Website

https://courses.engr.illinois.edu/bioe582

University of Illinois BIOE 582 (Fall 2017) Statistics and Algorithms in Genomic Biology

Home **Schedule**

Instructors

Sergei Maslov: maslov@illinois.edu
Office: IGB 3406
Office hours: by appointment

Until about mid-November when he will be replaced with
Paul Jensen: pjens@illinois.edu

University of Illinois BIOE 582 (Fall 2017) Statistics and Algorithms in Genomic Biology

Home **Schedule**

Date	Topic	Slides	Matlab code	Homework/Exams
8/22	Why probability and statistics in Genomics?	Lecture 0		

Syllabus

Topical Outline:

- Discrete, continuous, and multivariate probability distributions in genomics
- Parameter estimation and confidence intervals
- Hypotheses testing for one or multiple samples
- Linear regression
- Techniques of gene expression analysis: clustering algs, co-expression networks
- Graph-theoretical approaches to biological networks
- Intro to genomes
- Intro to sequencing technologies
- Microbial genome assembly
- How to find and annotate genes (annotation RAST and NCBI annotation pipeline)
- Final Project: Sequencing and assembling clinical isolates of oral streptococci.

Grading: midterm (40%), final project (60%).

Final Project: Teams of students will complete a de novo assembly for an unsequenced bacteria of clinical relevance. Students will load and run a Oxford Nanopore sequencer and apply standard tools to assemble, quality control, and annotate a genome. The final sequence will be deposited in the NCBI genome repository.

Foundations of Probability

Random experiments

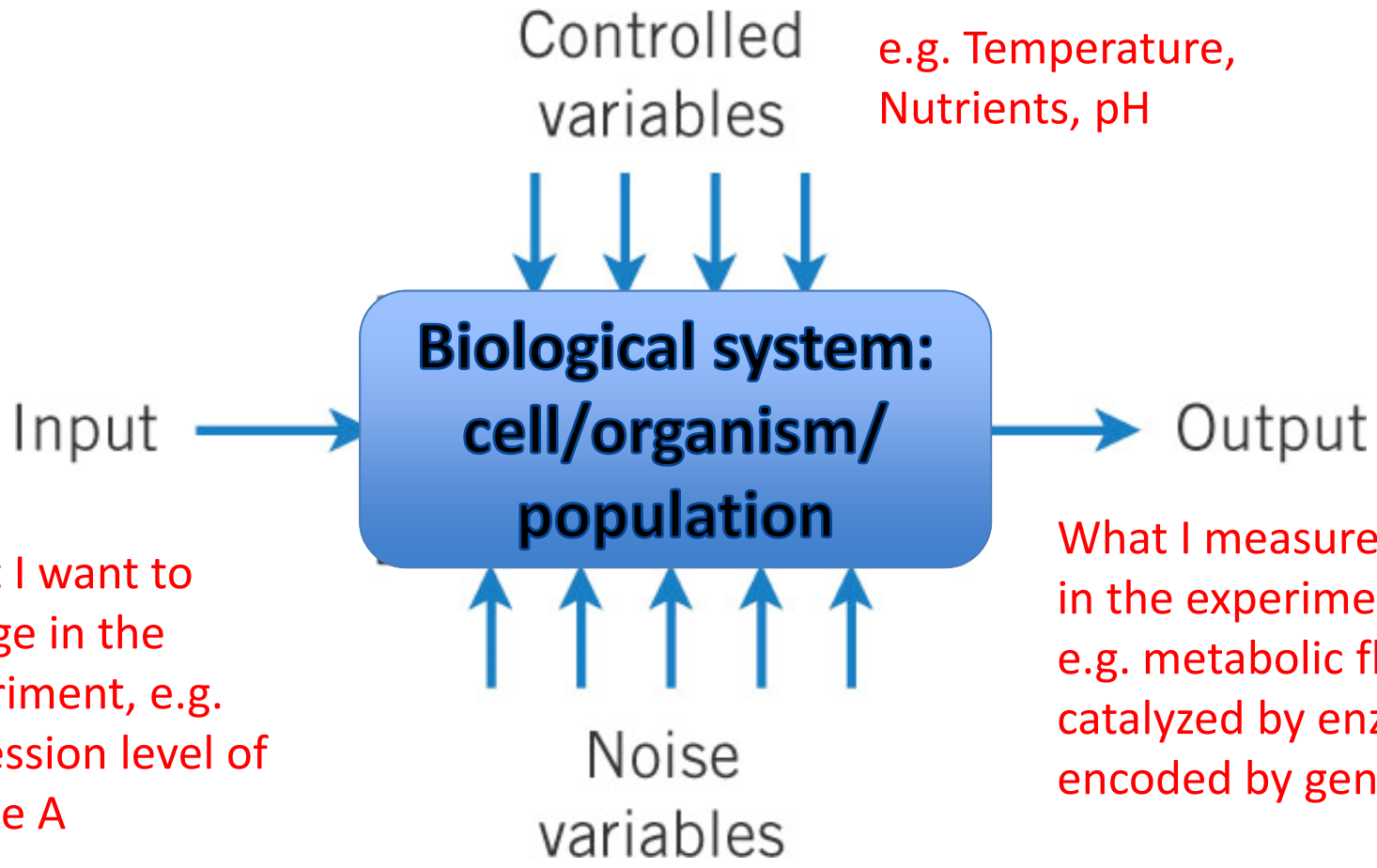
Sample spaces

Venn diagrams of
random events

Random Experiments

- An experiment is an operation or procedure, carried out under controlled conditions
  - Example: measure the metabolic flux through a reaction catalyzed by a given enzyme
- An experiment that can result in different outcomes, even if repeated in the same manner every time, is called a random experiment
  - Cell-to-cell variability due to history/genome variants
  - Noise in external parameters such as temperature, nutrients, pH, etc.
- Evolution offers ready-made random experiments
  - Genomes of different species
  - Genomes of different individuals within a species
  - Individual cancer cells

Changes in Input & Controlled parameters + Noise Produce Output Variation



e.g. Temperature, Nutrients, pH

What I want to change in the Experiment, e.g. expression level of a gene A

What I measure in the experiment, e.g. metabolic flux catalyzed by enzyme encoded by gene A

Internal state of individual cells, signals from neighbors
Uncontrolled fluctuations in T, pH, etc.

Sample Spaces

- Random experiments have unique outcomes.
- The set of all possible outcomes of a random experiment is called the sample space, *S*.
- *S* is discrete if it consists of a finite or countable infinite set of outcomes.
- *S* is continuous if it contains an interval (either a finite or infinite width) or multiple intervals of real numbers.

Examples of Sample Spaces

- Experiment measuring the abundance of mRNA expressed from a single gene
  $S = \{x | x > 0\}$: continuous.
- Bin it into four groups
  $S = \{$*below 10, 10-30, 30-100, above 100*$\}$: discrete.
- Is gene "on" (say >=30) or "off" (<30)?
  $S = \{$*true, false*$\}$: logical/Boolean.
- Time course of expression of 25,000 human genes measured every 1hr for 3 hours:
  $S = \{x_1(1), x_2(1), \ldots x_{25000}(1),$
  $x_1(2), x_2(2), \ldots x_{25000}(2),$
  $x_1(3), x_2(3), \ldots x_{25000}(3) \mid \text{all } x_i(t) > 0\}$

Event

An event (*E*) is a subset of the sample space of a random experiment, i.e., one or more outcomes of the sample space.

- The union of two events is the event that consists of all outcomes that are contained in either of the two events. We denote the union as $E_1 \cup E_2$

- The intersection of two events is the event that consists of all outcomes that are contained in both of the two events. We denote the intersection as $E_1 \cap E_2$

- The complement of an event in a sample space is the set of outcomes in the sample space that are not in the event. We denote the complement of the event E as E' (sometimes E^c or $\bar{E}$)

Examples

Discrete

1. Assume you toss a coin once. The sample space is $S = \{H, T\}$, where H = head and T = tail and the event of a head is $\{H\}$.
2. Assume you toss a coin twice. The sample space is $S = \{(H, H), (H, T), (T, H), (T, T)\}$, and the event of obtaining exactly one head is $\{(H, T), (T, H)\}$.

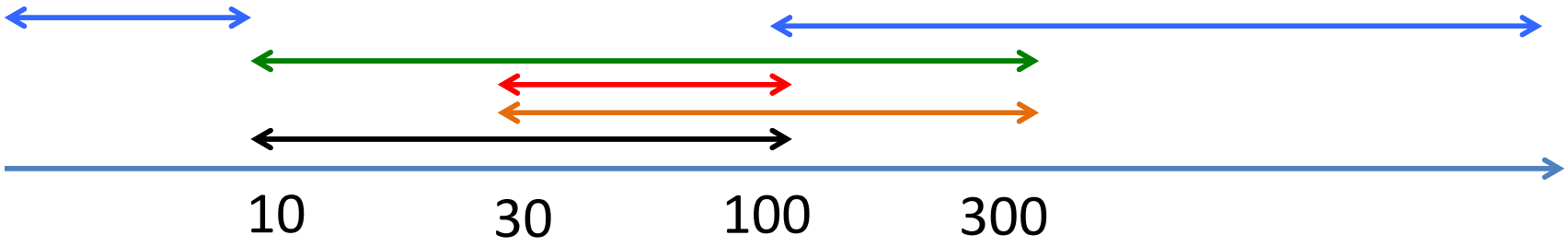
Continuous

Sample space for the expression level of a gene: $S = \{x | x > 0\}$

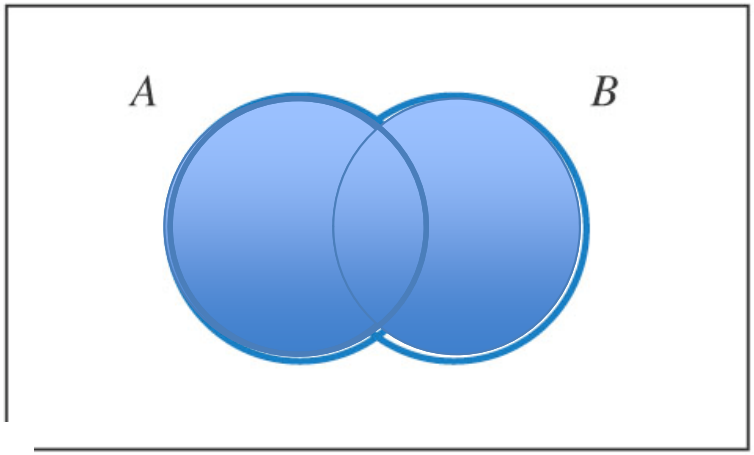
Two events:

- E1 = {x | 10 < x < 100}
- E2 = {x | 30 < x < 300}

- E1 ∩ E2 = {x | 30 < x < 100 }
- E1 U E2 = {x | 10 < x < 300}
- E1' = {x | x ≤ 10 or x ≥ 100}



Venn diagrams



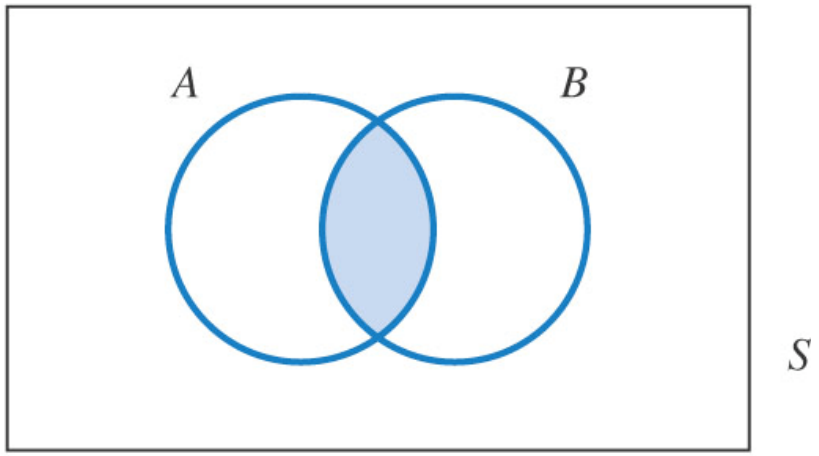
A U B

John Venn (1843-1923)
British logician

Find
5 differences
in beard and
hairstyle

John Venn (1990-)
Brooklyn hipster

Venn diagrams



Which formula describes the blue region?

A. A ∪ B

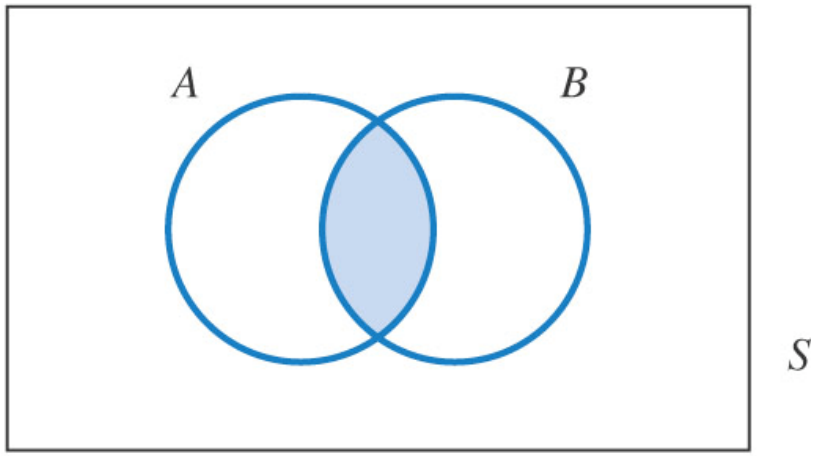
B. A ∩ B

C. A’

D. B’

Get your i-clickers

Venn diagrams



Which formula describes the blue region?

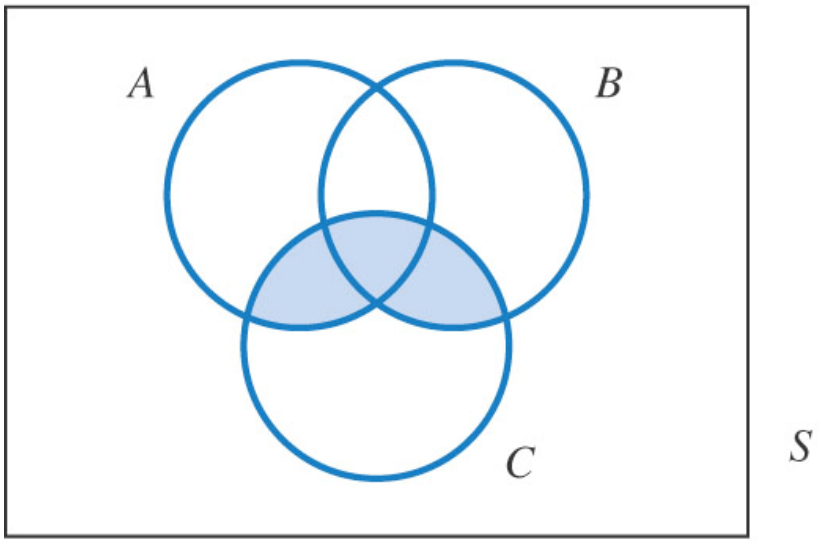
A. A ∪ B

B. A ∩ B

C. A’

D. B’

Venn diagrams



Which formula describes the blue region?

A. (A U B) ∩ C

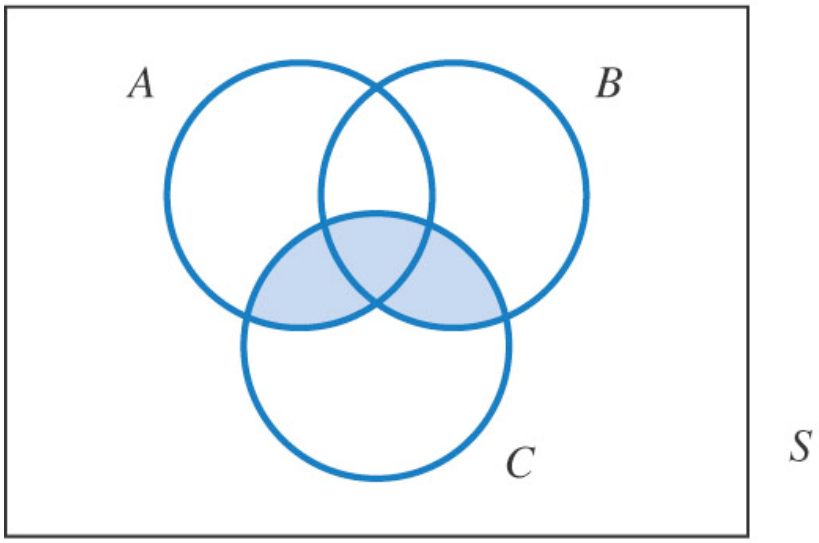
B. (A ∩ B) ∩ C

C. (A U B) U C

D. (A ∩ B) U C

Get your i-clickers

Venn diagrams



Which formula describes the blue region?

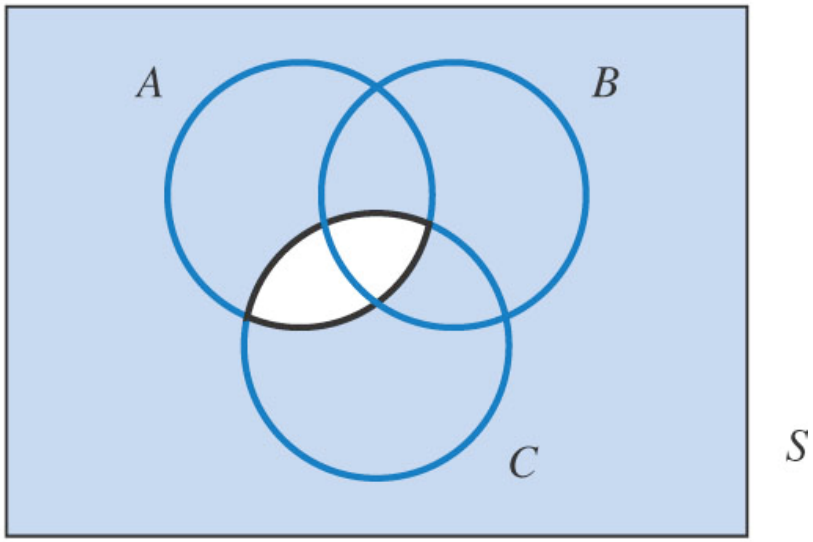
A. (A U B) ∩ C

B. (A ∩ B) ∩ C

C. (A U B) U C

D. (A ∩ B) U C

Venn diagrams



Which formula describes the blue region?

A. A ∩ C

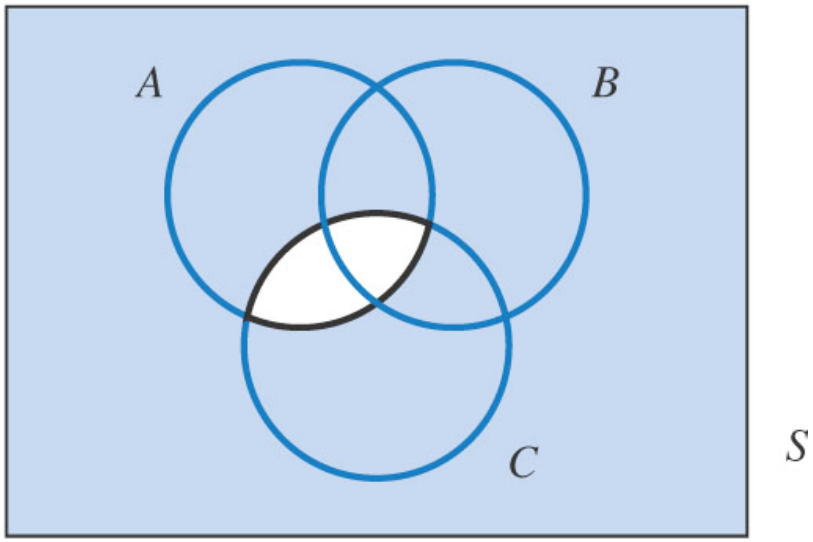
B. A' U C'

C. (A ∩ B ∩ C)'

D. (A ∩ B) ∩ C

Get your i-clickers

Venn diagrams



Which formula describes the blue region?

A. A ∩ C

B. A' U C'

C. (A ∩ B ∩ C)'

D. (A ∩ B) ∩ C

Definitions of Probability

Two definitions of probability

- (1) STATISTICAL PROBABILITY: the relative frequency with which an event occurs in the long run

- (2) INDUCTIVE PROBABILITY: the degree of belief which it is reasonable to place in a proposition on given evidence

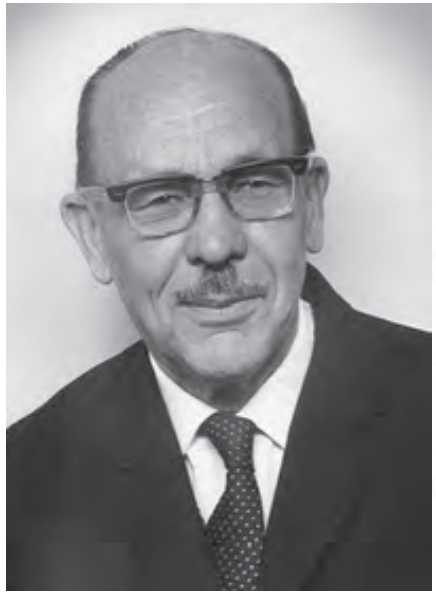
Bulmer, M. G.. Principles of Statistics (Dover Books on Mathematics)

Statistical Probability

A statistical probability of an event is the limiting value of the relative frequency with it occurs in a very large number of independent trials

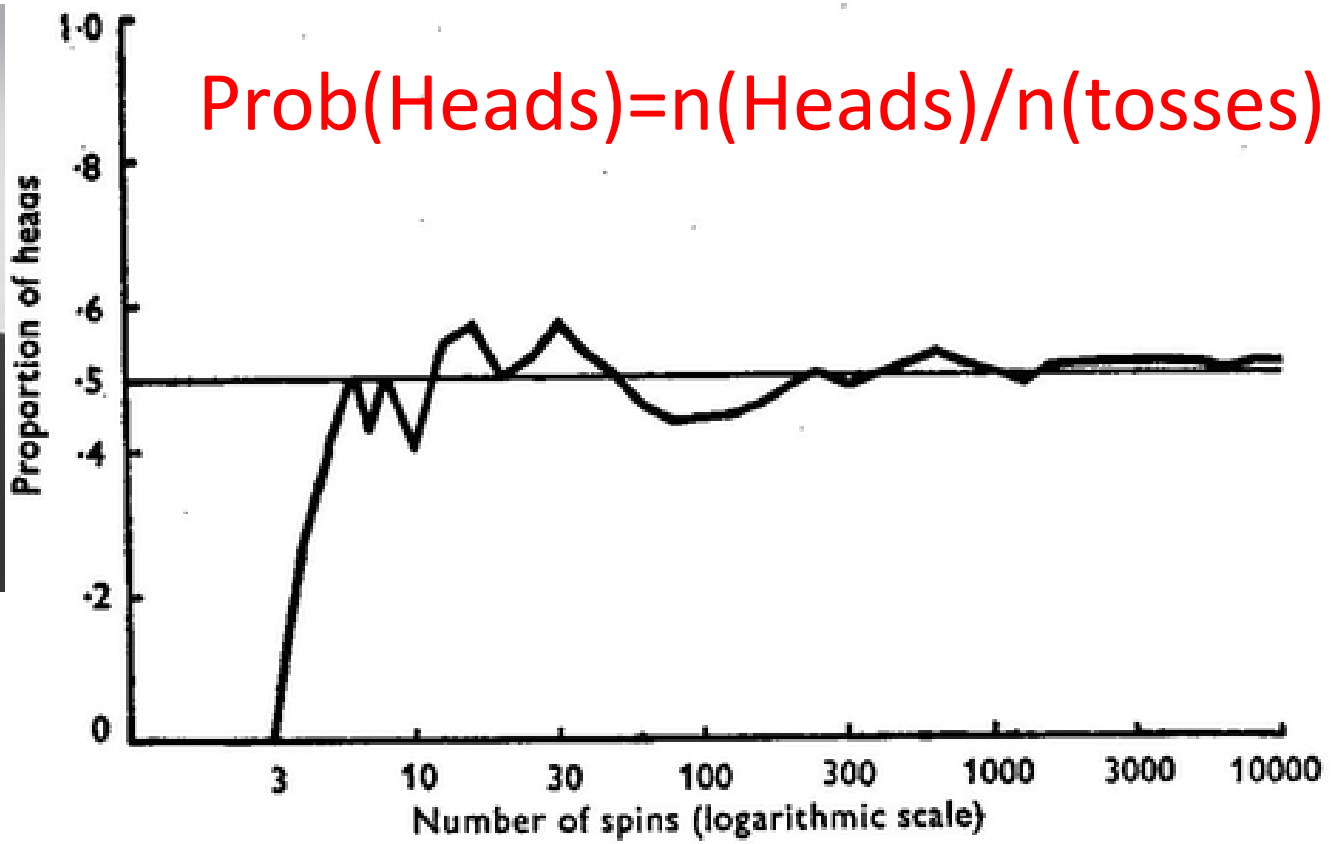
Empirical

Statistical Probability of Coin Toss



John Edmund Kerrich
(1903–1985)
British/South African
mathematician

Prob(Heads)=n(Heads)/n(tosses)



Proportion of heads among 10,000 coin tosses (Kerrich 1946)

Matlab is easy to learn

- Matlab is the lingua franca of all of **engineering**
- We are working with CS department to reinsta teaching Matlab in CS 101
- Meanwhile, use online tutorials e.g.: https://www.youtube.com/watch?v=82TGgQApFIQ
- Matlab is designed to work with Matrices → symbols * and / are understood as matrix multiplication and division
- Use .* and ./ for regular (non-matrix) multiplication
- Add ; in the end of the line to avoid displaying the output on the screen
- Loops: for i=1:100; f(i)=2.*rand; end;
- Conditional statements: if rand>0.5; count=count+1; end;
- Plotting: plot(x,y,'ko-'); or semilogx(x,y,'ko-'); or loglog(x,y,'ko-'); .
  To keep adding plots onto the same axes use: hold on;
  To create a new axes use figure;
- Generating matrices: rand(100) – generates square matrix 100x100.
  Confusing! Use rand(100,1) or zeros(30,20), or randn(1,40);
- If Matlab complains multiplying matrices check sizes using whos and if needed use transpose operation: x=x';

A Matlab Cheat-sheet (MIT 18.06, Fall 2007)

Basics:

`save 'file.mat'` save variables to *file.mat*
`load 'file.mat'` load variables from *file.mat*
`diary on` record input/output to file *diary*
`diary off` stop recording
`whos` list all variables currenly defined
`clear` delete/undefine all variables
`help command` quick help on a given *command*
`doc command` extensive help on a given *command*

Defining/changing variables:

`x = 3` define variable x to be 3
`x = [1 2 3]` set x to the 1×3 row-vector (1,2,3)
`x = [1 2 3];` same, but don't echo x to output
`x = [1;2;3]` set x to the 3×1 column-vector (1,2,3)
`A = [1 2 3 4;5 6 7 8;9 10 11 12];`
set A to the 3×4 matrix with rows 1,2,3,4 etc.
`x(2) = 7` change x from (1,2,3) to (1,7,3)
`A(2,1) = 0` change $A_{2,1}$ from 5 to 0

Arithmetic and functions of numbers:

`3*4, 7+4, 2-6 8/3` multiply, add, subtract, and divide numbers
`3^7, 3^(8+2i)` compute 3 to the 7th power, or 3 to the 8+2i power
`sqrt(-5)` compute the square root of –5
`exp(12)` compute e^{12}
`log(3), log10(100)` compute the natural log (ln) and base-10 log ($\log_{10}$)
`abs(-5)` compute the absolute value |–5|
`sin(5*pi/3)` compute the sine of 5π/3
`besselj(2,6)` compute the Bessel function $J_2(6)$

Arithmetic and functions of vectors and matrices:

`x * 3` multiply every element of x by 3
`x + 2` add 2 to every element of x
`x + y` element-wise addition of two vectors x and y
`A * y` product of a matrix A and a vector y
`A * B` product of two matrices A and B
`x * y` not allowed if x and y are two column vectors!
`x .* y` element-wise product of vectors x and y
`A^3` the square matrix A to the 3rd power
`x^3` not allowed if x is not a square matrix!
`x.^3` every element of x is taken to the 3rd power
`cos(x)` the cosine of every element of x
`abs(A)` the absolute value of every element of A
`exp(A)` e to the power of every element of A
`sqrt(A)` the square root of every element of A
`expm(A)` the matrix exponential e^A
`sqrtm(A)` the matrix whose square is A

Constructing a few simple matrices:

`rand(12,4)` a 12×4 matrix with uniform random numbers in [0,1)
`randn(12,4)` a 12×4 matrix with Gaussian random (center 0, variance 1)
`zeros(12,4)` a 12×4 matrix of zeros
`ones(12,4)` a 12×4 matrix of ones
`eye(5)` a 5×5 identity matrix I ("eye")
`eye(12,4)` a 12×4 matrix whose first 4 rows are the 4×4 identity
`linspace(1.2,4.7,100)`
row vector of 100 equally-spaced numbers from 1.2 to 4.7
`7:15` row vector of 7,8,9,…,14,15
`diag(x)` matrix whose diagonal is the entries of x (and other elements = 0)

Portions of matrices and vectors:

`x(2:12)` the 2nd to the 12th elements of x
`x(2:end)` the 2nd to the last elements of x
`x(1:3:end)` every third element of x, from 1st to the last
`x(:)` all the elements of x
`A(5,:)` the row vector of every element in the 5th row of A
`A(5,1:3)` the row vector of the first 3 elements in the 5th row of A
`A(:,2)` the column vector of every element in the 2nd column of A
`diag(A)` column vector of the diagonal elements of A

Solving linear equations:

`A \ b` for A a matrix and b a column vector, the solution x to $Ax=b$
`inv(A)` the inverse matrix A^{-1}
`[L,U,P] = lu(A)` the LU factorization $PA=LU$
`eig(A)` the eigenvalues of A
`[V,D] = eig(A)` the columns of V are the eigenvectors of A, and the diagonals `diag(D)` are the eigenvalues of A

Plotting:

`plot(y)` plot y as the y axis, with 1,2,3,… as the x axis
`plot(x,y)` plot y versus x (must have same length)
`plot(x,A)` plot columns of A versus x (must have same # rows)
`loglog(x,y)` plot y versus x on a log-log scale
`semilogx(x,y)` plot y versus x with x on a log scale
`semilogy(x,y)` plot y versus x with y on a log scale
`fplot(@(x) …expression…,[a,b])`
plot some expression in x from $x=a$ to $x=b$
`axis equal` force the x and y axes of the current plot to be scaled equally
`title('A Title')` add a title *A Title* at the top of the plot
`xlabel('blah')` label the x axis as *blah*
`ylabel('blah')` label the y axis as *blah*
`legend('foo','bar')` label 2 curves in the plot *foo* and *bar*
`grid` include a grid in the plot
`figure` open up a new figure window

Transposes and dot products:

`x.', A.'` the transposes of x and A
`x', A'` the complex-conjugate of the transposes of x and A
`x' * y` the dot (inner) product of two *column* vectors x and y
`dot(x,y), sum(x.*y)` …two other ways to write the dot product
`x * y'` the *outer* product of two *column* vectors x and y

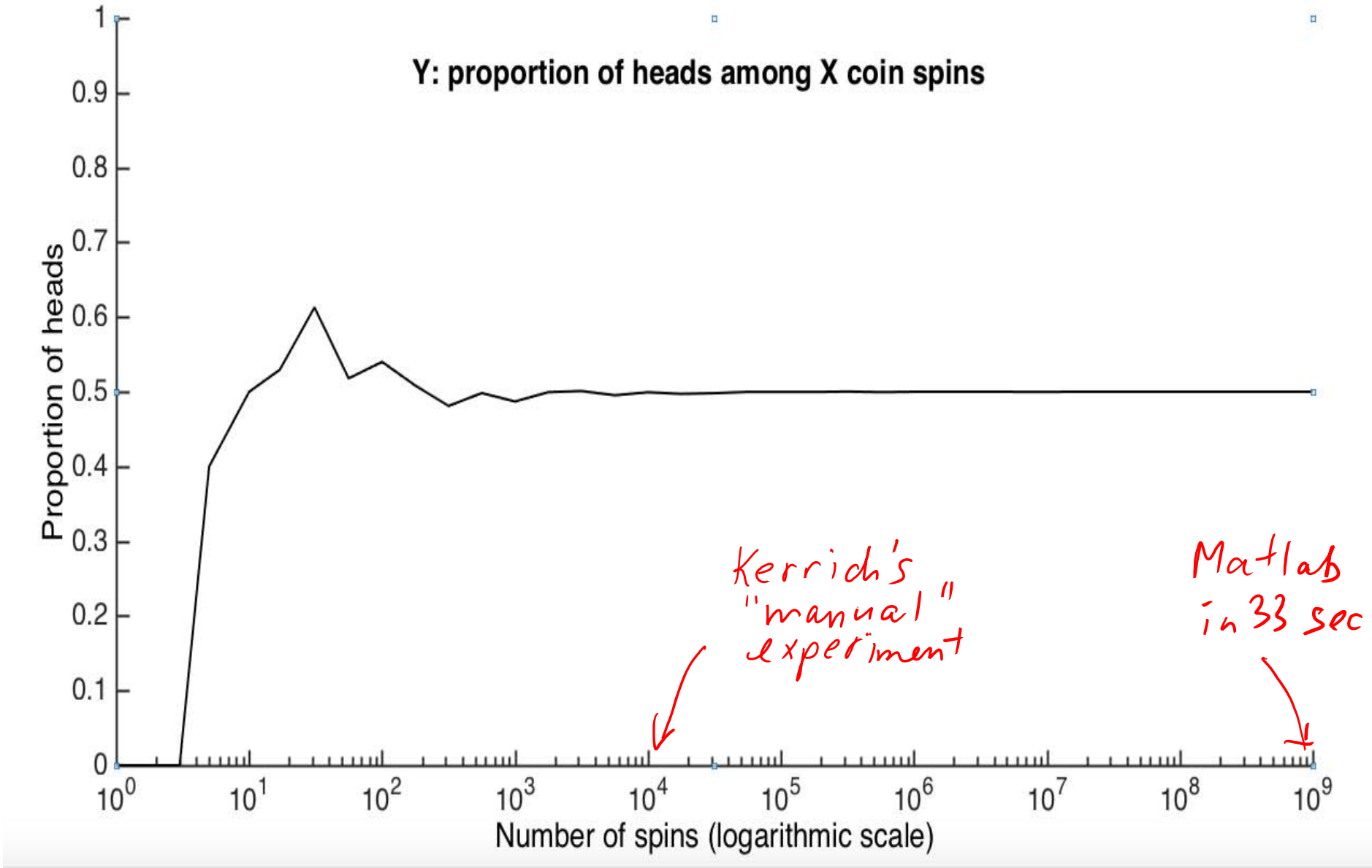
http://web.mit.edu/18.06/www/Spring09/matlab-cheatsheet.pdf

Matlab group exercise

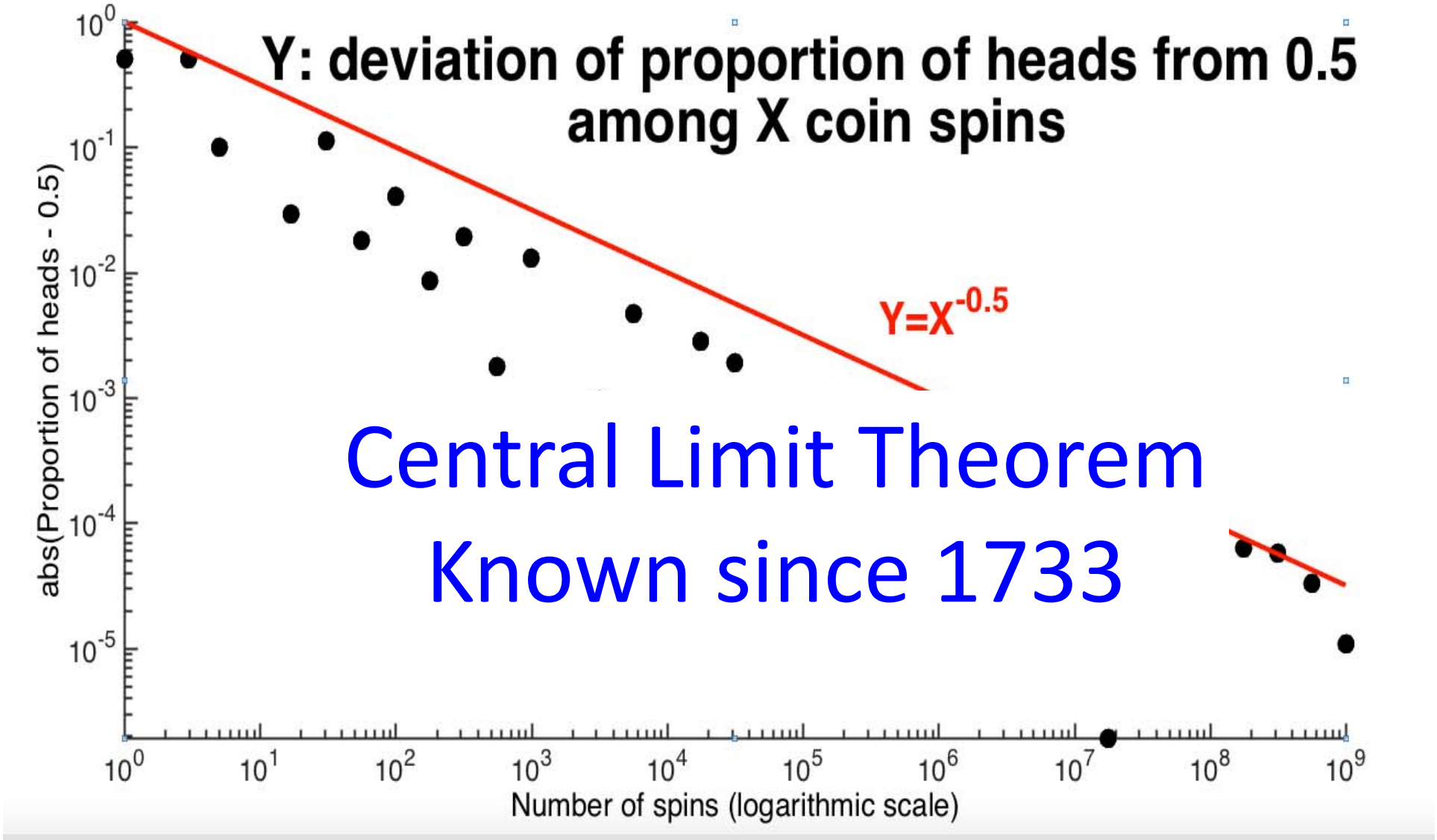
- I want each one of you to write a matlab script to:
  - Simulate a fair coin toss experiment. For this Use floor(2.*rand) to generate a fair coin toss:
    1 – heads, 0 - tails
  - Calculate the fraction of heads f_heads(t) in
    t=10;
      1000;
      100,000;
      10,000,000 throws
  - Plot f_heads(t) vs t with a logarithmic t-axis
  - Plot abs(f_heads(t)-0.5) vs t on a log-log plot
    (both axes are logarithmic)

How I did it

- Stats=1e7;
- r0=rand(Stats,1);
- r1=floor(2.*r0);
- n_heads(1)=r1(1);
- for t=2:Stats; n_heads(t)=n_heads(t-1)+r1(t); end;
- tp=[1, 10,100,1000, 10000, 100000, 1000000, 10000000]
- np=n_heads(tp)
- fp=np./tp
- figure; semilogx(tp,fp,'ko-');
- hold on; semilogx([1,10000000],[0.5,0.5],'r--');
- figure; loglog(tp,abs(fp-0.5),'ko-');
- hold on; loglog(tp,0.5./sqrt(tp),'r--');



Proportion of heads among 1,000,000,000 coin tosses
(10^5 more than Kerrich) took me 33 seconds on my Surface Book



ABS(Proportion of heads-0.5)
among 100,000,000 coin tosses

QUESTIONS

FOUND IN GOOGLE AUTOCOMPLETE

WHY DO WHALES JUMP
WHY ARE WITCHES GREEN
WHY ARE THERE MIRRORS ABOVE BEDS
WHY DO I SAY UH
WHY IS SEA SALT BETTER
WHY ARE THERE TREES IN THE MIDDLE OF FIELDS
WHY IS THERE NOT A POKEMON MMO
WHY IS THERE LAUGHING IN TV SHOWS
WHY ARE THERE DOORS ON THE FREEWAY
WHY ARE THERE SO MANY SVCHOST.EXE RUNNING
WHY AREN'T THERE ANY COUNTRIES IN ANTARCTICA
WHY ARE THERE SCARY SOUNDS IN MINECRAFT
WHY IS THERE KICKING IN MY STOMACH
WHY ARE THERE TWO SLASHES AFTER HTTP
WHY ARE THERE CELEBRITIES
WHY DO SNAKES EXIST
WHY DO OYSTERS HAVE PEARLS
WHY ARE DUCKS CALLED DUCKS
WHY DO THEY CALL IT THE CLAP
WHY ARE KYLE AND CARTMAN FRIENDS
WHY IS THERE AN ARROW ON AANG'S HEAD
WHY ARE TEXT MESSAGES BLUE
WHY ARE THERE MUSTACHES ON CLOTHES
WHY ARE THERE MUSTACHES ON CARS
WHY ARE THERE MUSTACHES EVERYWHERE
WHY ARE THERE SO MANY BIRDS IN OHIO
WHY IS THERE SO MUCH RAIN IN OHIO
WHY IS OHIO WEATHER SO WEIRD
WHY ARE THERE MALE AND FEMALE BIKES
WHY ARE THERE BRIDESMAIDS
WHY DO DYING PEOPLE REACH UP
WHY AREN'T THERE VARICOSE ARTERIES
WHY ARE OLD KLINGONS DIFFERENT
WHY ARE THERE TINY SPIDERS IN MY HOUSE
WHY DO SPIDERS COME INSIDE
WHY ARE THERE HUGE SPIDERS IN MY HOUSE
WHY ARE THERE LOTS OF SPIDERS IN MY HOUSE
WHY ARE THERE SPIDERS IN MY ROOM
WHY ARE THERE SO MANY SPIDERS IN MY ROOM
WHY DO SPIDER BITES ITCH
WHY IS DYING SO SCARY
WHY IS PROGRAMMING SO HARD
WHY IS THERE A 0 OHM RESISTOR
WHY DO AMERICANS HATE SOCCER
WHY DO RHYMES SOUND GOOD
WHY DO TREES DIE
WHY IS THERE NO SOUND ON CNN
WHY AREN'T POKEMON REAL
WHY AREN'T BULLETS SHARP
WHY DO DREAMS SEEM SO REAL
WHY DO TESTICLES MOVE
WHY ARE THERE PSYCHICS
WHY ARE HATS SO EXPENSIVE
WHY IS THERE CAFFEINE IN MY SHAMPOO
WHY DO YOUR BOOBS HURT
WHY ARE THERE SLAVES IN THE BIBLE
WHY DO TWINS HAVE DIFFERENT FINGERPRINTS
WHY ARE AMERICANS AFRAID OF DRAGONS
WHY IS HTTPS CROSSED OUT IN RED
WHY IS THERE A LINE THROUGH HTTPS
WHY IS THERE A RED LINE THROUGH HTTPS ON FACEBOOK
WHY IS HTTPS IMPORTANT
WHY AREN'T MY ARMS GROWING
WHY ARE THERE SWARMS OF GNATS
WHY IS THERE PHLEGM
WHY AREN'T ECONOMISTS RICH
WHY DO AMERICANS CALL IT SOCCER
WHY ARE MY EARS RINGING
WHY ARE THERE SO MANY AVENGERS
WHY ARE THE AVENGERS FIGHTING THE X MEN
WHY IS WOLVERINE NOT IN THE AVENGERS
WHY ARE THERE SO MANY CROWS IN ROCHESTER, MN
WHY IS PSYCHIC WEAK TO BUG
WHY DO CHILDREN GET CANCER
WHY IS POSEIDON ANGRY WITH ODYSSEUS
WHY IS THERE ICE IN SPACE
WHY ARE THERE ANTS IN MY LAPTOP
WHY IS EARTH TILTED
WHY IS SPACE BLACK
WHY IS OUTER SPACE SO COLD
WHY ARE THERE PYRAMIDS ON THE MOON
WHY IS NASA SHUTTING DOWN
WHY ARE THERE GHOSTS
WHY IS THERE AN OWL IN MY BACKYARD
WHY IS THERE AN OWL OUTSIDE MY WINDOW
WHY IS THERE AN OWL ON THE DOLLAR BILL
WHY DO OWLS ATTACK PEOPLE
WHY ARE AK 47s SO EXPENSIVE
WHY ARE THERE HELICOPTERS CIRCLING MY HOUSE
WHY ARE THERE GODS
WHY ARE THERE TWO SPOCKS
WHY IS MT VESUVIUS THERE
WHY DO THEY SAY T MINUS
WHY ARE THERE OBELISKS
WHY ARE WRESTLERS ALWAYS WET
WHY ARE OCEANS BECOMING MORE ACIDIC
WHY IS ARWEN DYING
WHY AREN'T MY QUAIL LAYING EGGS
WHY AREN'T MY QUAIL EGGS HATCHING
WHY AREN'T THERE ANY FOREIGN MILITARY BASES IN AMERICA
WHY ARE MY BOOBS ITCHY
WHY ARE CIGARETTES LEGAL
WHY ARE THERE DUCKS IN MY POOL
WHY IS JESUS WHITE
WHY IS THERE LIQUID IN MY EAR
WHY DO Q TIPS FEEL GOOD
WHY DO GOOD PEOPLE DIE
WHY AREN'T THERE GUNS IN HARRY POTTER
WHY ARE ULTRASOUNDS IMPORTANT
WHY ARE ULTRASOUND MACHINES EXPENSIVE
WHY IS STEALING WRONG
WHY ARE THERE SQUIRRELS
WHY ARE TREES TALL
WHY DO IGUANAS DIE
WHY AREN'T THERE DINOSAUR GHOSTS
WHY IS THERE HELL IF GOD FORGIVES
WHY IS YKK ON ALL ZIPPERS
WHY ARE THERE WEEKS
WHY DO I FEEL DIZZY
WHY IS THERE LAVA
WHY ARE DOGS AFRAID OF FIREWORKS
WHY IS THERE NO KING IN ENGLAND
WHY IS LIFE SO BORING
WHY ARE THERE FEMALE MR MIMES

Credit: XKCD comics