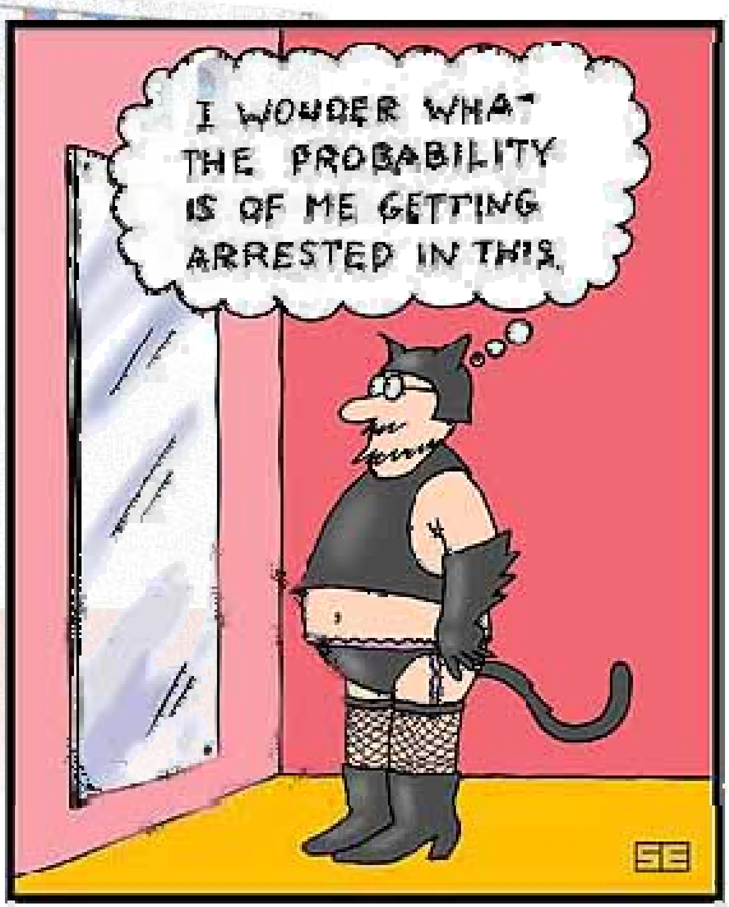
Introduction to Informatics

Lecture 19:

Probability



Schrodinger's cat suit.

Luis M.Rocha and Santiago Schnell

Readings until now

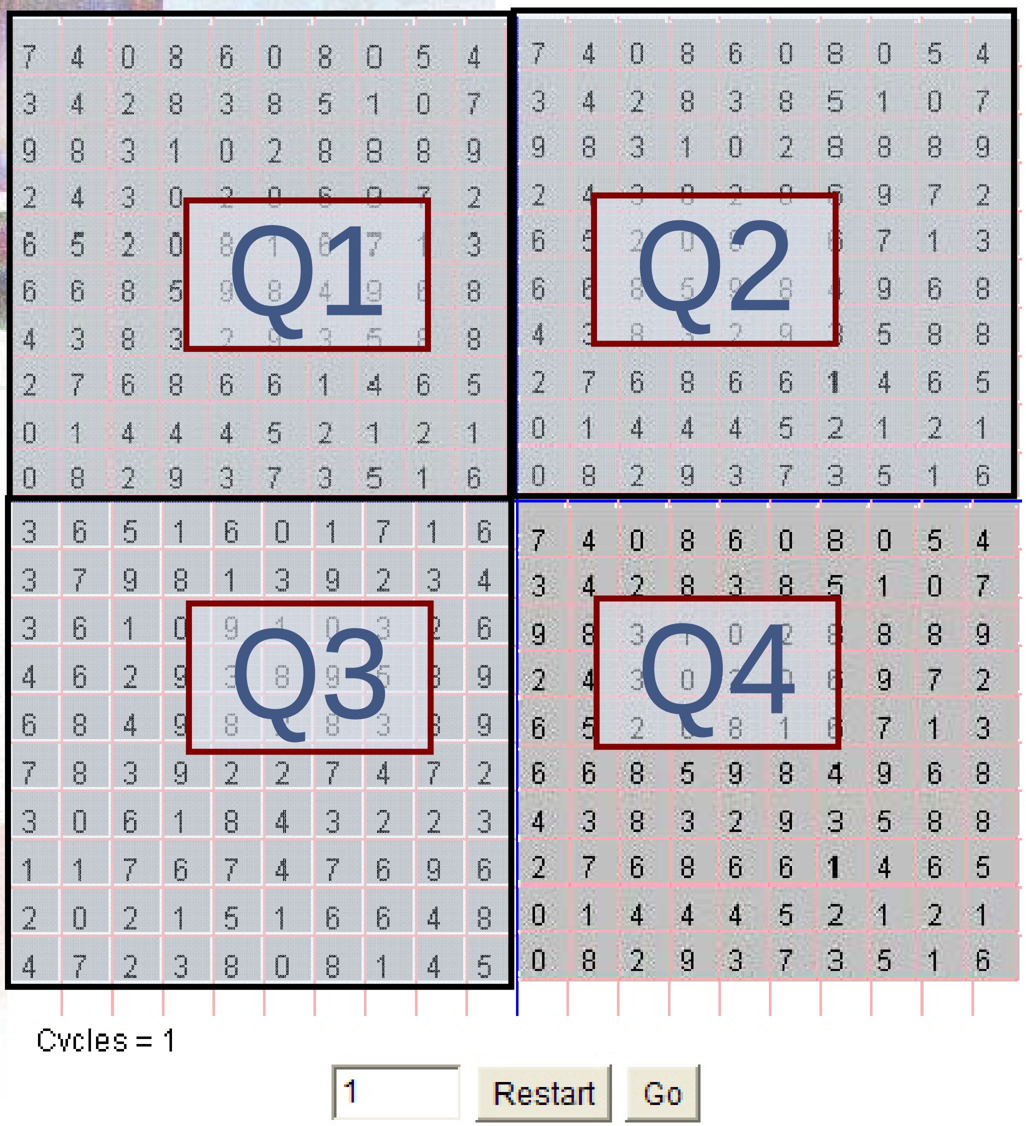
- Lecture notes
  - Posted online
    - http://informatics.indiana.edu/rocha/i101
      - *The Nature of Information*
      - *Technology*
      - *Modeling the World*
  - @ *infoport*
    - *http://infoport.blogspot.com*
  - From course package
    - Von Baeyer, H.C. [2004]. *Information: The New Language of Science*. Harvard University Press.
      - Chapters 1, 4 (pages 1-12)
    - From Andy Clark's book "*Natural-Born Cyborgs*"
      - Chapters 2 and 6 (pages 19 - 67)
    - From Irv Englander's book "*The Architecture of Computer Hardware and Systems Software*"
      - Chapter 3: Data Formats (pp. 70-86)
    - Klir, J.G., U. St. Clair, and B.Yuan [1997]. Fuzzy Set Theory: foundations and Applications. Prentice Hall
      - Chapter 2: Classical Logic (pp. 87-97)
      - Chapter 3: Classical Set Theory (pp. 98-103)
    - Norman, G.R. and D.L. Streinrt [2000]. *Biostatistics: The Bare Essentials*.
      - Chapters 1-3 (pages 105-129)
      - OPTIONAL: Chapter 4 (pages 131-136)
      - Chapter 13 (pages 147-155)
      - Chapter 5 (pages 141-144)

Assignment Situation

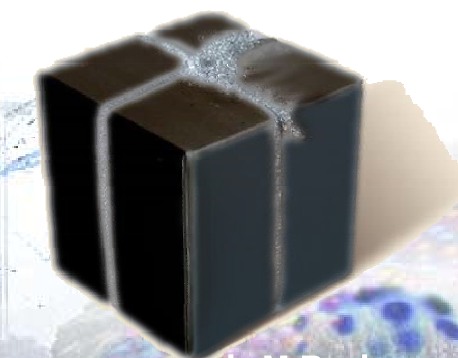
- Labs
  - Past
    - Lab 1: Blogs
      - Closed (Friday, January 19): Grades Posted
    - Lab 2: Basic HTML
      - Closed (Wednesday, January 31): Grades Posted
    - Lab 3: Advanced HTML: Cascading Style Sheets
      - Closed (Friday, February 2): Grades Posted
    - Lab 4: More HTML and CSS
      - Closed (Friday, February 9): Grades Posted
    - Lab 5: Introduction to Operating Systems: Unix
      - Closed (Friday, February 16): Grades Posted
    - Lab 6: More Unix and FTP
      - Closed (Friday, February 23): Grades Posted
    - Lab 7: Logic Gates
      - Closed (Friday, March 9): Grades Posted
    - Lab 8: Intro to Statistical Analysis using Excel
      - Due Friday, March 30
  - Next: Lab 9
    - Data analysis with Excel (linear regression)
      - April 29 and 30, Due Friday, April 6

- Assignments
  - Individual
    - First installment
      - Closed: February 9: Grades Posted
    - Second Installment
      - Past: March 2, Being Grades Posted
    - Third installment
      - Presented on March 8th, Due on March 30th
  - Group
    - First Installment
      - Past: March 9th, Being graded
    - Second Installment
      - March 29; Due Friday, April 6

Individual Assignment – Part III



- Step by step analysis of "dying" squares
  - 3rd Installment
    - Presented: March 8th
    - Due: March 30th
  - 4th Installment
    - Presented: April 5th
    - Due: April 20th
- Use descriptive statistics
  - To uncover rules inductively
    - E.g. the behavior of evens and odds, individual numbers, or ranges of cycles, etc.



Deduction vs. Induction

- Deductive Inference Logic
  - If the premises are true, we have absolute ***certainty*** of the conclusion
- Inductive Inference Uncertainty
  - Conclusion supported by ***good evidence*** (significant number of examples/observations) but not full certainty -- ***likelihood***

In the previous classes...

- We learned a little bit of statistics and Induction!
  - Histograms
  - Measurements of Central Tendency
  - Measurements of Dispersion
  - Regression

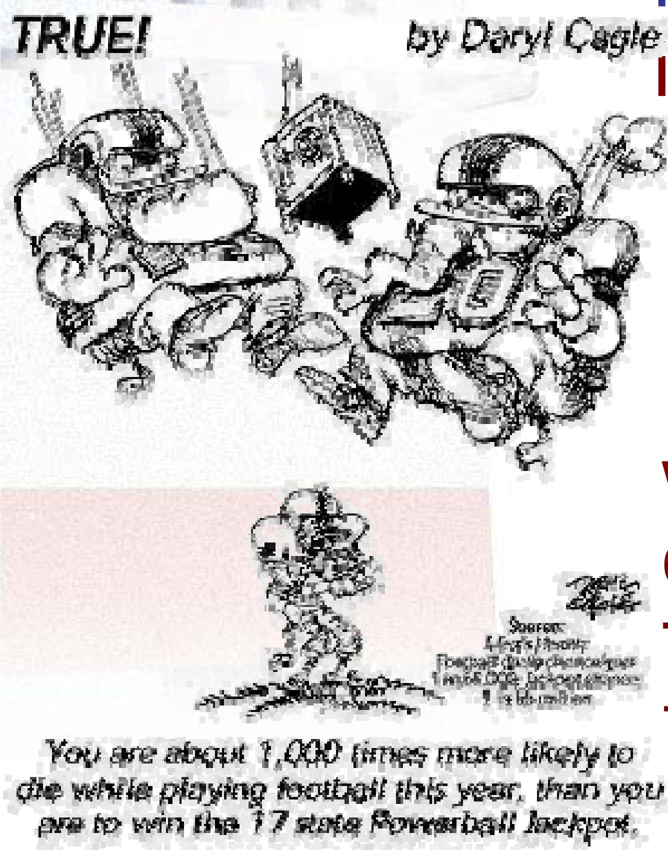
Let's toss a coin!

- Possible outcomes (events):
  - The Queen/Heads (H) or Peter Pan/Tails (T)

- Results of tossing a coin
  - Once: H
  - Twice: H, H
  - *n* times: H, H, T, H, T, T, T,...

How do we assign probability to an event?

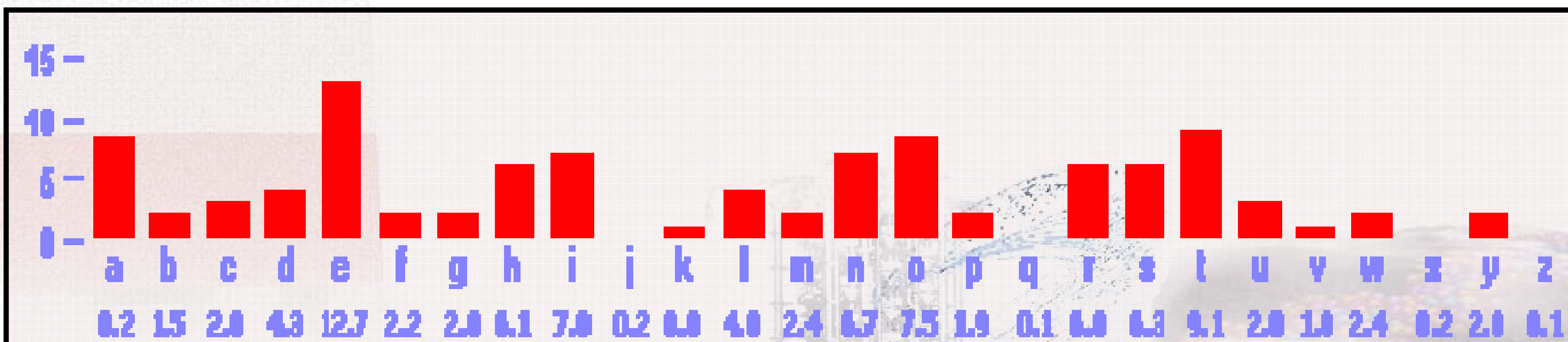
- The probability of an event *A* in an experiment is supposed to measure how ***frequently*** A is about to occur if we make many trials.
  - If we flip our coin many times, H and T will appear *about* equally often – we say that H and T are “equally likely”.



We regard ***probability*** as the counterpart of ***relative frequency***!
The latter is used to estimate the former.

English Letter Frequency

- Six most common letters in English
  - E T A O I N
- Spanish
  - E A O S R N
- Other Languages
  - http://people.bath.ac.uk/tab21/forcrypt.html

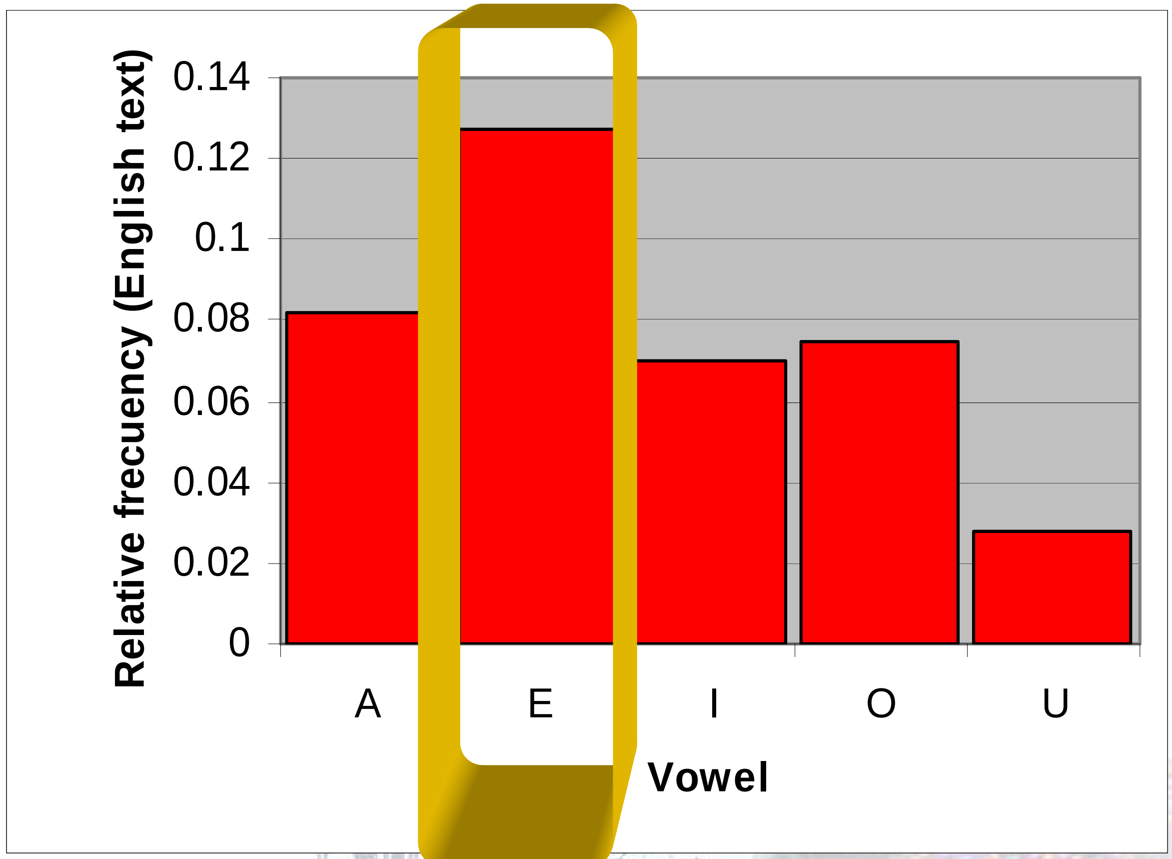


Luis M.Rocha and Santiago Schnell

Relative Frequency-Probability

What is the *estimated* probability of finding the letter "e" in an English text?

Vowel	f_{rel}
A	0.082
E	0.127
I	0.070
O	0.075
U	0.028



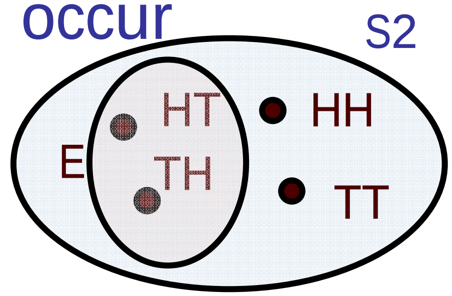
Probability Notions

- Experiment
  - Any activity that yields a result or an outcome
    - Tossing a coin

[Chase and Brown, "General Statistics"]

Probability Notions

- Experiment
  - Any activity that yields a result or an outcome
    - Tossing a coin
- Sample Space
  - The set of all possible outcomes of an experiment.
  - One and only one of the outcomes must occur
    - Flipping one coin: S1 = {H, T}
    - Flipping two coins: S2 = {HH, HT, TH, TT}



- Event
  - Subset of sample space
  - The event occurs if when we perform the experiment one of its elements occurs.
    - Non-match in two coin experiment is an event E = {HT, TH}

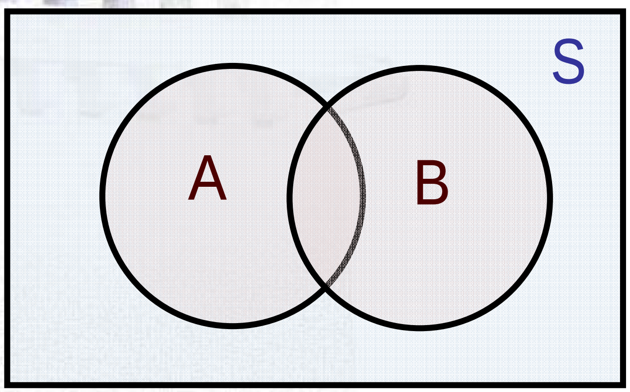
[Chase and Brown, "General Statistics"]

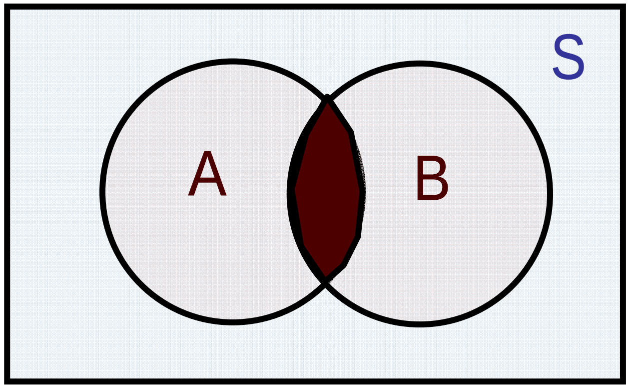
Probability of an Event

- P(A)
  - The expected proportion of occurrences of an event A if the experiment were to be ***repeated many times***
  - **$0 \leq P(A) \leq 1$**
  - $P(A) = P(\{a_1\}) + P(\{a_2\}) + \ldots. + P(\{a_n\})$
    - $A = \{a_1, a_2, \ldots, a_n\}$
      - e.g. dice faces {1, 5}
  - ***Theoretical*** probability: $P(A) = |A|/|S|$ ⇨ $|A|$ = Cardinality of A (number of elements)
    - S1: $P(\{T\}) = ½$ and $P(\{H\}) = ½$
    - S2: P(*nonmatch*)
      - $P(\{HT, TH\}) = P(\{HT\}) + P(\{TH\}) = ¼ + ¼ = ½$
    - A: $P(\{1,5\}) = P(\{1\})+P(\{5\})= 1/6 +1/6 = 1/3$
- Estimated from limited experiments
  - ***Empirical*** probability
    - $\{T,T,H,T\} \Rightarrow P(\{T\}) = 0.75$ and $P(\{H\}) \Rightarrow 0.25$
- Guessed Subjective probability
  - “there is a 90% chance that I will pass this course”

The Addition rule

- If A,B are events from some sample space
  - $P(A \vee B) = P(A) + P(B) - P(A \wedge B)$





- $P(A) = |A|/|S|$
- $P(B) = |B|/|S|$
- $P(A \wedge B) = |A \wedge B|/|S|$
- $P(A \vee B) = |A \vee B|/|S| = (|A| + |B| - | A \wedge B|)/|S|$

Addition Rule example

$P(E1) = |E1|/|S2| = 2/4 = ½$

$P(E2) = |E2|/|S2| = 2/4 = ½$

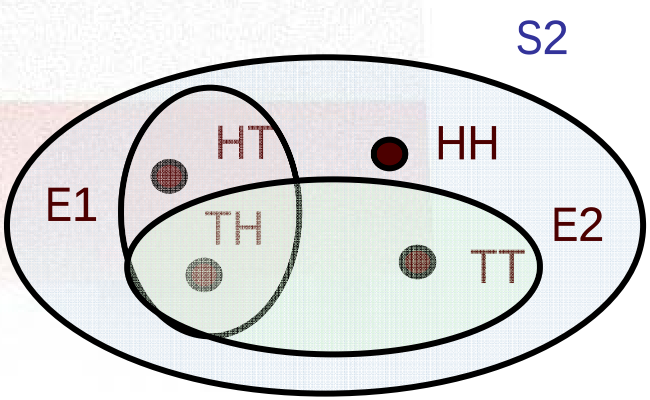
$P(E1 \wedge E2) = |E1 \wedge E2|/|S2| = ¼$

$P(E1 \vee E2) = |E1 \vee E2|/|S2| =$

$= (|E1| + |E2| - | E1 \wedge E2|)/|S2| =$

$= (2 + 2 – 1)/4 = ¾$

$$P(E1 \vee E2) = P(E1) + P(E2) – P(E1 \wedge E2)$$



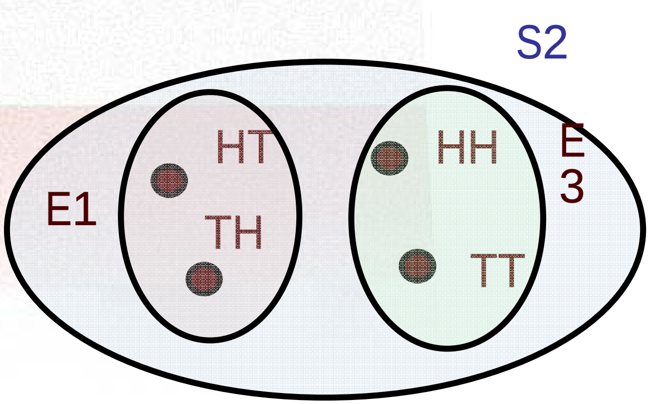
http://www.stat.sc.edu/~west/applets/Venn1.html

Mutually Exclusive Events

- The occurrence of one precludes the occurrence of the other
  - E3=Match and E1=nonmatch in two coin example
  - Addition Rule is just sum of exclusive events

$$P\ (E1 \vee E2) = P(E1) + P(E2) - \cancel{P(E1 \wedge E2)}$$

$$\boxed{P\ (E1 \vee E2) = P(E1) + P(E2)}$$

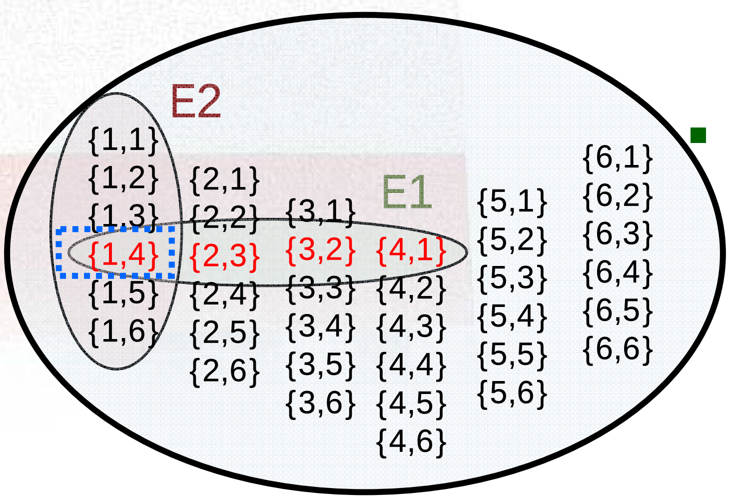


Conditionally Dependent Events: The outcome of one depends on the occurrence of the other

$$\boxed{\mathbf{P(E1 \wedge E2) > 0}}$$

Example of Conditionally dependent events

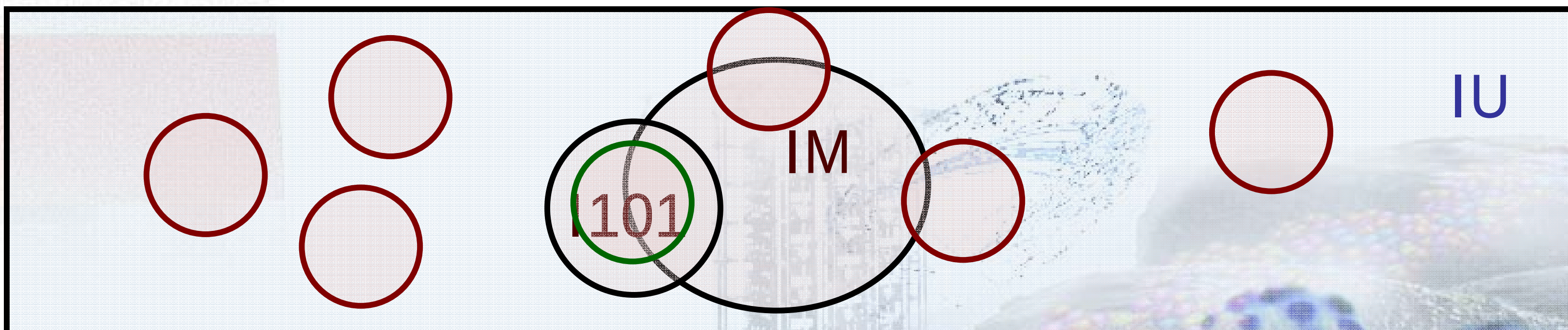
- 2 dice
  - E1= Sum of dice = 5
    - P(E1) = 4/36 = 1/9 =0.1111
      - 4 out 36 possibilities: {1,4}, {2,3}, {3,2}, {4,1}
  - E2 = “first dice is 1”
    - If E2
      - Probability of “5” = P(E1) = 1/6 = 0.1667
        - 1 out of 6 possibilities: {1,1}, {1,2}, {1,3}, {1,4}, {1,5}, {1,6}
  - Probability of E1 is <u>conditional</u> on value of first dice (E2)
    - $P(E1 \wedge E2)>0 \Rightarrow$ Not mutually Exclusive
    - $P(E1|E2) = |E1 \wedge E2|/|E2|=1/6$
      - Probability of E1 <u>given</u> E2



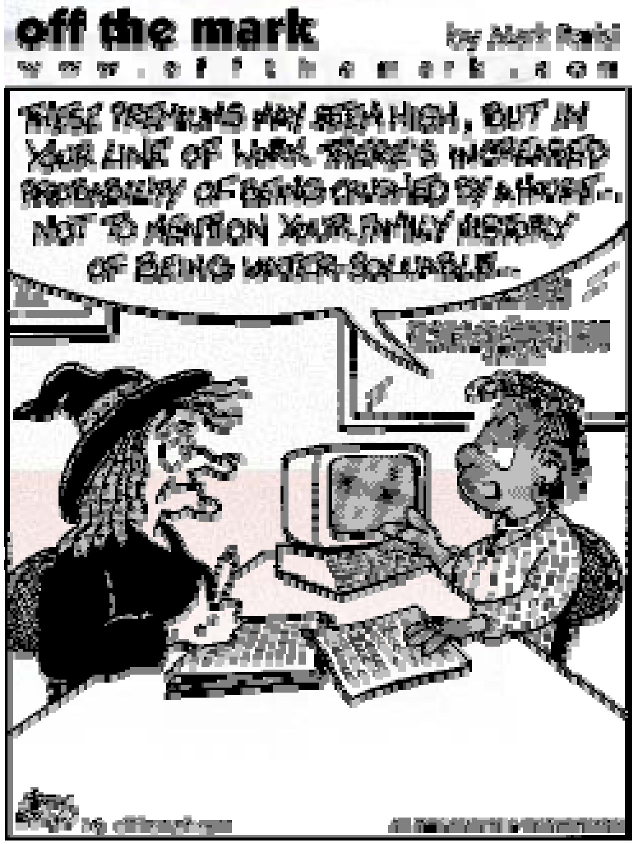
Luis M.Rocha and Santiago Schnell

Conditional Probability

- $\boxed{P(B|A) = |A \wedge B| / |A|}$
  - Probability of a IU student being an Informatics major, *given* that a student is enrolled in I101
    - $|I101| = 110$ students
    - $|IM| = |\{\text{informatics major}\}| = 400$
    - **P(IM|I101)** $= |IM \wedge I101| / |I101| = 55/110 = \boxed{\mathbf{0.5}}$
    - **P(IM)** $= 400/20000 = \boxed{\mathbf{0.02}}$
- Multiplication Rule for ***conditionally probable*** events
  - $P(A \wedge B) = P(A) \, . \, P(B|A)$



Independent Events



- Neither mutually exclusive nor conditionally probable events
- Two events A, B are ***independent*** if the occurrence of one has ***no effect on the probability*** of the occurrence of the other
  - $P(B|A) = P(B)$
  - Multiplication Rule
    - $P(A \wedge B) = P(A) \, . \, P(B|A) = P(A).P(B)$
  - Example
    - Tossing coins

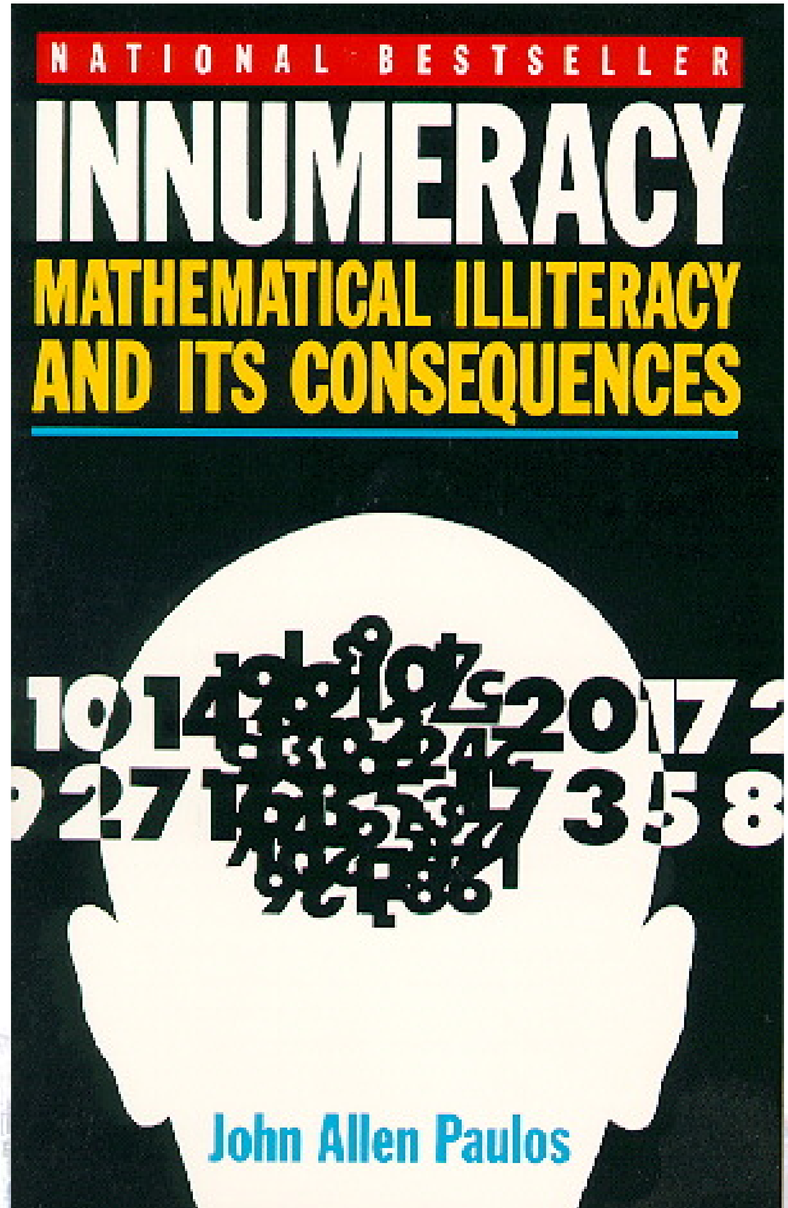
Interesting Probability

- The probability that you just inhaled a molecule which Julius Caesar inhaled in his last breath?
  - Assuming
    - Exhaled Caesar molecules are now uniformly spread around and still free in the atmosphere
  - N molecules of air in the World
  - Caesar exhaled A of them
  - Probability of any given air molecule having been exhaled by Caesar
    - A/N
  - If you inhale B molecules, the probability that none of them are from Caesar is $[1 - A/N]^B$
  - Hence, the probability of inhaling a molecule from Caesar is $1 - [1 - A/N]^B$
    - $A \cong B \cong 2.2 \times 10^{22}$, $N \cong 10^{44}$

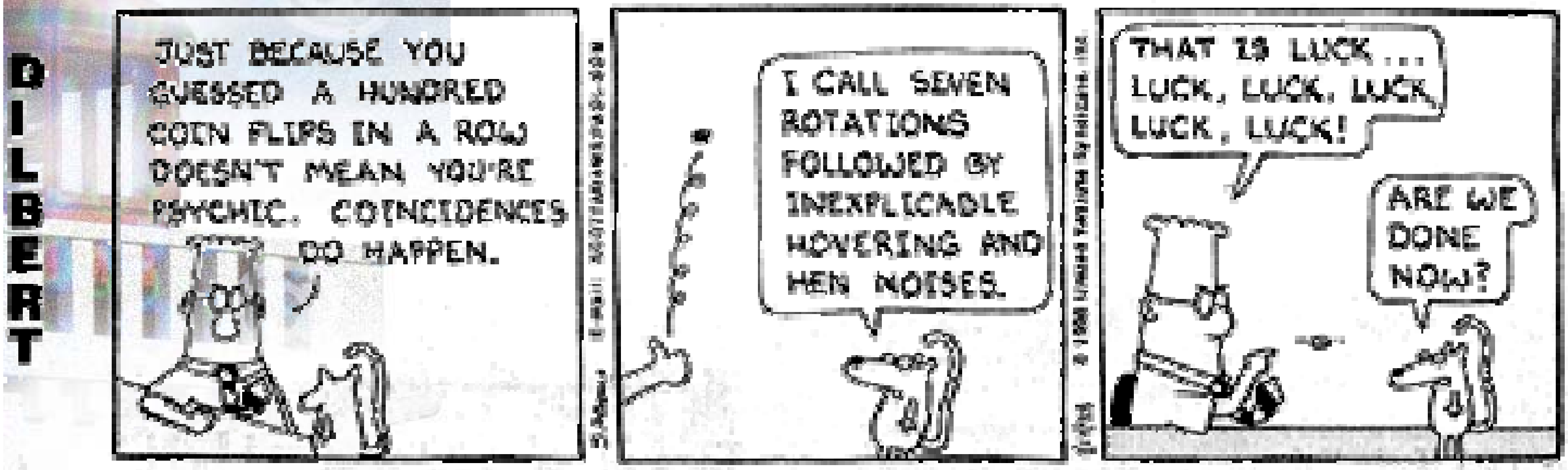
Greater than 99%!!!

"It is no great wonder if, in the long process of time, while fortune takes her course hither and thither, numerous coincidences should spontaneously occur". Plutarch

Et tu, Brutus?



Coincidences



- How many people would there have to be in a group in order for the probability to be 50% that at least two people in it have the same birthday?
  - 23
    - Half the time that 23 randomly selected people are gathered together, two or more will share a birthday

Next Class!

- Topics
  - More Inductive Reasoning Modeling
    - Information and Uncertainty
- Readings for Next week
  - @ *infoport*
  - From course package
    - Norman, G.R. and D.L. Streinrt [2000]. *Biostatistics: The Bare Essentials*.
      - Chapters 1-3 (pages 109-134)
      - OPTIONAL: Chapter 4 (pages 135-140)
      - Chapter 13 (pages 151-159)
      - Chapter 5 (pages 141-144)
    - Von Baeyer, H.C. [2004]. *Information: The New Language of Science*. Harvard University Press.
      - Chapter 10 (pages 13-17))
- Lab 9: Data analysis with Excel (linear regression)