AI: 15-780 / 16-731
Mar 1, 2007

Probability Theory & Uncertainty

Read Chapter 13 of textbook

What you will learn today

- fundamental role of uncertainty in AI
- probability theory can be applied to many of these problems
- probability as uncertainty
- probability theory is the calculus of reasoning with uncertainty
- probability and uncertainty in different contexts
- review of basis probabilistic concepts
  - discrete and continuous probability
  - joint and marginal probability
  - calculating probability

- next probability lecture: the process of probabilistic inference

What is the role of probability and inference in AI?

- Many algorithms are designed as if knowledge is perfect, but it rarely is.
- There are almost always things that are unknown, or not precisely known.

- Examples:
  - bus schedule
  - quickest way to the airport
  - sensors
  - joint positions
  - finding an H-bomb

- An agent making optimal decisions must take into account *uncertainty*.

Probability as frequency: *k* out of *n* possibilities

- Suppose we're drawing cards from a standard deck:
  - P(card is the Jack ♥ | standard deck) = 1/52
  - P(card is a ♣ | standard deck) = 13/52 = 1/4

- What's the probability of a drawing a pair in 5-card poker?
  - P(hand contains pair | standard deck) =

$$\frac{\text{\# of hands with pairs}}{\text{total \# of hands}}$$

  - Counting can be tricky (take a course in combinatorics)
  - Other ways to solve the problem?

- General probability of *event* given some conditions:

  P(event | conditions)

Making rational decisions when faced with uncertainty

- *Probability*

  the precise representation of knowledge and uncertainty
- *Probability theory*

  how to optimally update your knowledge based on new information
- *Decision theory: probability theory + utility theory*

  how to use this information to achieve maximum expected utility

- Consider again the bus schedule. What's the utility function?
  - Suppose the schedule says the bus comes at 8:05.
  - Situation A: You have a class at 8:30.
  - Situation B: You have a class at 8:30, and it's cold and raining.
  - Situation C: You have a final exam at 8:30.

Probability of uncountable events

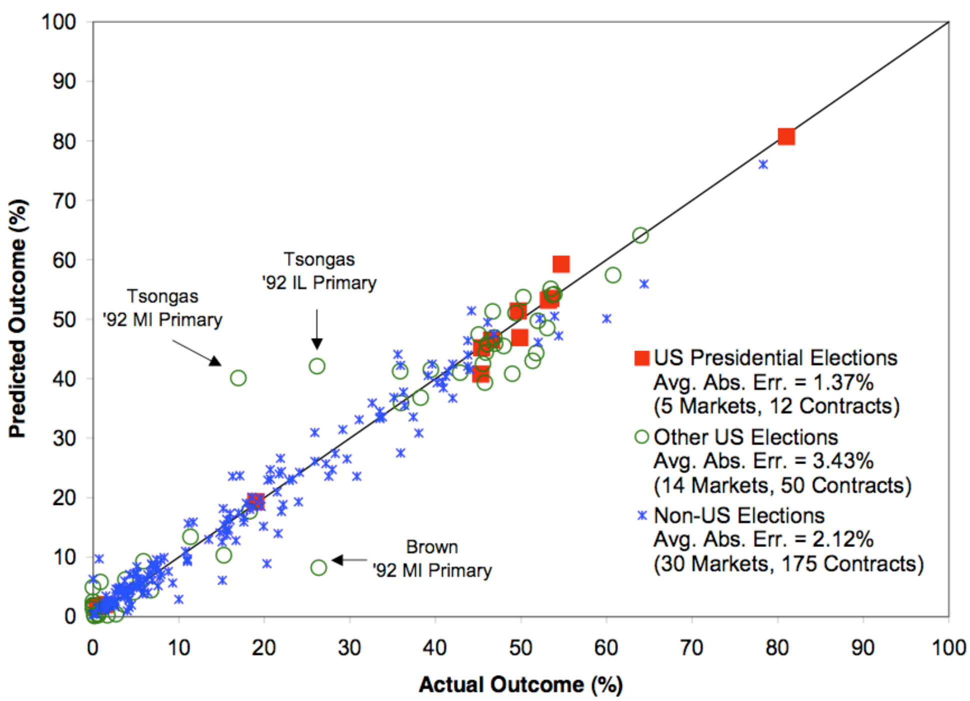
- How do we calculate probability that it will rain tomorrow?
  - Look at historical trends?
  - Assume it generalizes?

- What's the probability that there was life on Mars?
- What was the probability the sea level will rise 1 meter within the century?
- What's the probability that candidate X will win the election?

The Iowa Electronic Markets: placing probabilities on single events

- http://www.biz.uiowa.edu/iem/
- "The Iowa Electronic Markets are real-money futures markets in which contract payoffs depend on economic and political events such as elections."
- Typical bet: predict vote share of candidate X - "a vote share market"

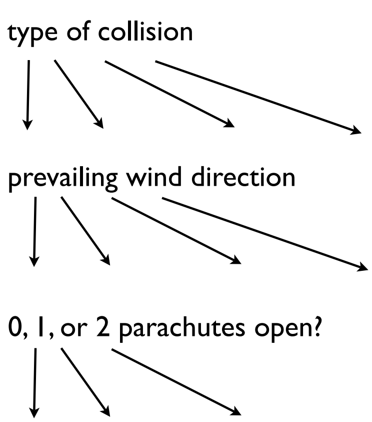
Political futures market predicted vs actual outcomes



John Craven and the missing H-Bomb

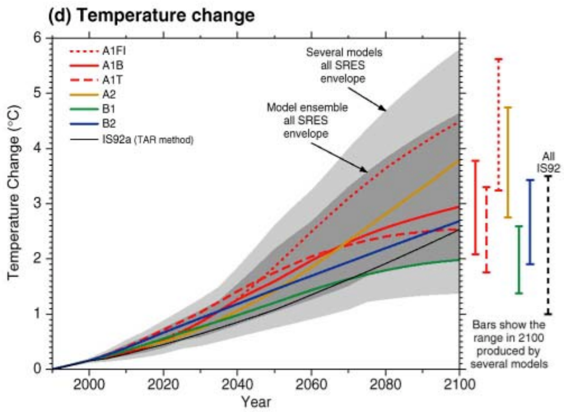
- In Jan. 1966, used Bayesian probability and subjective odds to locate H-bomb missing in the Mediterranean ocean.

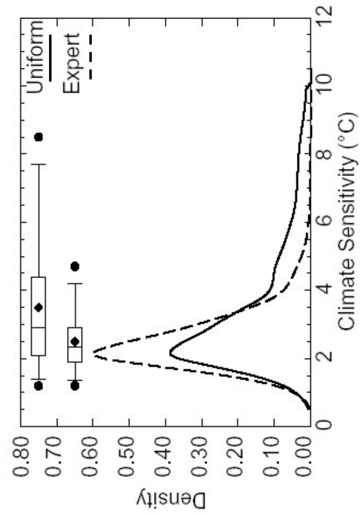
Probabilistic Methodology



Probabilistic assessment of dangerous climate change

from Mastrandrea and Schneider (2004)



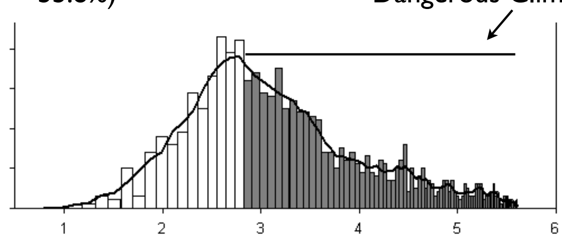


from Forrest et al (2001)

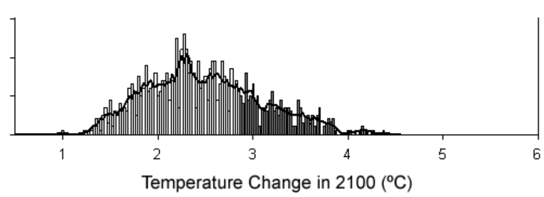
Factoring in Risk Using Decision Theory

P("DAI" = 55.8%)

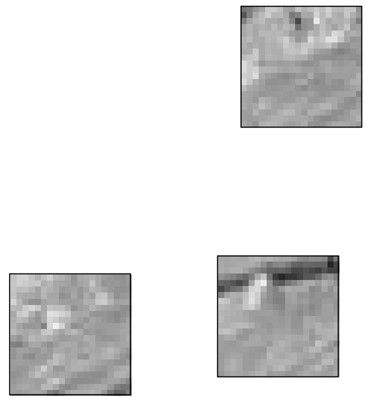
Dangerous Climate Change



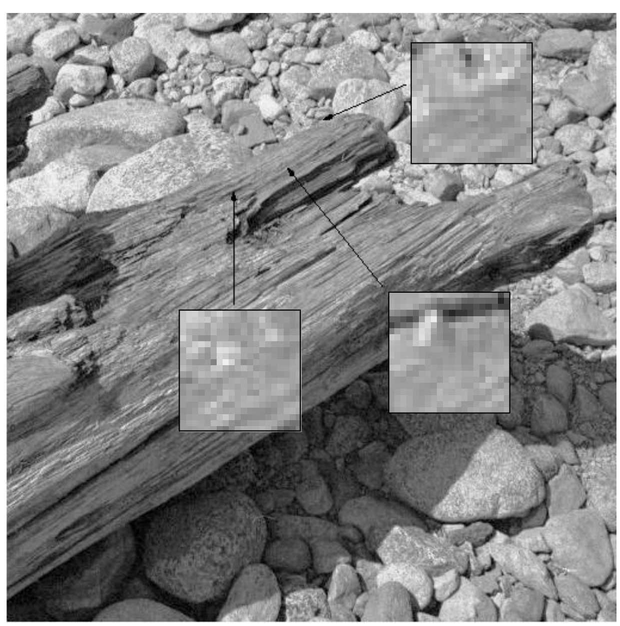
P("DAI" = 27.4%
Carbon Tax 2050
= $174/Ton



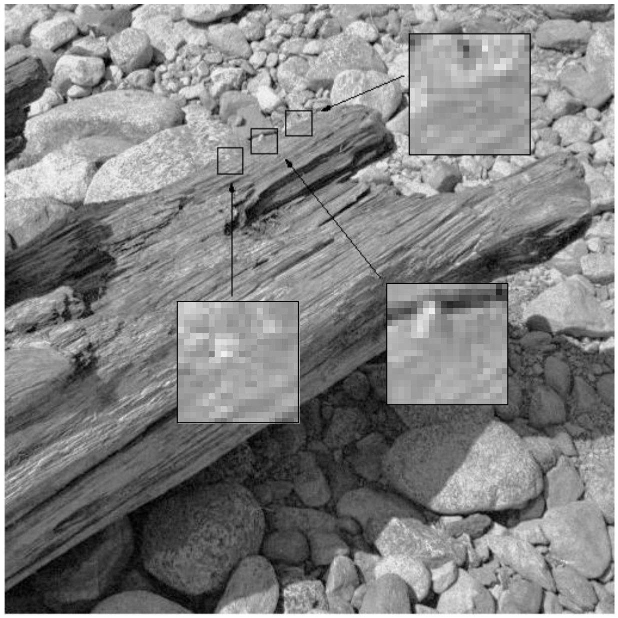
Uncertainty in vision: What are these?



Uncertainty in vision

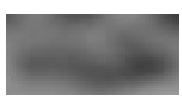


Edges are not as obvious they seem



An example from Antonio Torralba

What's this?



We constantly use other information to resolve uncertainty

Image interpretation is heavily context dependent

This phenomenon is even more prevalent in speech perception

- It is very difficult to recognize phonemes from naturally spoken speech when they are presented in isolation.
- All modern speech recognition systems rely heavily on context (as do we).
- HMMs model this contextual dependence explicitly.
- This allows the recognition of words, even if there is a great deal of uncertainty in each of the individual parts.

De Finetti's definition of probability

- Was there life on Mars?
- You promise to pay \$1 if there is, and \$0 if there is not.
- Suppose NASA will give us the answer tomorrow.
- Suppose you have an oppenent
  - You set the odds (or the "subjective probability") of the outcome
  - But your oppenent decides which side of the bet will be yours
- de Finetti showed that the price you set has to obey the axioms of probability or you face certain loss, i.e. you'll lose every time.

Axioms of probability

- Axioms (Kolmogorov):

  0 ≤ P(A) ≤ 1

  P(true) = 1

  P(false) = 0

  P(A or B) = P(A) + P(B) − P(A and B)

- Corollaries:
  - A single random variable must sum to 1:

$$\sum_{i=1}^{n} P(D = d_i) = 1$$

  - The joint probability of a set of variables must also sum to 1.
  - If A and B are mutually exclusive:

    P(A or B) = P(A) + P(B)

Rules of probability

- conditional probability

$$Pr(A|B) = \frac{Pr(A \text{ and } B)}{Pr(B)}, \qquad Pr(B) > 0$$

- corollary (Bayes' rule)

$$\begin{aligned} Pr(B|A)Pr(A) &= Pr(A \text{ and } B) = Pr(A|B)Pr(B) \\ \Rightarrow Pr(B|A) &= \frac{Pr(A|B)Pr(B)}{Pr(A)} \end{aligned}$$

Discrete probability distributions

- discrete probability distribution
- joint probability distribution
- marginal probability distribution
- Bayes' rule
- independence

The Joint Distribution

Example: Boolean variables A, B, C

Recipe for making a joint distribution of M variables:

1. Make a truth table listing all combinations of values of your variables (if there are M Boolean variables then the table will have 2^M rows).

A	B	C
0	0	0
0	0	1
0	1	0
0	1	1
1	0	0
1	0	1
1	1	0
1	1	1

All the nice looking slides like this one from now on are from Andrew Moore.

The Joint Distribution

Example: Boolean variables A, B, C

Recipe for making a joint distribution of M variables:

1. Make a truth table listing all combinations of values of your variables (if there are M Boolean variables then the table will have 2^M rows).
2. For each combination of values, say how probable it is.

A	B	C	Prob
0	0	0	0.30
0	0	1	0.05
0	1	0	0.10
0	1	1	0.05
1	0	0	0.05
1	0	1	0.10
1	1	0	0.25
1	1	1	0.10

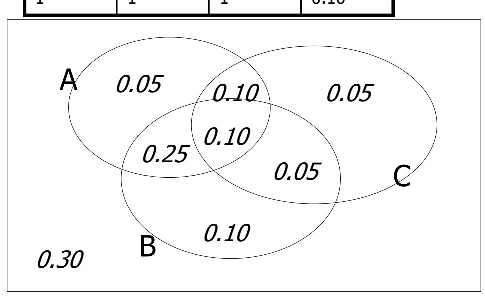
The Joint Distribution

Example: Boolean variables A, B, C

Recipe for making a joint distribution of M variables:

1. Make a truth table listing all combinations of values of your variables (if there are M Boolean variables then the table will have 2^M rows).
2. For each combination of values, say how probable it is.
3. If you subscribe to the axioms of probability, those numbers must sum to 1.

A	B	C	Prob
0	0	0	0.30
0	0	1	0.05
0	1	0	0.10
0	1	1	0.05
1	0	0	0.05
1	0	1	0.10
1	1	0	0.25
1	1	1	0.10



Using the Joint

gender	hours_worked	wealth	
Female	v0:40.5-	poor	0.253122
		rich	0.0245895
	v1:40.5+	poor	0.0421768
		rich	0.0116293
Male	v0:40.5-	poor	0.331313
		rich	0.0971295
	v1:40.5+	poor	0.134106
		rich	0.105933

One you have the JD you can ask for the probability of any logical expression involving your attribute

$$P(E) = \sum_{\text{rows matching } E} P(\text{row})$$

Using the Joint

gender	hours_worked	wealth	
Female	v0:40.5-	poor	0.253122
		rich	0.0245895
	v1:40.5+	poor	0.0421768
		rich	0.0116293
Male	v0:40.5-	poor	0.331313
		rich	0.0971295
	v1:40.5+	poor	0.134106
		rich	0.105933

P(Poor Male) = 0.4654

$$P(E) = \sum_{\text{rows matching } E} P(\text{row})$$

Using the Joint

gender	hours_worked	wealth	
Female	v0:40.5-	poor	0.253122
		rich	0.0245895
	v1:40.5+	poor	0.0421768
		rich	0.0116293
Male	v0:40.5-	poor	0.331313
		rich	0.0971295
	v1:40.5+	poor	0.134106
		rich	0.105933

P(Poor) = 0.7604

$$P(E) = \sum_{\text{rows matching } E} P(\text{row})$$

Inference with the Joint

gender	hours_worked	wealth	
Female	v0:40.5-	poor	0.253122
		rich	0.0245895
	v1:40.5+	poor	0.0421768
		rich	0.0116293
Male	v0:40.5-	poor	0.331313
		rich	0.0971295
	v1:40.5+	poor	0.134106
		rich	0.105933

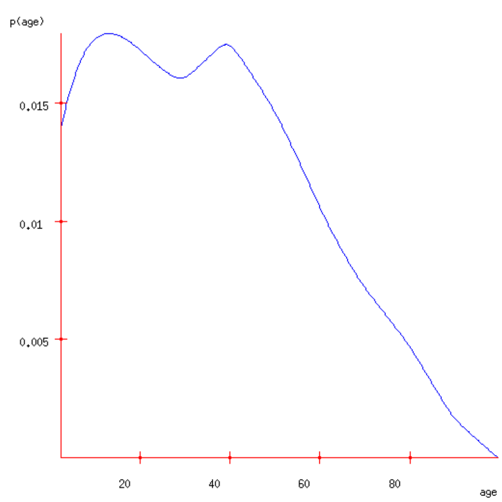
$$P(E_1 \mid E_2) = \frac{P(E_1 \wedge E_2)}{P(E_2)} = \frac{\sum_{\text{rows matching } E_1 \text{ and } E_2} P(\text{row})}{\sum_{\text{rows matching } E_2} P(\text{row})}$$

P(Male | Poor) = 0.4654 / 0.7604 = 0.612

Continuous probability distributions

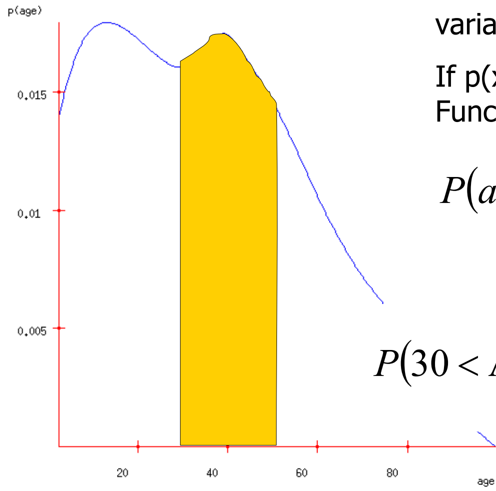
- probability density function (pdf)
- joint probability density
- marginal probability
- calculating probabilities using the pdf
- Bayes' rule

A PDF of American Ages in 2000



more of Andrew's nice slides

A PDF of American Ages in 2000



Let X be a continuous random variable.

If p(x) is a Probability Density Function for X then…

$$P(a < X \le b) = \int_{x=a}^{b} p(x)dx$$

$$P(30 < \text{Age} \le 50) = \int_{\text{age}=30}^{50} p(\text{age})d\text{age}$$

= 0.36

What does p(x) mean?

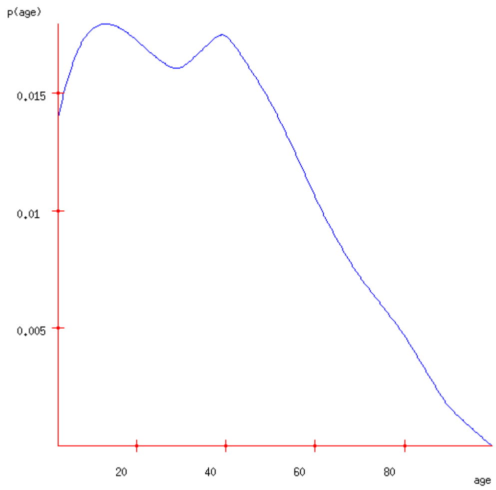
- It does *not* mean a probability!
- First of all, it's not a value between 0 and 1.
- It's just a value, and an arbitrary one at that.
- The likelihood of p(a) can only be compared *relatively* to other values p(b)
- It indicates the relative probability of the integrated density over a small delta:

If $$\frac{p(a)}{p(b)} = \alpha$$

then

$$\lim_{h \to 0} \frac{P(a-h < X < a+h)}{P(b-h < X < b+h)} = \alpha$$

Expectations

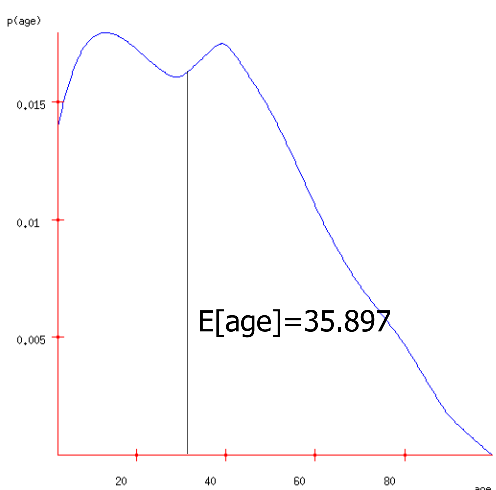


E[X] = the expected value of random variable X

= the average value we'd see if we took a very large number of random samples of X

$$= \int_{x=-\infty}^{\infty} x\, p(x)\, dx$$

Expectations



E[X] = the expected value of random variable X

= the average value we'd see if we took a very large number of random samples of X

$$= \int_{x=-\infty}^{\infty} x\, p(x)\, dx$$

= the first moment of the shape formed by the axes and the blue curve

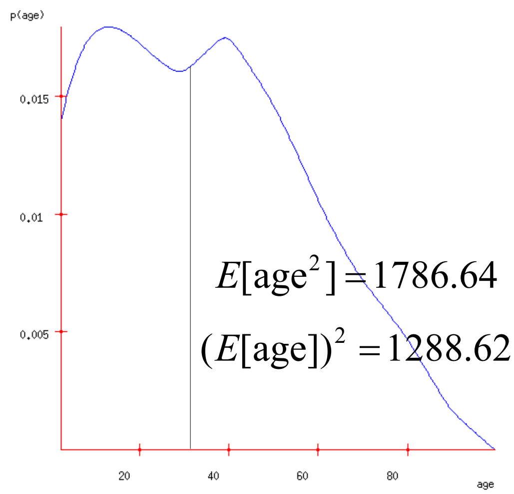
= the best value to choose if you must guess an unknown person's age and you'll be fined the square of your error

Expectation of a function

μ=E[f(X)] = the expected value of f(x) where x is drawn from X's distribution.

= the average value we'd see if we took a very large number of random samples of f(X)

$$\mu = \int_{x=-\infty}^{\infty} f(x)\, p(x)\, dx$$



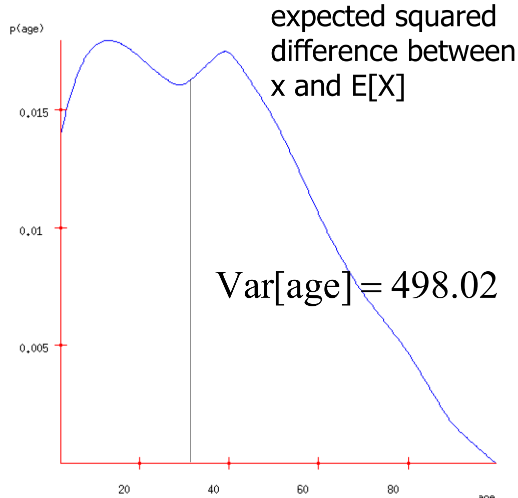
Note that in general:

$$E[f(x)] \neq f(E[X])$$

Variance

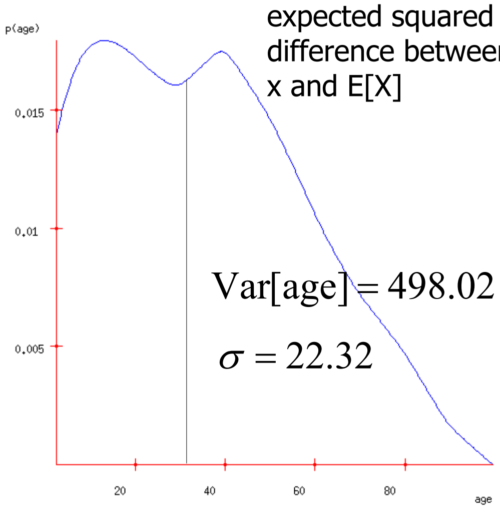
σ^2 = Var[X] = the expected squared difference between x and E[X]

$$\sigma^2 = \int_{x=-\infty}^{\infty} (x-\mu)^2\, p(x)\, dx$$



= amount you'd expect to lose if you must guess an unknown person's age and you'll be fined the square of your error, and assuming you play optimally

Standard Deviation



σ^2 = Var[X] = the expected squared difference between x and E[X]

$$\mathrm{Var[age]} = 498.02$$

$$\sigma = 22.32$$

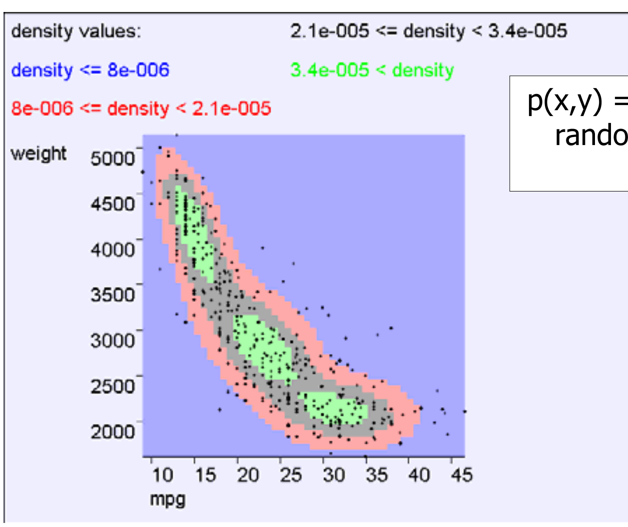
$$\sigma^2 = \int_{x=-\infty}^{\infty} (x-\mu)^2 \, p(x)\, dx$$

= amount you'd expect to lose if you must guess an unknown person's age and you'll be fined the square of your error, and assuming you play optimally

σ = Standard Deviation = "typical" deviation of X from its mean

$$\sigma = \sqrt{\mathrm{Var}[X]}$$

In 2 dimensions

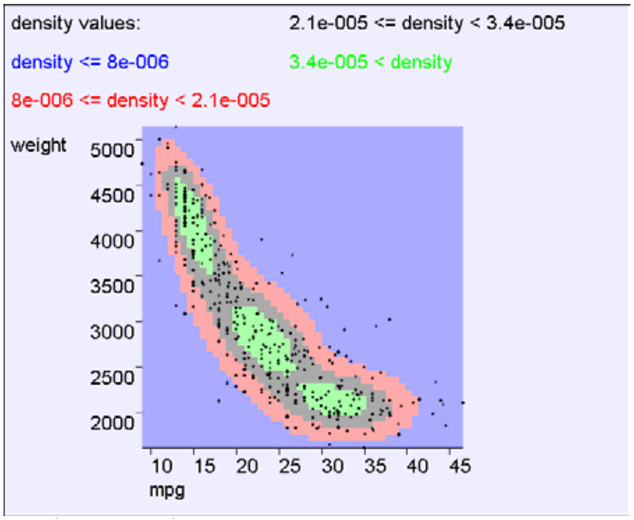


p(x,y) = probability density of random variables (X,Y) at location (x,y)

In 2 dimensions

Let *X,Y* be a pair of continuous random variables, and let R be some region of (X,Y) space…

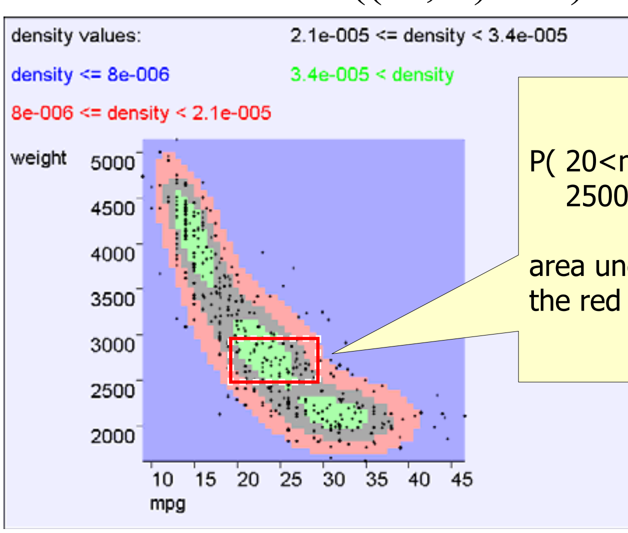
$$P((X,Y) \in R) = \iint_{(x,y)\in R} p(x,y)\,dy\,dx$$



In 2 dimensions

Let *X,Y* be a pair of continuous random variables, and let R be some region of (X,Y) space…

$$P((X,Y) \in R) = \iint_{(x,y)\in R} p(x,y)\,dy\,dx$$



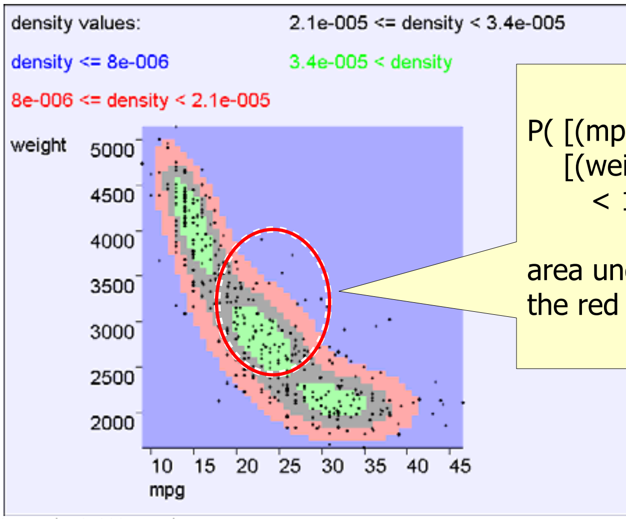
P(20<mpg<30 and 2500<weight<3000) =

area under the 2-d surface within the red rectangle

In 2 dimensions

Let X, Y be a pair of continuous random variables, and let R be some region of (X,Y) space…

$$P((X,Y)\in R) = \iint_{(x,y)\in R} p(x,y)\,dy\,dx$$



P($[(\text{mpg}-25)/10]^2$ + $[(\text{weight}-3300)/1500]^2$ < 1) =

area under the 2-d surface within the red oval

In 2 dimensions

Let X, Y be a pair of continuous random variables, and let R be some region of (X,Y) space…

$$P((X,Y)\in R) = \iint_{(x,y)\in R} p(x,y)\,dy\,dx$$

Take the special case of region R = "everywhere".

Remember that with probability 1, (X,Y) will be drawn from "somewhere".

So..

$$\int_{x=-\infty}^{\infty}\int_{y=-\infty}^{\infty} p(x,y)\,dy\,dx = 1$$

In m dimensions

Let $(X_1, X_2, \ldots X_m)$ be an *n*-tuple of continuous random variables, and let R be some region of $\mathbf{R}^m$ …

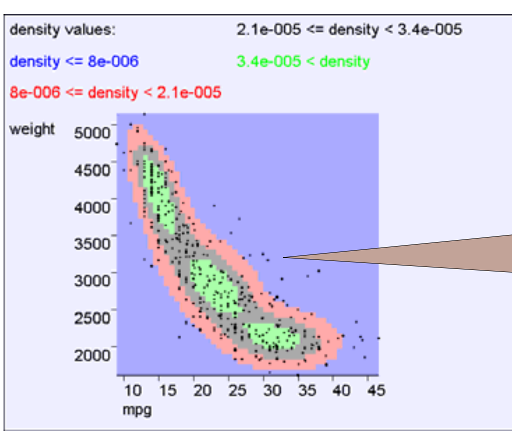
$$P((X_1, X_2, \ldots, X_m) \in R) =$$

$$\iint_{(x_1, x_2, \ldots, x_m) \in R} \ldots \int p(x_1, x_2, \ldots, x_m) dx_m, \ldots dx_2, dx_1$$

Independence

$$X \perp Y \text{ iff } \forall \text{x, y} : p(x, y) = p(x)p(y)$$

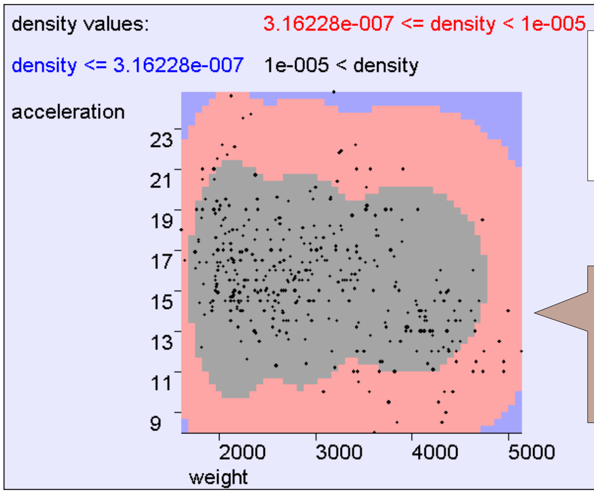
If X and Y are independent then knowing the value of X does not help predict the value of Y



mpg,weight NOT independent

Independence

$$X \perp Y \text{ iff } \forall x, y : p(x,y) = p(x)p(y)$$

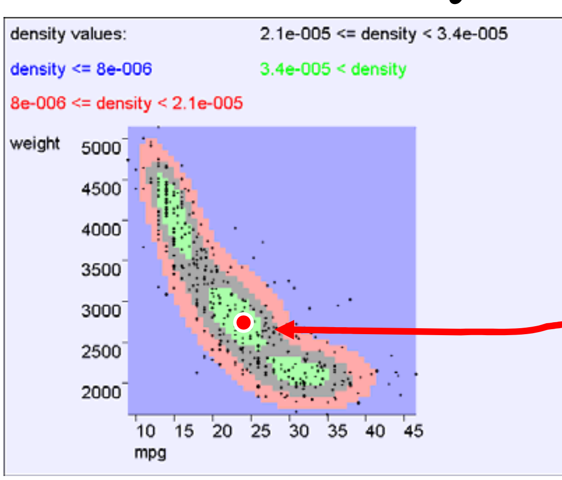


If X and Y are independent then knowing the value of X does not help predict the value of Y

the contours say that acceleration and weight are independent

Multivariate Expectation

$$\boldsymbol{\mu}_{\mathbf{X}} = E[\mathbf{X}] = \int \mathbf{x}\, p(\mathbf{x}) d\mathbf{x}$$



E[mpg,weight] = (24.5,2600)

The centroid of the cloud

Multivariate Expectation

$$E[f(\mathbf{X})] = \int f(\mathbf{x})\; p(\mathbf{x}) d\mathbf{x}$$

Test your understanding

Question : When (if ever) does $E[X+Y] = E[X] + E[Y]$?

- All the time?
- Only when X and Y are independent?
- It can fail even if X and Y are independent?

Bivariate Expectation

$$E[f(x,y)] = \int f(x,y)\ p(x,y)dydx$$

$$\text{if } f(x,y) = x \text{ then } E[f(X,Y)] = \int x\ p(x,y)dydx$$

$$\text{if } f(x,y) = y \text{ then } E[f(X,Y)] = \int y\ p(x,y)dydx$$

$$\text{if } f(x,y) = x + y \text{ then } E[f(X,Y)] = \int (x+y)\ p(x,y)dydx$$

$$E[X+Y] = E[X] + E[Y]$$

Bivariate Covariance

$$\sigma_{xy} = \text{Cov}[X,Y] = E[(X-\mu_x)(Y-\mu_y)]$$

$$\sigma_{xx} = \sigma^2{}_x = \text{Cov}[X,X] = Var[X] = E[(X-\mu_x)^2]$$

$$\sigma_{yy} = \sigma^2{}_y = \text{Cov}[Y,Y] = Var[Y] = E[(Y-\mu_y)^2]$$

Bivariate Covariance

$$\sigma_{xy} = \text{Cov}[X,Y] = E[(X-\mu_x)(Y-\mu_y)]$$

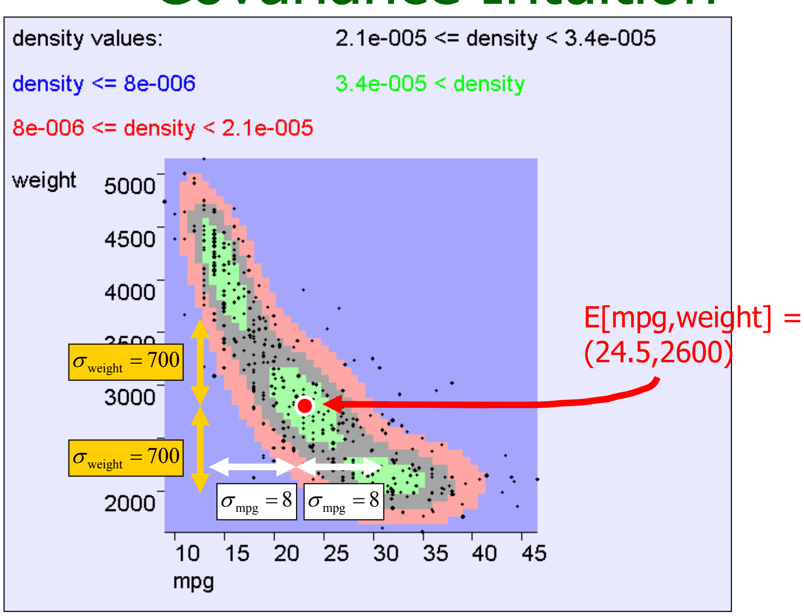
$$\sigma_{xx} = \sigma^2{}_x = \text{Cov}[X,X] = Var[X] = E[(X-\mu_x)^2]$$

$$\sigma_{yy} = \sigma^2{}_y = \text{Cov}[Y,Y] = Var[Y] = E[(Y-\mu_y)^2]$$

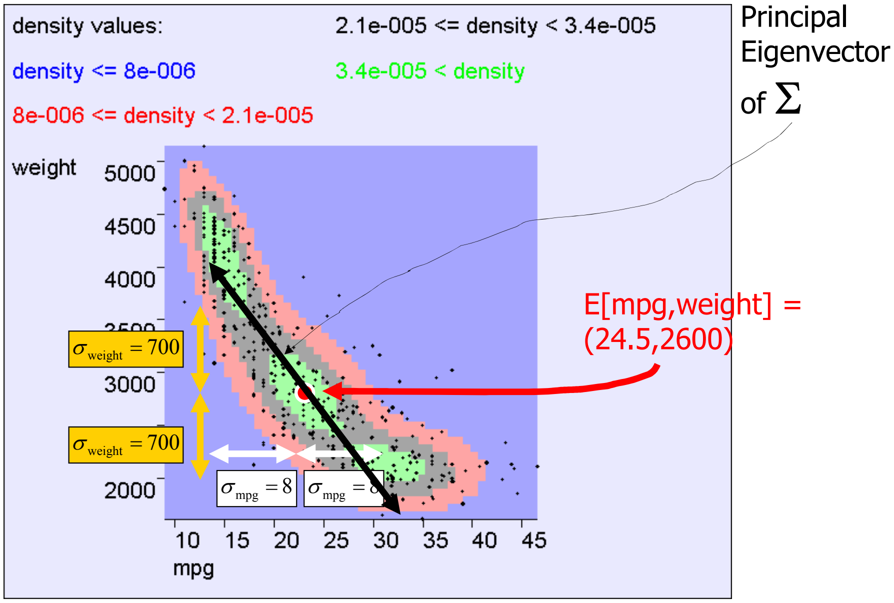
$$\text{Write } \mathbf{X} = \begin{pmatrix} X \\ Y \end{pmatrix}, \text{ then}$$

$$\mathbf{Cov}[\mathbf{X}] = E[(\mathbf{X}-\boldsymbol{\mu}_x)(\mathbf{X}-\boldsymbol{\mu}_x)^T] = \boldsymbol{\Sigma} = \begin{pmatrix} \sigma^2{}_x & \sigma_{xy} \\ \sigma_{xy} & \sigma^2{}_y \end{pmatrix}$$

Covariance Intuition



Covariance Intuition



Covariance Fun Facts

$$\mathbf{Cov}[\mathbf{X}] = E[(\mathbf{X}-\boldsymbol{\mu}_x)(\mathbf{X}-\boldsymbol{\mu}_x)^T] = \boldsymbol{\Sigma} = \begin{pmatrix} \sigma^2{}_x & \sigma_{xy} \\ \sigma_{xy} & \sigma^2{}_y \end{pmatrix}$$

- True or False: If $\sigma_{xy} = 0$ then X and Y are independent
- True or False: If X and Y are independent then $\sigma_{xy} = 0$
- True or False: If $\sigma_{xy} = \sigma_x \sigma_y$ then X and Y are deterministically related
- True or False: If X and Y are deterministically related then $\sigma_{xy} = \sigma_x \sigma_y$

How could you prove or disprove these?

General Covariance

Let $\boldsymbol{X} = (X_1, X_2, \ldots X_k)$ be a vector of k continuous random variables

$$\mathbf{Cov}[\ \mathbf{X}] = E[(\mathbf{X}-\boldsymbol{\mu}_x)(\mathbf{X}-\boldsymbol{\mu}_x)^T] = \boldsymbol{\Sigma}$$

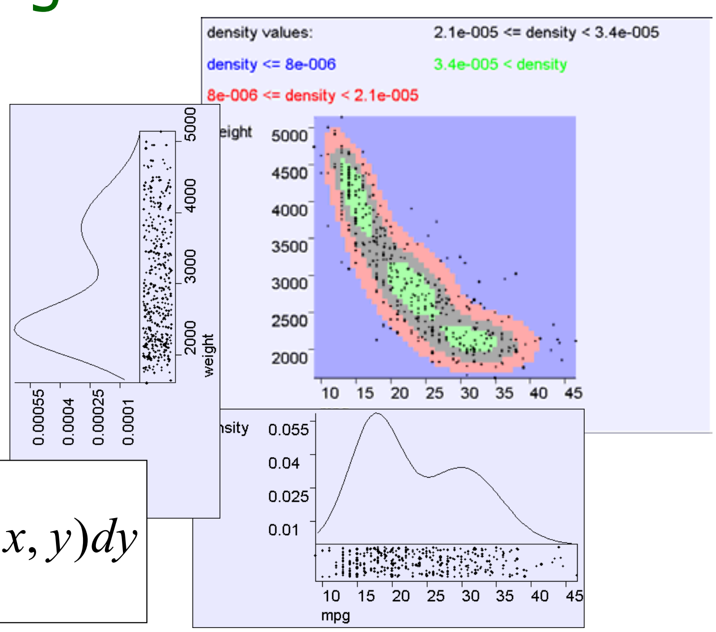
$$\boldsymbol{\Sigma}_{ij} = Cov[X_i, X_j] = \sigma_{x_i x_j}$$

S is a k x k symmetric non-negative definite matrix

If all distributions are linearly independent it is positive definite

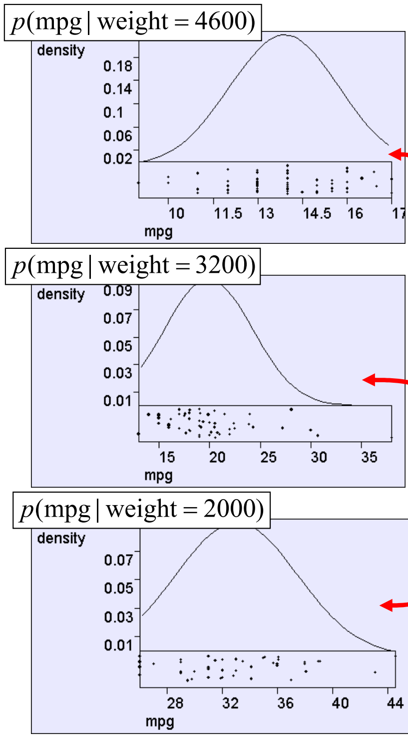
If the distributions are linearly dependent it has determinant zero

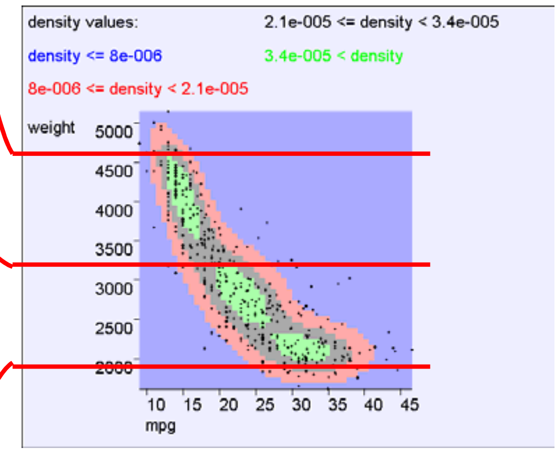
Marginal Distributions



$$p(x) = \int_{y=-\infty}^{\infty} p(x,y)\,dy$$

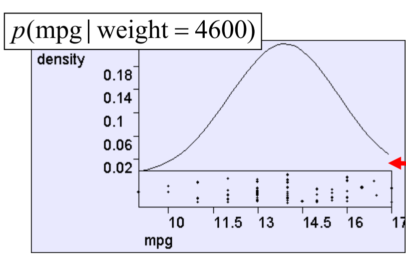
Conditional Distributions





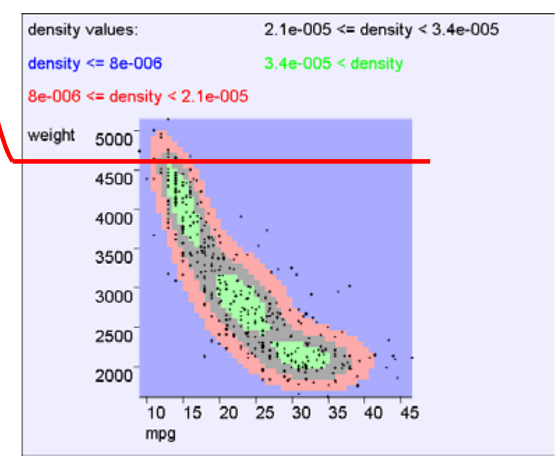
$$p(x \mid y) =$$
p.d.f. of X when $Y = y$

Conditional Distributions



$$p(x \mid y) = \frac{p(x, y)}{p(y)}$$

Why?



$$p(x \mid y) =$$
p.d.f. of X when $Y = y$

Independence Revisited

$$X \perp Y \text{ iff } \forall \mathrm{x}, \mathrm{y}: p(x, y) = p(x)p(y)$$

It's easy to prove that these statements are equivalent...

$$\forall \mathrm{x}, \mathrm{y}: p(x, y) = p(x)p(y)$$

$$\Leftrightarrow$$

$$\forall \mathrm{x}, \mathrm{y}: p(x \mid y) = p(x)$$

$$\Leftrightarrow$$

$$\forall \mathrm{x}, \mathrm{y}: p(y \mid x) = p(y)$$

More useful stuff

$$\int_{x=-\infty}^{\infty} p(x \mid y) dx = 1$$

(These can all be proved from definitions on previous slides)

$$p(x \mid y, z) = \frac{p(x, y \mid z)}{p(y \mid z)}$$

$$p(x \mid y) = \frac{p(y \mid x) p(x)}{p(y)}$$

Bayes Rule

Next time: The process of probabilistic inference

1. *define* model of problem
2. *derive* posterior distributions and estimators
3. *estimate* parameters from data
4. *evaluate* model accuracy