CS395T Computational Statistics with Application to Bioinformatics
and
CAM 383M Statistical and Discrete Methods for Scientific Computing

Prof. William H. Press
Spring Term, 2011
The University of Texas at Austin

Lecture 1

Why two course numbers and titles?

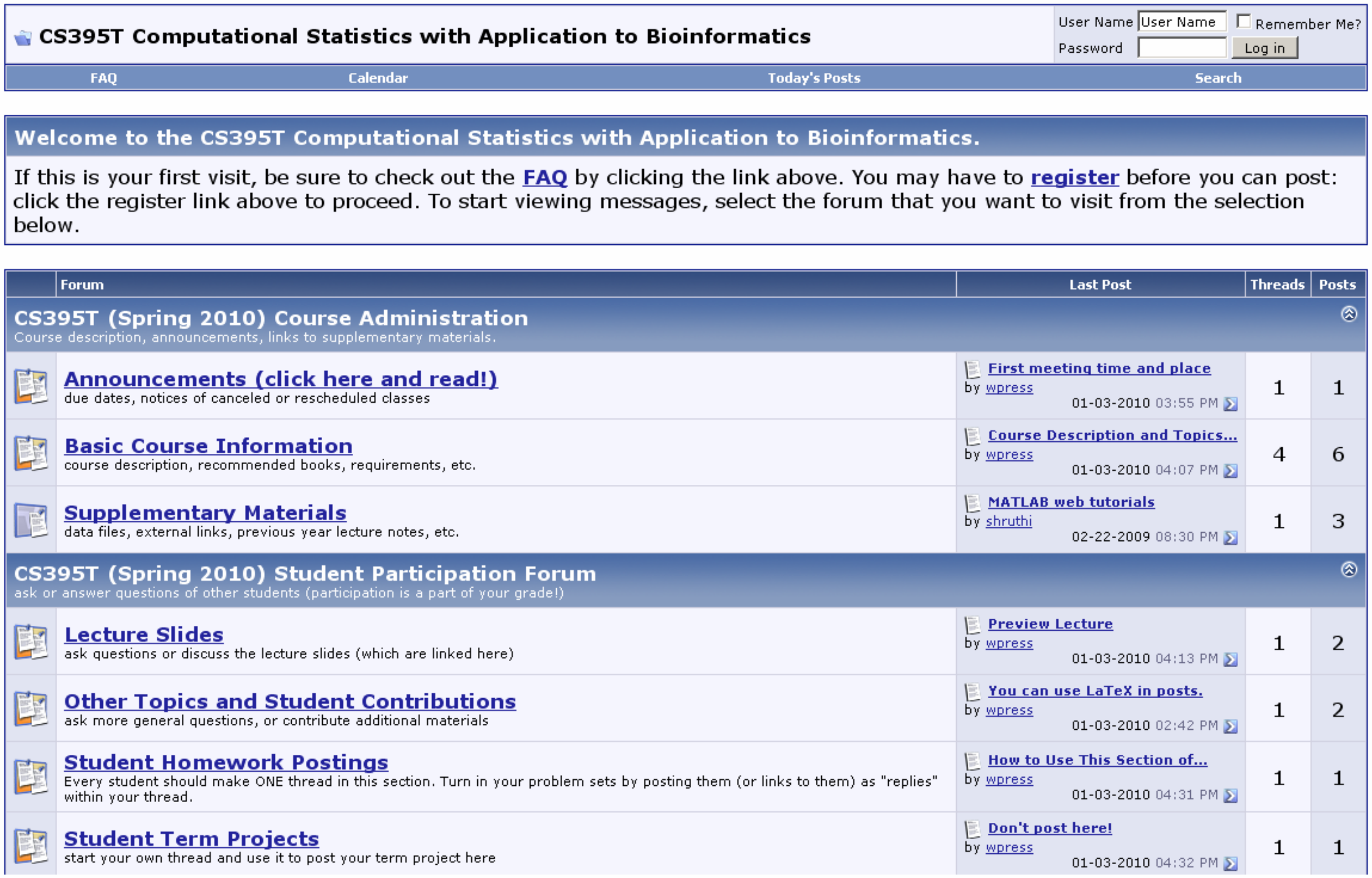
- CS395T was formerly cross-listed as CAM395T
  - "Computational Statistics with Application to Bioinformatics"
  - same course for CS and CSEM (CAM) students
- CSEM graduate program realigning (and generally broadening) its course requirements
  - need for a course covering statistical methods
  - also need to introduce some basic CS structures/algorithms to students whose background is applied math
  - "Statistical and Discrete Methods for Scientific Computing"
- So, this year, a blended course
  - a couple of different problem sets
  - maybe a couple of different lectures
  - CS students are often weak on analysis (= calculus year 3+)
- Next year (Spring, 2012) will likely be taught as the CAM course

What is Computational Statistics, anyway?

- It’s not different (mathematically) from “regular” statistics.
- Has less distinction between “statisticians” and “users of statistics”
  - since users have access to lots of computing power
- Heavy use of simulation (e.g., Monte Carlo) and resampling techniques
  - instead of analytic calculation of distributions of the null hypothesis
- Somewhat more Bayesian, but not exclusively so
- Somewhat more driven by specifics of unique data sets
  - this can be dangerous (“shopping for significance”)
  - or powerful!
- Closely related to machine learning, but a somewhat different culture

http://wpressutexas.net/forum

- Course web forum is the central hub of the course
  - you should register using same email as on sign-up sheet
  - start threads or add comments under Lecture Slides or Other Topics
  - add comments to Course Administration topics



CS395T Computational Statistics with Application to Bioinformatics

Welcome to the CS395T Computational Statistics with Application to Bioinformatics.

If this is your first visit, be sure to check out the FAQ by clicking the link above. You may have to register before you can post: click the register link above to proceed. To start viewing messages, select the forum that you want to visit from the selection below.

Forum	Last Post	Threads	Posts
CS395T (Spring 2010) Course Administration Course description, announcements, links to supplementary materials.			
Announcements (click here and read!) due dates, notices of canceled or rescheduled classes	First meeting time and place by wpress 01-03-2010 03:55 PM	1	1
Basic Course Information course description, recommended books, requirements, etc.	Course Description and Topics... by wpress 01-03-2010 04:07 PM	4	6
Supplementary Materials data files, external links, previous year lecture notes, etc.	MATLAB web tutorials by shruthi 02-22-2009 08:30 PM	1	3
CS395T (Spring 2010) Student Participation Forum ask or answer questions of other students (participation is a part of your grade!)			
Lecture Slides ask questions or discuss the lecture slides (which are linked here)	Preview Lecture by wpress 01-03-2010 04:13 PM	1	2
Other Topics and Student Contributions ask more general questions, or contribute additional materials	You can use LaTeX in posts. by wpress 01-03-2010 02:42 PM	1	2
Student Homework Postings Every student should make ONE thread in this section. Turn in your problem sets by posting them (or links to them) as "replies" within your thread.	How to Use This Section of... by wpress 01-03-2010 04:31 PM	1	1
Student Term Projects start your own thread and use it to post your term project here	Don't post here! by wpress 01-03-2010 04:32 PM	1	1

What should you learn in this course?

- A lot of conventional statistics at a 1st year graduate level
  - mostly by practical example, not proving theorems
  - but you should also learn to read the statistics and/or machine learning and/or pattern recognition textbook literature
- A lot about real, as opposed to idealized, data sets
  - we'll supply and discuss some
  - you can also use and/or share your own
- A bunch of important computational algorithms
  - often stochastic
- Some bioinformatics, especially genomics
  - although that is not the main point of the course
- Some programming methodology and languages
  - computer with MATLAB or Octave (free) is <u>required</u>
    - MATLAB Student Version at computer store in Flawn Academic Center is a bargain at $100. (Permanent license will install on 2 machines, I think.)
  - you'll get at least a reading knowledge of Mathematica
  - a bit of data parallel methods, notated in MATLAB but more general in concept
  - we won't use R (or S), which would make you a "real" statistician

Laws of Probability

"There is this thing called *probability*. It obeys the laws of an axiomatic system. When identified with the real world, it gives (partial) information about the future."

- What axiomatic system?
- How to identify to real world?
  - Bayesian or frequentist viewpoints are somewhat different "mappings" from axiomatic probability theory to the real world
  - yet both are useful

"And, it gives a consistent and complete calculus of inference."

- This is only a Bayesian viewpoint
  - It's sort of true and sort of not true, as we will see!
- R.T. Cox (1946) showed that reasonable assumptions about "degree of belief" uniquely imply the axioms of probability (and Bayes)
  - belief in a proposition's negation increases as belief in the proposition decreases
  - "composable" (belief in AB depends only on A and B|A)
  - belief in a proposition independent of the order in which supporting evidence is adduced (path-independence of belief)

Axioms:

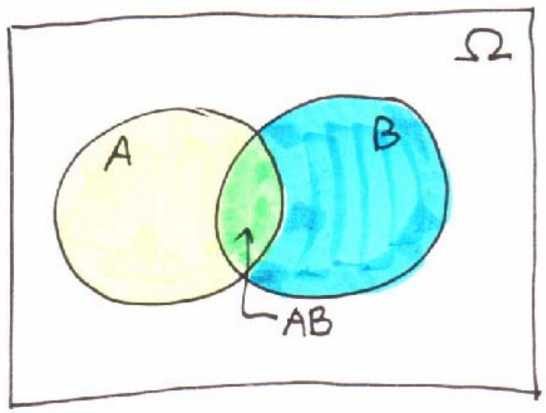
I. $P(A) \geq 0$ for an event A
II. $P(\Omega) = 1$ where Ω is the set of all possible outcomes
III. if $A \cap B = \emptyset$, then $P(A \cup B) = P(A) + P(B)$

Example of a theorem:

Theorem: $P(\emptyset) = 0$
Proof: $A \cap \emptyset = \emptyset$, so
$P(A) = P(A \cup \emptyset) = P(A) + P(\emptyset)$, q.e.d.

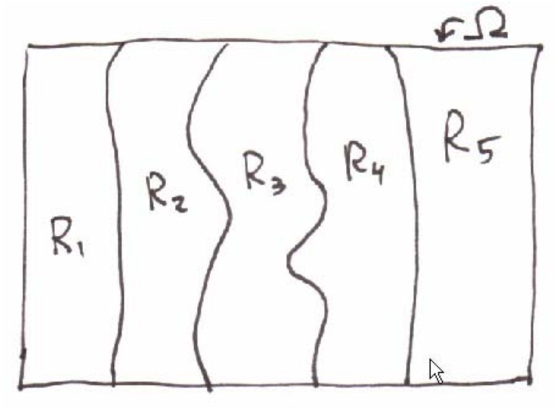
Basically this is a theory of measure on Venn diagrams, so we can (informally) cheat and prove theorems by inspection of the appropriate diagrams, as we now do.

Additivity or “Law of Or-ing”



$$P(A \cup B) = P(A) + P(B) - P(AB)$$

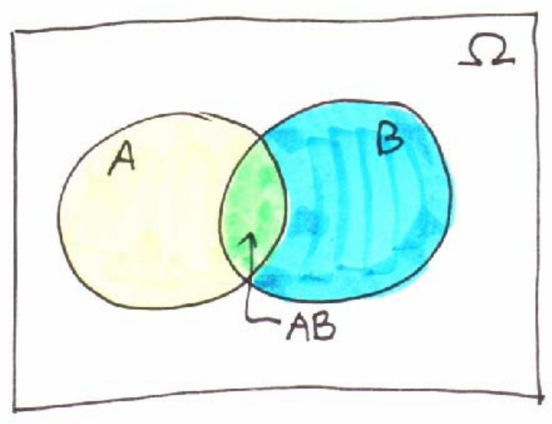
“Law of Exhaustion”



If R_i are exhaustive and mutually exclusive (EME)

$$\sum_i P(R_i) = 1$$

Multiplicative Rule or “Law of And-ing”



(same picture as before)

“given”

$$P(AB) = P(A)P(B|A) = P(B)P(A|B)$$

$$P(B|A) = \frac{P(AB)}{P(A)}$$

“conditional probability”

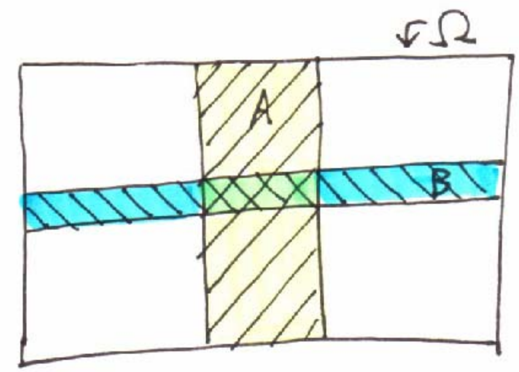
“renormalize the outcome space”

Similarly, for multiple And-ing:

$$P(ABC) = P(A)P(B|A)P(C|AB)$$

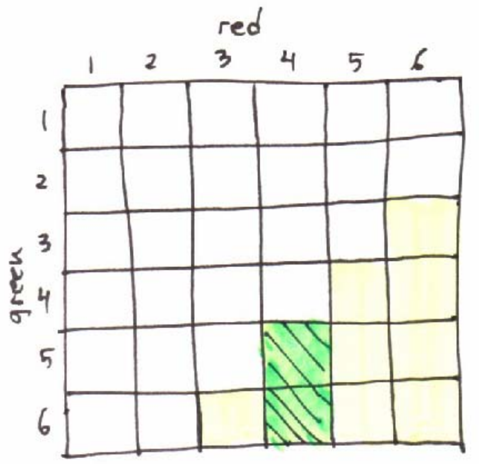
Independence:

Events A and B are independent if
$P(A|B) = P(A)$
so $P(AB) = P(B)P(A|B) = P(A)P(B)$



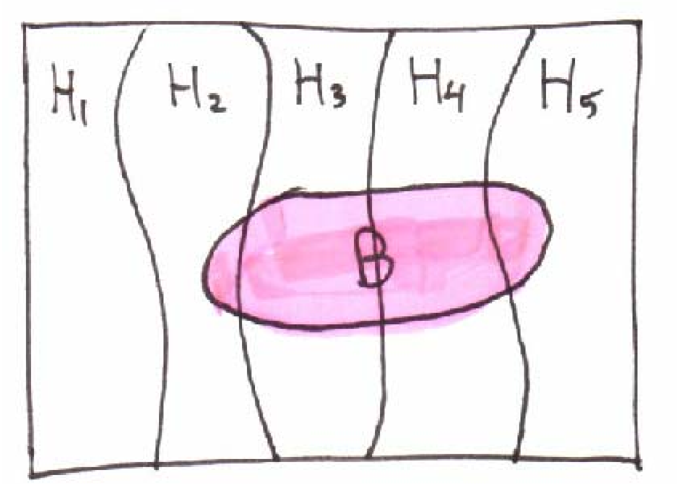
A symmetric die has
$P(1) = P(2) = \ldots = P(6) = \frac{1}{6}$
Why? Because $\sum_i P(i) = 1$ and $P(i) = P(j)$.
Not because of “frequency of occurence in N trials”.
That comes later!

The sum of faces of two dice (red and green) is > 8.
What is the probability that the red face is 4?



$$P(R4\,|\,{>}8) = \frac{P(R4 \cap {>}8)}{P({>}8)} = \frac{2/36}{10/36} = 0.2$$

Law of Total Probability or “Law of de-Anding”



H’s are exhaustive and mutually exclusive (EME)

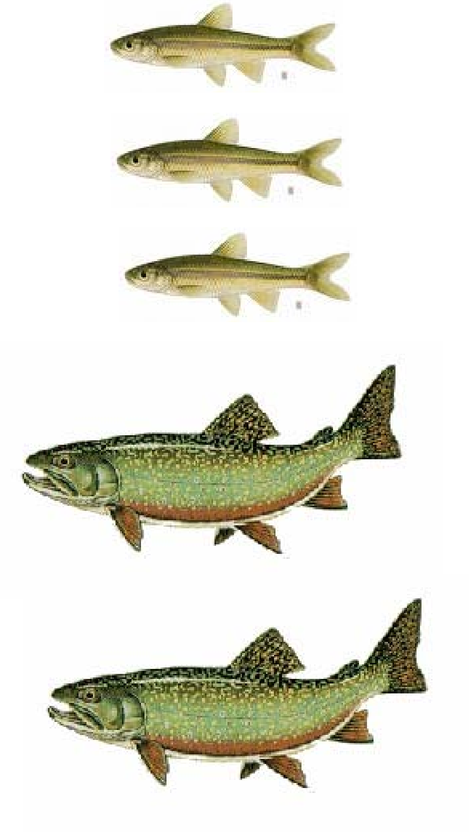
$$P(B) = P(BH_1) + P(BH_2) + \ldots = \sum_i P(BH_i)$$

$$P(B) = \sum_i P(B|H_i)P(H_i)$$

“How to put Humpty-Dumpty back together again.”

Example: A barrel has 3 minnows and 2 trout, with equal probability of being caught. Minnows must be thrown back. Trout we keep.

What is the probability that the 2nd fish caught is a trout?

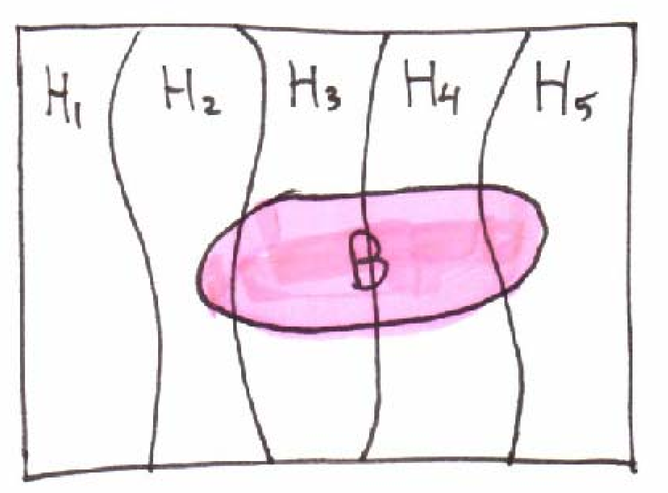


$$\begin{aligned} H_1 &\equiv \text{1st caught is minnow, leaving } 3+2 \\ H_2 &\equiv \text{1st caught is trout, leaving } 3+1 \\ B &\equiv \text{2nd caught is a trout} \end{aligned}$$

$$\begin{aligned} P(B) &= P(B|H_1)P(H_1) + P(B|H_2)P(H_2) \\ &= \tfrac{2}{5} \cdot \tfrac{3}{5} + \tfrac{1}{4} \cdot \tfrac{2}{5} = 0.34 \end{aligned}$$

Course preview question: About how many times would you have to do this experiment to distinguish the true value from a claim that P=1/3 ?

Bayes Theorem



Thomas Bayes
1702 - 1761

(same picture as before)

$$P(H_i|B) = \frac{P(H_iB)}{P(B)}$$

$$= \frac{P(B|H_i)P(H_i)}{\sum_j P(B|H_j)P(H_j)}$$

Law of And-ing

Law of de-Anding

We usually write this as

$$P(H_i|B) \propto P(B|H_i)P(H_i)$$

this means, “compute the normalization by using the completeness of the H_i’s”

- As a theorem relating probabilities, Bayes is unassailable
- But we will also use it in inference, where the H's are hypotheses, while B is the data
  - "what is the probability of an hypothesis, given the data?"
  - some (defined as frequentists) consider this dodgy
  - others (Bayesians like us) consider this fantastically powerful and useful
  - in real life, the "war" between Bayesians and frequentists is long since over, and most statisticians adopt a mixture of techniques appropriate to the problem
    - for a view of the "war", see Efron paper on the forum
- Note that you generally have to know a complete set of EME hypotheses to use Bayes for inference
  - perhaps its principal weakness

Let’s work a couple of examples using Bayes Law:

Example: Trolls Under the Bridge

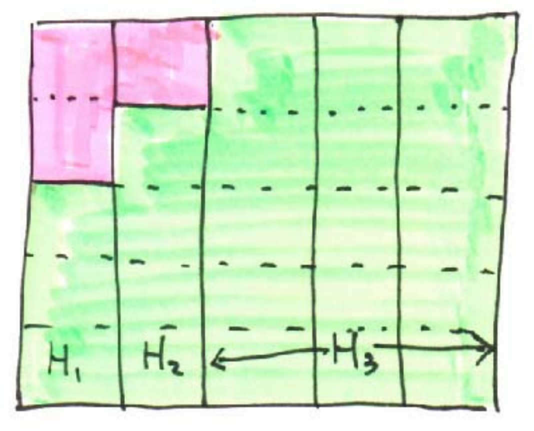
Trolls are bad. Gnomes are benign.
Every bridge has 5 creatures under it:

20% have TTGGG (H_1)
20% have TGGGG (H_2)
60% have GGGGG (benign) (H_3)

Before crossing a bridge, a knight captures one of the 5 creatures at random. It is a troll. “I now have an 80% chance of crossing safely,” he reasons, “since only the case

20% had TTGGG (H1) → now have TGGG

is still a threat.”



$$P(H_i|T) \propto P(T|H_i)P(H_i)$$

so, $$P(H_1|T) = \frac{\frac{2}{5}\cdot\frac{1}{5}}{\frac{2}{5}\cdot\frac{1}{5}+\frac{1}{5}\cdot\frac{1}{5}+0\cdot\frac{3}{5}} = \frac{2}{3}$$

The knight’s chance of crossing safely is actually only 33.3%
Before he captured a troll (“saw the data”) it was 60%.
Capturing a troll actually made things worse!
(80% was never the right answer!)

Data changes probabilities!
Probabilities after assimilating data are called <u>posterior probabilities</u>.

Congratulations! You are now a Bayesian.

Bayesian viewpoint:
Probabilities are modified by data. This makes them intrinsically subjective, because different observers have access to different amounts of data (including their "background information" or "background knowledge").

Notice in particular that the connection of probability to "frequency of occurrence of repeated events" is now complicated! (Would have to "repeat" the exact state of knowledge of the observer.)