimizes $\sum_{i=1}^{n} \hat{\varepsilon}_i^2 = \hat{\varepsilon}_1^2 + \hat{\varepsilon}_2^2 + \hat{\varepsilon}_3^2 + \hat{\varepsilon}_4^2$



Coefficient Equations

- Prediction equation

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

- Sample slope

$$\hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

- Sample Y - intercept

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Derivation of Parameters (1)

- Least Squares (L-S):

  Minimize squared error

$$\sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} \left(y_i - \beta_0 - \beta_1 x_i\right)^2$$

$$0 = \frac{\partial \sum \varepsilon_i^2}{\partial \beta_0} = \frac{\partial \sum \left(y_i - \beta_0 - \beta_1 x_i\right)^2}{\partial \beta_0}$$

$$= -2\left(n\bar{y} - n\beta_0 - n\beta_1 \bar{x}\right)$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Derivation of Parameters (1)

- Least Squares (L-S):

  Minimize squared error

$$0 = \frac{\partial \sum \varepsilon_i^2}{\partial \beta_1} = \frac{\partial \sum \left(y_i - \beta_0 - \beta_1 x_i\right)^2}{\partial \beta_1}$$

$$= -2 \sum x_i \left(y_i - \beta_0 - \beta_1 x_i\right)$$

$$= -2 \sum x_i \left(y_i - \bar{y} + \beta_1 \bar{x} - \beta_1 x_i\right)$$

$$\beta_1 \sum x_i \left(x_i - \bar{x}\right) = \sum x_i \left(y_i - \bar{y}\right)$$

$$\beta_1 \sum \left(x_i - \bar{x}\right)\left(x_i - \bar{x}\right) = \sum \left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)$$

$$\hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}}$$

Computation Table

X_i	Y_i	X_i^2	Y_i^2	X_iY_i
X_1	Y_1	X_1^2	Y_1^2	X_1Y_1
X_2	Y_2	X_2^2	Y_2^2	X_2Y_2
:	:	:	:	:
X_n	Y_n	X_n^2	Y_n^2	X_nY_n
ΣX_i	ΣY_i	ΣX_i^2	ΣY_i^2	ΣX_iY_i

Interpretation of Coefficients

Interpretation of Coefficients

- 1. Slope ($\hat{\beta}_1$)
  - Estimated Y Changes by $\hat{\beta}_1$ for Each 1 Unit Increase in X
    - If $\hat{\beta}_1 = 2$, then Y Is Expected to Increase by 2 for Each 1 Unit Increase in X

Interpretation of Coefficients

- 1. Slope ($\hat{\beta}_1$)
  - Estimated Y Changes by $\hat{\beta}_1$ for Each 1 Unit Increase in X
    - If $\hat{\beta}_1 = 2$, then Y Is Expected to Increase by 2 for Each 1 Unit Increase in X
- 2. Y-Intercept ($\hat{\beta}_0$)
  - Average Value of Y When $X = 0$
    - If $\hat{\beta}_0 = 4$, then Average Y Is Expected to Be 4 When X Is 0

References and Reading

- https://www.mathsisfun.com/data/index.html#stats
- Bayes for Beginners: https://www.fil.ion.ucl.ac.uk/mfd_archive/2011/page1/mfd2011_bayes.pptx
- The Bayesian Trap: https://www.youtube.com/watch?v=R13BD8qKeTg

- Reading:
  - PRML, Bishop: Chapter 8, Section 3.3. **Figure 1.27 provides a nice illustration for why Naïve Bayes may perform well despite making a naïve assumption and not modelling the actual (joint) likelihood/posterior.**