office hours

The PoCSverse

Volume 1

now on cassette !!!

2022–2023

P o C S What's The Story?

Principles of Complex Systems, Vols. 1, 2, & 3D, CSYS/MATH 300, 303, &
University of Vermont, Fall 2022
Assignment 01
code name: "I Aten't Dead"

Due: Friday, September 2, by 11:59 pm
Relevant clips, episodes, and slides are listed on the assignment's page:
https://pdodds.w3.uvm.edu/teaching/courses/2022-2023pocsverse//assignments/01/
Some useful reminders:
Deliverator: Prof. Peter Sheridan Dodds (contact through Teams)
Assistant Deliverator: Dylan Casey (contact through Teams)
Office: The Ether
Office hours: TBD
Course website: https://pdodds.w3.uvm.edu/teaching/courses/2022-2023pocsverse

---

All parts are worth 3 points unless marked otherwise. Please show all your workingses clearly and list the names of others with whom you collaborated.

For coding, we recommend you improve your skills with Python, R, and/or Julia. The Deliverator uses Matlab.

Graduate students are requested to use LaTeX (or related TeX variant). If you are new to LaTeX, please endeavor to submit at least n questions per assignment in LaTeX, where n is the assignment number.

Assignment submission:

fix.

1. Please send to both the Deliverator and Assistant Deliverator via direct message on Teams.
2. PDF only! Please name your file as follows (where the number is to be padded by a 0 if less than 10 and names are all lowercase): `CSYS300assignment%02d$firstname-$lastname.pdf` as in `CSYS300assignment06michael-palin.pdf`

---

1. An amuse-bouche for scaling, to signal the flavors ahead:

   Examine current weight lifting records for the snatch, clean and jerk, and the total for scaling with body mass (three regressions). Do so for both women and men's records.

   For weight classes, take the upper limit for the mass of the lifter.

   Wikipedia is an excellent source.

   (a) How well does 2/3 scaling hold up?

(b) Normalized by the scaling you determine, who holds the overall, rescaled world record?

Normalization here means relative:

$$100 \times \left(\frac{M_{\text{worldrecord}}}{cM^{\beta}_{\text{weightclass}}.} - 1 \right),$$

where c and β are the parameters determined from a linear fit.

2. Some kitchen table preparation for for power-law size distributions:

   Consider a random variable X with a probability distribution given by

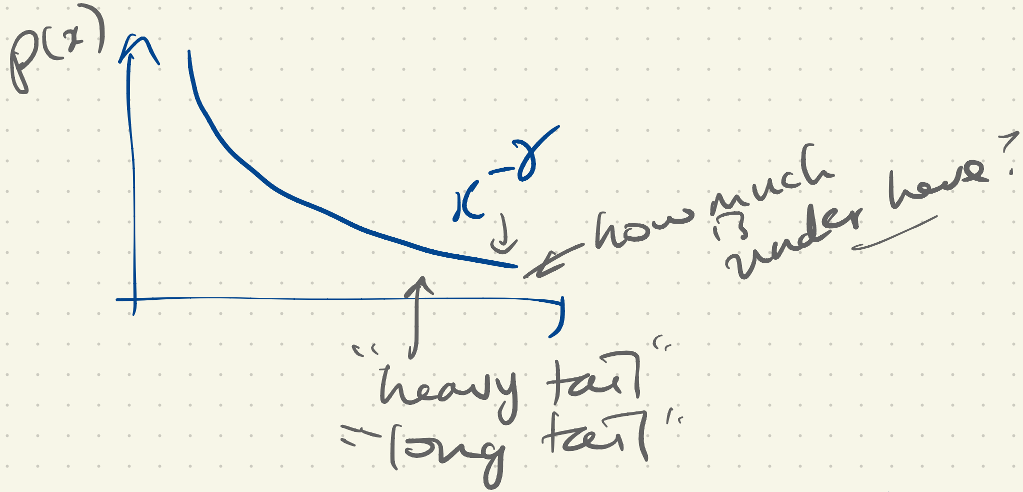
$$P(x) = cx^{-\gamma}$$

   where c is a normalization constant, and $0 < a \leq x \leq b$. (a and b are the lower and upper cutoffs respectively.) Assume that $\gamma > 1$.

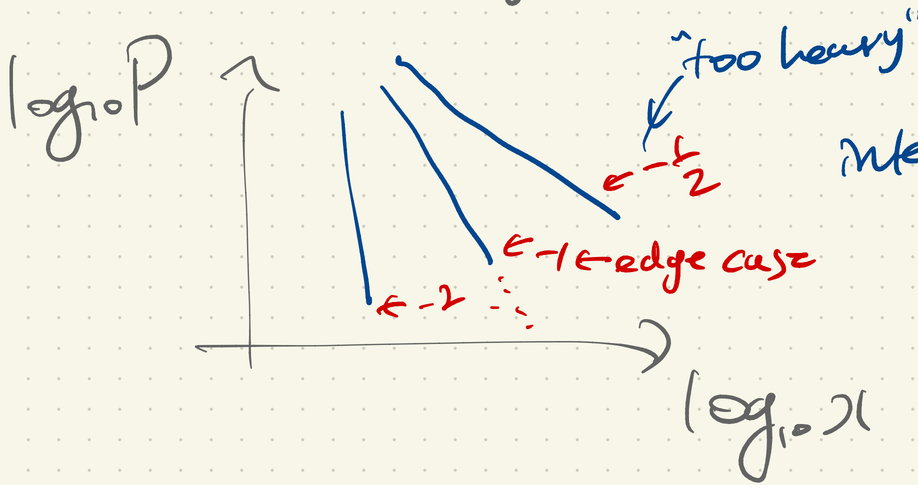
   what happens when $b \to \infty$?

   (a) Determine c.

   (b) Why did we assume $\gamma > 1$?



$\frac{1}{x}$ is the edge case



integral diverges

$\gamma = -1$

$$1 = c\int_a^b \frac{1}{x}\,dx$$

for a healthy prob.

$$= c\ln x \Big|_a^b$$

$$= c\ln b - c\ln a$$

→ diverges as $b \to \infty$.

$\ln x$.

$\ln x \to \infty$ as $x \to \infty$

Super slow

"x^0"-ish.

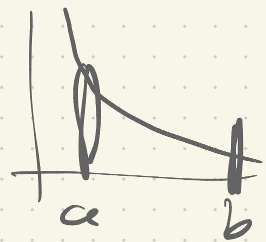
Q2. We must have $\int_{-\infty}^{\infty} P(x)\,dx = 1$.

and $P(x) \geq 0$ for all x

mean

$$\langle x \rangle = \int_{-\infty}^{\infty} x P(x)\,dx$$

for $P(x) = c x^{-\gamma}$



$\gamma > 1$

$$1 = \int_a^b P(x)\,dx$$

$$= c \int_a^b x^{-\gamma}\,dx = c \left. \frac{x^{-\gamma+1}}{-\gamma+1} \right|_a^b = c \left. \frac{x^{-(\gamma-1)}}{-(\gamma-1)} \right|_a^b$$

γ is +ve

clear this is a decaying power law

good tidiness

If $0 < a < b < \infty$,
then c is fine
as integral works

Concern is $b \to \infty$.
full tail

$$1 = c\left(\frac{a^{-(\gamma-1)} - b^{(\gamma-1)}}{(\gamma-1)}\right)$$

again, nice to keep $(\gamma-1)$

$\leftarrow$ +ve as $\gamma > 1$

rate than

$\underbrace{-\gamma+1}$ $\leftarrow$ -ve as $\gamma > 1$,

$l, \quad m, \quad g, \quad \tau$

$L \quad M \quad L/T^2 \quad T$

m: doesn't matter

eyeball:

$$\Pi_1 = \frac{l \, \cancel{\tau^2}}{g\,\tau^2}$$

only one dimensionless parameter

$$\Pi_1 = \text{constant}$$

$$\Rightarrow \tau^2 \propto \frac{l}{g}$$

$$[\Pi_1] = \frac{[l]\,\cancel{[\tau^2]}}{[g][\tau^2]} = \frac{L\,\cancel{T^2}}{\frac{L}{T^2}\,\cancel{T^2}} = 1$$

$[\Pi_1]$: dimension operator

$$\tau \propto \sqrt{l}$$

PoCS What's The Story?

Principles of Complex Systems, Vols. 1, 2, & 3D
CSYS/MATH 300, 303, & 394
University of Vermont, Fall 2022
Assignment 02
Curry with Named Meat 15p

Due: Friday, September 9, by 11:59 pm
Relevant clips, episodes, and slides are listed on the assignment's page:
https://pdodds.w3.uvm.edu/teaching/courses/2022-2023pocsverse//assignments/02/
Some useful reminders:
Deliverator: Prof. Peter Sheridan Dodds (contact through Teams)
Assistant Deliverator: Dylan Casey (contact through Teams)
Office: The Ether
Office hours: TBD
Course website: https://pdodds.w3.uvm.edu/teaching/courses/2022-2023pocsverse

---

All parts are worth 3 points unless marked otherwise. Please show all your workingses clearly and list the names of others with whom you collaborated.

For coding, we recommend you improve your skills with Python, R, and/or Julia. The Deliverator uses Matlab.

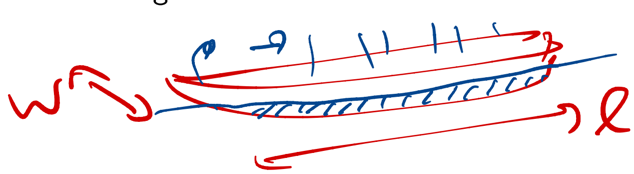
Graduate students are requested to use LaTeX (or related TeX variant). If you are new to LaTeX, please endeavor to submit at least n questions per assignment in LaTeX, where n is the assignment number.

Assignment submission:

When available, via Blackboard.

---

1. Use a back-of-an-envelope scaling argument to show that maximal rowing speed V_{max} increases as the number of oarspeople N as $V_{max} \propto N^{1/9}$.
   Assume the following:
   (a) Rowing shells are geometrically similar (isometric). The table below taken from McMahon and Bonner [1] shows that shell width is roughly proportional to shell length ℓ.



w

choose variable designation

isometry: $w \propto \ell$

Volume ∝ W·H·L ∝ ℓ^3 b/c Isometry

big assumption

Shell dimensions and performances.

No. of oarsmen	Modifying description	Length, l (m)	Beam, b (m)	l/b	Boat mass per oarsman (kg)	Time for 2000 m (min) I	II	III	IV
8	Heavyweight	18.28	0.610	30.0	14.7	5.87	5.92	5.82	5.73
8	Lightweight	18.28	0.598	30.6	14.7				
4	With coxswain	12.80	0.574	22.3	18.1				
4	Without coxswain	11.75	0.574	21.0	18.1	6.33	6.42	6.48	6.13
2	Double scull	9.76	0.381	25.6	13.6				
2	Pair-oared shell	9.76	0.356	27.4	13.6	6.87	6.92	6.95	6.77
1	Single scull	7.93	0.293	27.0	16.3	7.16	7.25	7.28	7.17

$\ell \propto V^2$?

(b) The resistance encountered by a shell is due largely to drag on its wetted surface.

$D_f \propto V^2 \ell^2$

(c) Drag force is proportional to the product of the square of the shell's speed (V^2) and the area of the wetted surface ($\propto \ell^2$ due to shell isometry).

(d) Power $\propto$ drag force $\times$ speed (in symbols: $P \propto D_f \times V$). ✓ $P \propto D_f \propto V$

(e) Volume displacement of water by a shell is proportional to the number of oarspeople N (i.e., the team's combined weight). $N \propto \ell^3 \propto$ Volume

(f) Assume the depth of water displacement by the shell grows isometrically with boat length ℓ. → $P \propto \ell^3$

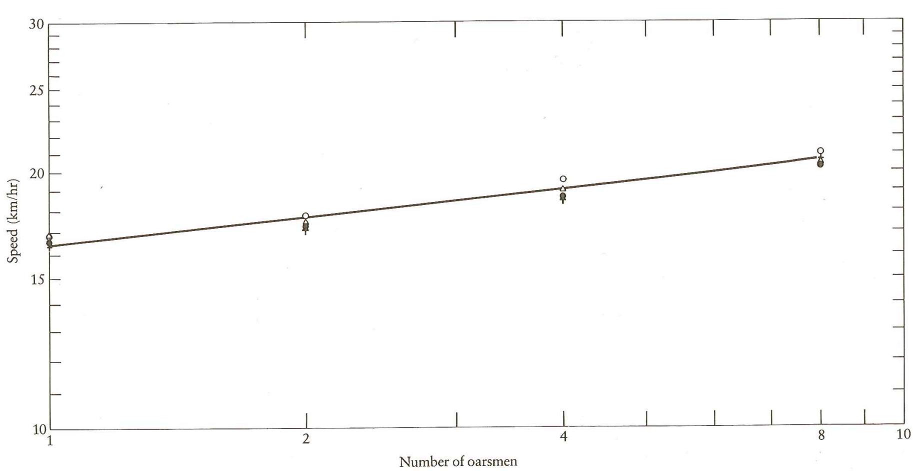
is (f) too much informat?

(g) Power is proportional to the number of oarspeople N. → $P \propto N$

2. Find the modern day world record times for 2000 metre races and see if this scaling still holds up. Of course, our relationship is approximate as we have neglected numerous factors, the range is extremely small (1–8 oarspeople), and the scaling is very weak (1/9). But see what you can find. The figure below shows data from McMahon and Bonner.

eliminate P, ℓ, D_f keep N, V

game



3. Finish the calculation for the platypus on a pendulum problem so show that a simple pendulum's period τ is indeed proportional to $\sqrt{\ell/g}$.

Basic plan from lectures: Create a matrix A where ijth entry is the power of dimension i in the jth variable, and solve by row reduction to find basis null vectors.

In lectures, we arrived at:

$$A\vec{x} = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & -2 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

You only have to take a few steps from here.

From Lecture 3: the Buckingham π theorem (20 minutes).

4. Show that the maximum speed of animals $V_{\max}$ is proportional to their length L [2]. Here are five dimensionful parameters:

   - $V_{\max}$, maximum speed.
   - ℓ, animal length.
   - ρ, organismal density.
   - σ, maximum applied force per unit area of tissue.
   - b, maximum metabolic rate per unit mass (b has the dimensions of power per unit mass).

   And here are the three dimensions: L, M, and T.

   Use a back-of-the-envelope calculation to express $V_{\max}/\ell$ in terms of ρ, σ, and b.

   Note: It's argued in [2] that these latter three parameters vary little across all organisms (we're mostly thinking about running organisms here), and so finding $V_{\max}/\ell$ as a function of them indicates that $V_{\max}/\ell$ is also roughly constant.

5. Use the Buckingham π theorem to reproduce G. I. Taylor's finding the energy of an atom bomb E is related to the density of air ρ and the radius of the blast wave R at time t:

   $$E = \text{constant} \times \rho R^5/t^2.$$

   In constructing the matrix, order parameters as E, ρ, R, and t and dimensions as L, T, and M.

6. Use the Buckingham π theorem to derive Kepler's third law, which states that the square of the orbital period of a planet is proportional to the cube of its semi-major axis.

Let's shed some enlightenment and assume circular orbits.

Parameters:

- Planet's mass m; $[m] = M$
- Sun's mass M_s; $[M_s] = M$
- Orbital period T; $[T] = T$
- Orbital radius r; $[R] = L$
- Graviational constant G. $[G] = \frac{L^3}{MT^2}$

M
T
L

$F = \overset{\text{const}}{G}\frac{m_1 M_2}{r^2}$

(a) What are the dimensions of these five quantities?

(b) You will find that there are two dimensionless parameters using the Buckingham π theorem, and that you can choose one to be $\pi_2 = m/M$. Find the other dimensionless parameter, π_1.

$\pi_1 = \left(\frac{M_s G T^2}{R^3}\right)$

OR

$\frac{m}{M_s} = \pi_2$

(c) Now argue that $T^2 \propto r^3$.

(d) For our solar system's nine (9) planets (yes, Pluto is on the team here), plot T^2 versus r^3, and using basic linear regression report on how well Kepler's third law holds up.

$f(\pi_1, \pi_2) = c$

or $\pi_1 = f(\pi_2)$

$\sim f(0)$

$\pi_2 = \frac{m}{M_s}$

References

[1] T. A. McMahon and J. T. Bonner. *On Size and Life*. Scientific American Library, New York, 1983.

[2] N. Meyer-Vernet and J.-P. Rospars. How fast do living organisms move: Maximum speeds from bacteria to elephants and whales. *American Journal of Physics*, pages 719–722, 2015. pdf

Assume $f(x) \to$ const

as $x \to 0$

$\frac{m}{M_s}$

how things could go wrong $\rightarrow$

Worry is:

$f(x) \sim x^{-\beta}$ as $x \to 0$

$$A\vec{x} = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & -2 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

RREF

$$\left[\begin{array}{cccc|c} 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & -2 & 1 & 0 \end{array}\right]$$

$\sim$ is equivalent to

$R_3' = -\frac{1}{2}R_3$
$$\left[\begin{array}{cccc|c} 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & -\frac{1}{2} & 0 \end{array}\right]$$

"remove"

pivot

$\sim$

$R_1' = R_1 - R_3$
$$\left[\begin{array}{cccc|c} 1 & 0 & 0 & +\frac{1}{2} & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & -\frac{1}{2} & 0 \end{array}\right]$$

$\underbrace{}_{\text{I}}$

$$\overbrace{x_1}^{\text{pivot}} \quad + \overbrace{\tfrac{1}{2}x_4}^{\text{free}} = 0$$
$$x_2 \quad = 0$$
$$x_3 + -\tfrac{1}{2}x_4 = 0$$

express pivots in terms of free var

$$x_1 = -\frac{1}{2}x_4$$
$$x_2 = 0 = 0 \cdot x_4$$
$$x_3 = \frac{1}{2}x_4$$

pivots

$$\vec{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} -\frac{1}{2}x_4 \\ 0 \cdot x_4 \\ \frac{1}{2}x_4 \\ x_4 \end{bmatrix} = x_4 \begin{bmatrix} -\frac{1}{2} \\ 0 \\ \frac{1}{2} \\ 1 \end{bmatrix}$$

free

$x_4 \in R$

nullspace is 1d subspace in 4d.

$$\Pi_1 = \left((l)^{-\frac{1}{2}}(m)^0(g)^{1/2}(\tau)^1\right)^{x_4 \in \text{free choice}}$$

any power

for example choose $x_4 = 2$

$$\Pi_1 = \frac{g\tau^2}{l}$$

Because only one dimensionless parameter,

must have pendulum equation:

$$f(\Pi_1) = \text{const}$$

we don't know what this is

$$\Rightarrow \Pi_1 = \text{const.}$$

$$\Pi_1 = \frac{g\tau^2}{l} \rightarrow \tau \propto \sqrt{\frac{l}{g}}$$

amazing!

not true:

$$\Pi_1^3 = 47$$

to figure out

$$f(\Pi_1, \Pi_2) = 0$$

Propose

$$\pi = E^{x_1} \rho^{x_2} R^{x_3} t^{x_4}$$

Want $[\pi] = 1$

$$1 = [\pi] = [E^{x_1}][\rho^{x_2}][R^{x_3}][t^{x_4}]$$

$$= \left(\frac{ML^2}{T^2}\right)^{x_1} \left(\frac{M}{L^3}\right)^{x_2} L^{x_3} T^{x_4}$$

$$= M^{x_1 + x_2} L^{(\quad)} T^{(\quad)}$$

$0 \qquad 0 \qquad 0.$

$$F = \frac{G M_1 M_2}{r^2}$$

aside $\frac{P_1 P_2}{d}$

gravity model

$$[F] = [G]\frac{MM}{L^2}$$

$\downarrow$

ma

$\downarrow$

$$\frac{ML}{T^2}$$

//

$$\frac{ML}{T^2} = [G]\frac{M^2}{L^2}$$

$$[G] = \frac{L^3}{MT^2}$$

Principles of Complex Systems, Vols. 1, 2, & 3D
CSYS/MATH 300, 303, & 394
University of Vermont, Fall 2022
Assignment 03
kingons, or possibly queons

Due: Friday, September 16, by 11:59 pm
https://pdodds.w3.uvm.edu/teaching/courses/2022-2023pocsverse/assignments/03/
Some useful reminders:
Deliverator: Prof. Peter Sheridan Dodds (contact through Teams)
Assistant Deliverator: Dylan Casey (contact through Teams)
Office: The Ether
Office hours: TBD
Course website: https://pdodds.w3.uvm.edu/teaching/courses/2022-2023pocsverse

---

All parts are worth 3 points unless marked otherwise. Please show all your workingses clearly and list the names of others with whom you ~~conspired~~ collaborated.

For coding, we recommend you improve your skills with Python, R, and/or Julia. The (evil) Deliverator uses (evil) Matlab.

Graduate students are requested to use LaTeX (or related TeX variant). If you are new to LaTeX, please endeavor to submit at least n questions per assignment in LaTeX, where n is the assignment number.

Assignment submission:

Via Blackboard.

All about power law size distributions (basic computations and some real life data from Google).

Note 1: Please do not use Mathematica, etc. for any symbolic work—you can do all of these calculations by hand. Yes you can!

Note 2: Otherwise, use whatever tools you like for the data analysis.

1. As in assignment 1, consider a random variable X with a probability distribution given by
$$P(x) = cx^{-\gamma},$$
where c is a normalization constant you determined in the first assignment, and $0 < a \leq x \leq b$. (a and b are the lower and upper cutoffs respectively.) Assume that $\gamma > 1$.

Note: For all answers you obtain for the questions below, replace c by the expression you obtained in the first assignment, and simplify expressions as much as possible.

Compute the nth moment of X which is in general defined as:

$$\langle x^n \rangle = \int_a^b x^n P(x)\,\mathrm{d}x$$

"expected value of $<x^n>$"

bring c in from asst 1

2. In the limit $b \to \infty$, how does the nth moment behave as a function of γ?

3. (a) Find σ, the standard deviation of X for finite a and b, then obtain the limiting form of σ as $b \to \infty$, noting any constraints we must place on γ for the mean and the standard deviation to remain finite as $b \to \infty$.

   Some help: the form of σ^2 as $b \to \infty$ should reduce to

   $$= \frac{(\gamma - c_1)}{(\gamma - c_2)(\gamma - c_3)^2} a^2$$

   where c_1, c_2, and c_3 are simple, meaningful constants to be determined (by you).

   (b) For the case of $b \to \infty$, how does σ behave as a function of γ, given the constraints you have already placed on γ? More specifically, how does σ behave as γ reaches the ends of its allowable range?

4. Drawing on a Google vocabulary data set (see below for links)

   (a) Plot the frequency distribution N_k representing how many distinct words appear k times in this particular corpus as a function of k.

   (b) Repeat the same plot in log-log space (using base 10, i.e., plot $\log_{10} N_k$ as a function of $\log_{10} k$).

5. Using your eyeballs, indicate over what range power-law scaling appears to hold and, estimate, using least squares regression over this range, the exponent in the fit $N_k \sim k^{-\gamma}$ (we'll return to this estimate later).

measurement from data question

6. Compute the mean and standard deviation for the entire sample (not just for the restricted range you used in the preceding question). Based on your answers to the following questions and material from the lectures, do these values for the mean and standard deviation make sense given your estimate of γ?

   Hint: note that we calculate the mean and variance from the distribution N_k; a common mistake is to treat the distribution as the set of samples. Another routine misstep is to average numbers in log space (oops!) and to average only over the range of k values you used to estimate γ.

   The data for N_k and k (links are clickable):

- Compressed text file (first column $= k$, second column $= N_k$):
  https://pdodds.w3.uvm.edu/teaching/courses/2022-2023pocsverse/docs/vocab_cs_mod.txt.gz
- Uncompressed text file (first column $= k$, second column $= N_k$):
  https://pdodds.w3.uvm.edu/teaching/courses/2022-2023pocsverse/docs/vocab_cs_mod.txt
- Matlab file (`wordfreqs` $= k$, `counts` $= N_k$):
  https://pdodds.w3.uvm.edu/teaching/courses/2022-2023pocsverse/docs/google_vocab_freqs.mat

Easier!

The raw frequencies of individual words:

- https://pdodds.w3.uvm.edu/teaching/courses/2022-2023pocsverse/docs/google_vocab_rawwordfreqs.txt.gz
- https://pdodds.w3.uvm.edu/teaching/courses/2022-2023pocsverse/docs/google_vocab_rawwordfreqs.txt
- https://pdodds.w3.uvm.edu/teaching/courses/2022-2023pocsverse/docs/google_vocab_rawwordfreqs.mat

mean of all numbers in sample

Note: 'words' here include any separate textual object including numbers, websites, html markup, etc.

Note: To keep the file to a reasonable size, the minimum number of appearances is $k_{\min} = 200$ *corresponding to* $N_{200} = 48030$ *distinct words that each appear 200 times.*

how does
$x \ln x$ behave as $x \to 0$?
$0 \cdot \cancel{\ln 0}\ \infty$?

$$\lim_{x\to 0} x\ln x = \lim_{x\to 0} \frac{\ln x}{\frac{1}{x}} = \lim_{x\to 0} \frac{\frac{1}{x}}{\frac{-1}{x^2}} = 0.$$

use l'Hôpital's rule

Moments

$\langle x^1 \rangle$ = mean of $P(x)$ = $\int_a^b \underline{x^1} P(x)\,dx$

$\langle x^2 \rangle \hookrightarrow$ $\sigma^2 = \langle x^2 \rangle - \langle x \rangle^2$

σ^2 = Variance = (std dev)^2

$P(x)$ can be rebuilt by the moments (string of numbers)

$$F(\omega) = \int_{-\infty}^{\infty} \underbrace{e^{-i\omega t}}_{\sim x^n} f(t)$$

Fourier
Laplace

breaking structures into fundamental parts

$\gamma > 1.$

$$\int_{a}^{b} x^n \underbrace{c\,x^{-\gamma}}_{p(x)} \, dx$$

$a \to -\infty$, $b \to \infty$

$$= c \int_a^b x^{-\gamma+n} \, dx$$

← make decaying power law clear

$$= c \cdot \frac{x^{-\gamma+n+1}}{-\gamma+n+1} \Bigg|_a^b$$

$$= c \cdot \frac{x^{-(\gamma-n-1)}}{-(\gamma-n-1)} \Bigg|_a^b$$

$$= \frac{c}{(\gamma-n-1)} \left[a^{-(\gamma-n-1)} - b^{-\overbrace{(\gamma-n-1)}} \right]$$

trouble if $(\gamma-n-1) \le 0$ as $b \to \infty$.

good enough.

where in general: $a <<< b$

$c: n=0$

$x^{-\gamma}$ make decay very clear

everything revolves around this term:

$$b^{-(\gamma-n-1)}$$

[γ is fixed
vary n.

$$\lim_{b\to\infty} b^{\underbrace{-(\gamma-n-1)}} = \begin{cases} 0 & \text{if } (\gamma-n-1)>0 \\ \infty & \text{if } (\gamma-n-1)\leq 0 \end{cases}$$

ok in $(\gamma-n-1)>0$

disaster if $(\gamma-n-1)\leq 0$

don't worry about

full effort requires looking at $\gamma-n-1=0$ case

$$(37 + x^2)$$

$$37\left(1 + \frac{1}{37}x^2\right)$$

$$\text{const}\,(1 + \sim\sim)$$

when thinking about asymptotics.

- just good form.
- will help over and over

Asst 1. Normal.

$$1 = \int_a^b p(x)\,dx = \frac{c}{\gamma-1}\left(a^{-(\gamma-1)} - b^{-(\gamma-1)}\right).$$

← require that $p(x)$ is a prob distr...

$$c = \frac{\gamma-1}{a^{-(\gamma-1)} - b^{-(\gamma-1)}}$$

$$\lim_{b\to\infty} c = \lim_{b\to\infty} \frac{\gamma-1}{a^{-(\gamma-1)} - \cancel{b^{-\overbrace{(\gamma-1)}^{\text{always +ve}}}}} = \frac{\gamma-1}{a^{-(\gamma-1)}}$$

$\to 0$

$\gamma > 1$

$$= (\gamma-1)a^{\gamma-1}$$

$$7,\ 3,\ 3,\ 1,\ 1,\ 1,\ 1 \xrightarrow{\text{mean}} \frac{1}{7}\left(\overset{1.7}{7}+\overbrace{3+3}^{2.3}+\overbrace{1+1+1+1}^{4.1}\right) = \langle k \rangle$$

7 1

Z 2

$$\left.\begin{aligned} N_7 &= 1 \\ N_3 &= 2 \\ N_1 &= 4 \end{aligned}\right\}$$

thing

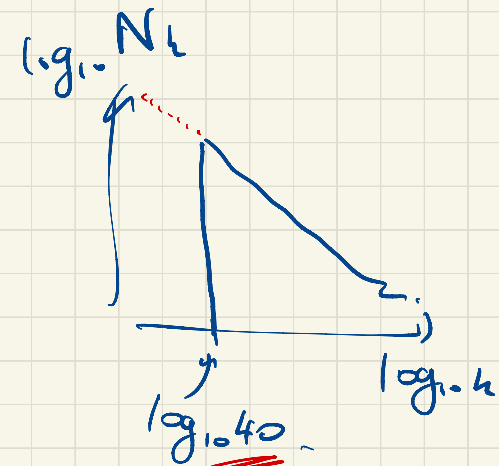
weight

$$\frac{\sum\limits_{k=1} k N_k}{\underbrace{\sum\limits_{k'} N_{k'}}_{\text{total \# of things}}} = \langle k \rangle$$

$$P_k = \frac{N_k}{\sum\limits_{k'} N_{k'}}$$

normalised

$$\langle x \rangle = \int_{-a}^{a} \underset{\text{metres}}{x}\, \overbrace{P(x)}^{\text{number}}\, dx$$



$$\sigma^2 = \text{Variance} = \underbrace{\langle k^2 \rangle}_{\text{second moment}} - \underbrace{\langle k \rangle^2}_{\text{first moment squared}}$$

$$\langle k^2 \rangle = \frac{\sum_k k^2 N_k}{\sum_k N_k}$$

definition

$$= \int_{-\infty}^{\infty} (x - \overset{\text{mean}}{\underset{\text{const}}{\mu}})^2 \, p(x)\, dx \qquad \mu \equiv \langle x \rangle$$

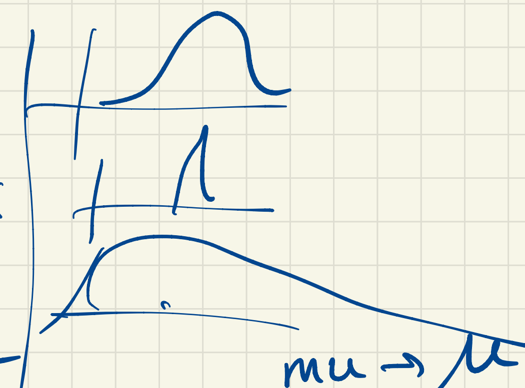
(b/c $\int_{-\infty}^{\infty} |x-\mu| \, p(x)\,dx$ is not "fun" mathematical)

$$\| $$

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

$$= \underbrace{\int_{-\infty}^{\infty} x^2 p(x)\, dx}_{\langle x^2 \rangle} - 2\mu \underbrace{\int_{-\infty}^{\infty} x\, p(x)\, dx}_{\mu} + \mu^2 \underbrace{\int_{-\infty}^{\infty} 1 \cdot p(x)\, dx}_{1}$$

$$= \langle x^2 \rangle \overset{-\mu^2}{\cancel{-2\mu^2 + \mu^2}} = \langle x^2 \rangle - \langle x \rangle^2$$



mu → μ

$\bar{x}$, μ, $\langle x \rangle$, $\langle x^1 \rangle$, μ_x

$\mathbb{A}$, $\mathbb{B}$ $\qquad$ $\mathbb{U}\,\mathbb{\Sigma}\,\mathbb{V}^T$

$\qquad\qquad\qquad\qquad\qquad\qquad \mathbb{\Lambda}$

$\vec{x}$, $\vec{y}$, $\vec{z}$

a

```
\newcommand{\mathexponent}{a}
$ x^{\mathex      t}
```

P o C S What's The Story?

Principles of Complex Systems, Vols. 1, 2, & 3D
CSYS/MATH 300, 303, & 394
University of Vermont, Fall 2022
Assignment 04
++?????++ Out of Cheese Error. Redo From Start

Due: Friday, September 23, by 11:59 pm
https://pdodds.w3.uvm.edu/teaching/courses/2022-2023pocsverse/assignments/04/
Some useful reminders:
Deliverator: Prof. Peter Sheridan Dodds (contact through Teams)
Assistant Deliverator: Dylan Casey (contact through Teams)
Office: The Ether
Office hours: TBD
Course website: https://pdodds.w3.uvm.edu/teaching/courses/2022-2023pocsverse

---

All parts are worth 3 points unless marked otherwise. Please show all your workingses clearly and list the names of others with whom you ~~conspired~~ collaborated.

For coding, we recommend you improve your skills with Python, R, and/or Julia. The (evil) Deliverator uses (evil) Matlab.

Graduate students are requested to use LaTeX (or related TeX variant). If you are new to LaTeX, please endeavor to submit at least n questions per assignment in LaTeX, where n is the assignment number.

Assignment submission:

Via Blackboard.

For Q1–5, you'll further explore the Google data set you examined earlier.

Q6 prepares for allotaxonometry.

1. Plot the complementary cumulative distribution function (CCDF).

2. Using standard linear regression, measure the exponent $\gamma - 1$ where γ is the exponent of the underlying distribution function. Identify and use a range of frequencies for which scaling appears consistent. Report the 95% confidence interval for your estimate.

   You will find two scaling regimes—please examine them both.

3. Size-rank plots:

Using the alternate data set providing the raw word frequencies, plot word frequency as a function of rank in the manner of Zipf.

Hint: you will not be able to plot all points (there are close to 14 million) so think about how to plot a subsample that still shows the full form.

4. Using standard linear regression, measure α, Zipf's exponent. Report the 95% confidence interval for your estimate.

   using your preferred software

   Again, you will find two regimes.

5. For each scaling regime, write down how γ and α are related (per lectures) and check how this expression works for your estimates here.

6. (3 + 3) **Baby name frequencies in the US:**

   Note: We will use this data set again in the next assignment.

   (a) Plot the Complementary Cumulative Frequency Distributions and Size-rank plots (Zipf's law) for the following:

      i. Baby girl names in 1952.
      ii. Baby boy names in 1952.
      iii. Baby girl names in 2002.
      iv. Baby boy names in 2002.

      Note that you will have counts that will make the Zipf distribution easy to plot straight away.

      From these counts, you will have to create the distributions N_k and $N_{\geq k}$.

   (b) As you did for the Google data set, fit regression lines and report values of γ and the Zipf exponent α.

      BUT: Only fit lines if fitting lines make sense!

      You may only have one region of scaling or zero.

   We will revisit these distributions in following assignments.

   Download:

   Data for 1880 through 2018:

   http://pdodds.w3.uvm.edu/permanent-share/pocs-babynames.zip (8.0M)

   Files:

   For each year, Zipf distribution of counts are stored in: `names-girlsYYYY.txt` and `names-boyYYYY.txt`.

For normalization to estimate rates, total number of births per year: `births_per_year.txt`. For this question, you do not need to determine rates, and this file is included for completeness.

For privacy, names with less than 5 counts are excluded.

The rare are legion and, for baby names, hidden.

Notes:

You should be able to re-use scripts from previous assignments.

Data is based on names registered through Social Security within the US.

Source:

Baby name dataset available here: https://catalog.data.gov/dataset?tags=baby-names. Separate dataset for total births available here: https://ssa.gov/oact/babynames/numberUSbirths.html.

7. Install Matlab on your machine of choice.

   We'll use Matlab in the next assignment.

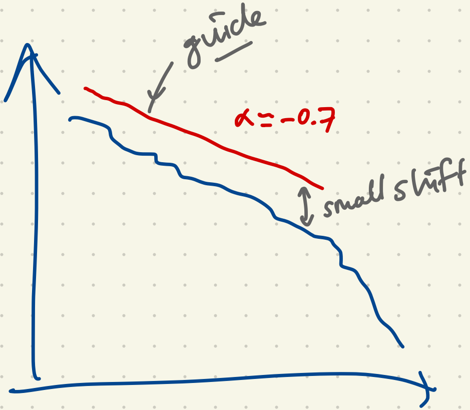
   3 points for just getting this done and reporting faithfully that you did succeed.

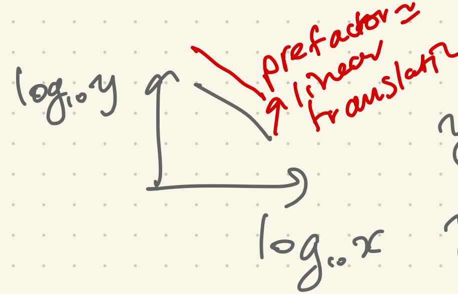
   If you already have Matlab, then this is a freebie.

   For what we're going to do, there will be some roadblocks if you are not using a UNIX machine.

   Linux machines will work, Apple/Mac will work (because Mac OS is UNIX underneath), but Windows will present some problems.

   Note: For future happiness, we encourage students to use Python in general. Maybe R. Julia too. We're using a complicated piece of machinery that only exists fully realized in Matlab.





$$y = x^{-\alpha}$$

$$y' = 10x^{-\alpha}.$$

In principle

$\{N_k\}$, $N_1, N_2, \ldots$ ← not normalized (partial data set $k \geq \cancel{40}$ 200).

vs

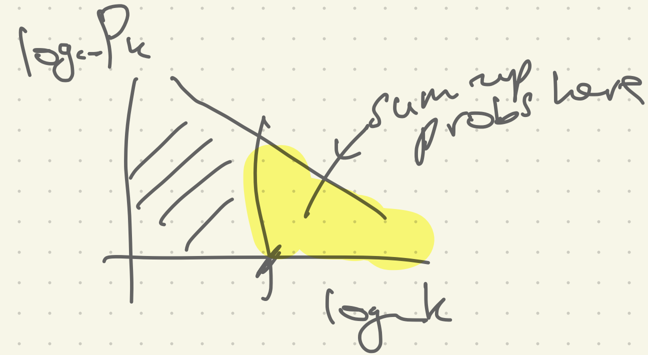
$\{p_k\}$, $p_1, p_2,$

↑ normalized

for Google data set

$N_{\cancel{40}\,200}, N_{\cancel{41}\,201}, \ldots$

$k_{min} = \cancel{40}$ 200



CCDF — $P_{\geq k} = \sum_{j=k}^{\infty} p_j$

↑ decreases with k.

$N_{\geq k} = \sum_{j=k}^{\text{"}\infty\text{"}} N_j$

↑ just counts

$$\text{counts} = [N_{200}, N_{201}, \ldots\ldots, N_{k-1}, N_k, N_{k+1}, \ldots\ldots, N_{kmax}]$$

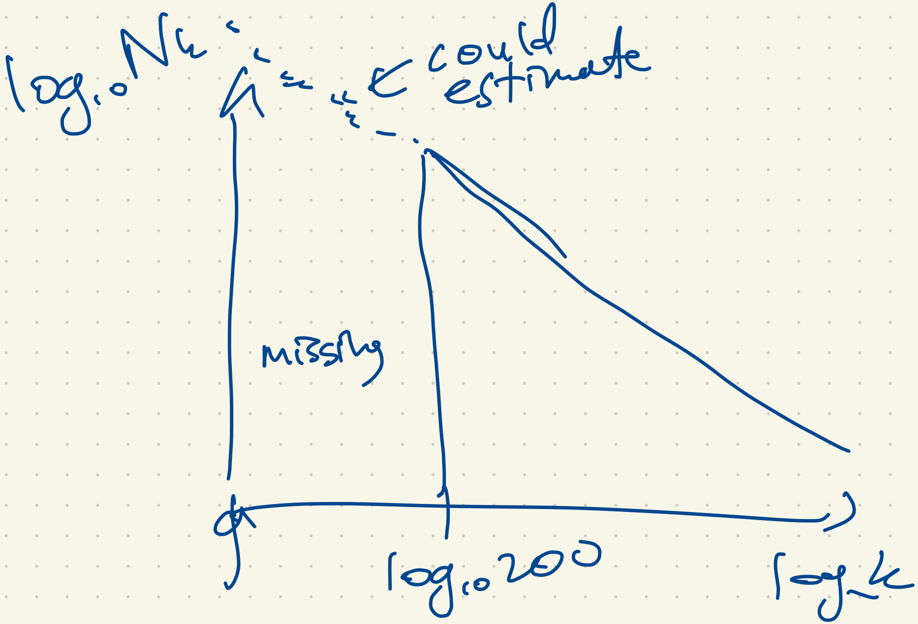
```
for j = 1 : length(counts)
    CCDF(j) = sum(counts(j:end));
end
```

$$N_{\text{occurences}} = [200, 201, \ldots, \ldots]$$

j=1   j=2 . - - -

Faster:

```
CCDF(1) = sum(counts).
for j = 2 : length(counts)
    CCDF(j) = CCDF(j-1) - counts(j-1);
end
```



P_k only possible if we know

$$\sum_{j=1} N_j = \# \text{ types}$$

CCDF   CDF

$$P_{\geq k} = 1 - P_{<k} \quad \checkmark$$

$$\cancel{N_{\geq k} \stackrel{?}{=} 1 - N_{<k}}$$

$N_1 = 7$

$N_2 = 3$

$N_3 = 1$

$N_7 = 1$

$N_{occurrences} = [1 \quad 2 \quad 3 \quad 7]$

$count = [7 \quad 3 \quad 1 \quad 1]$

CCDF

$N_{\geq 1} = N_1 + \underline{N_2 + N_3} + N_7$

$N_{\geq 2} = N_2 + \underline{N_3} + N_7$

$N_{\geq 3} = \underline{N_3 + N_7}$

$N_{\geq 4} = \underline{N_7}$

$N_{\geq 5} = \underline{N_7}$

$N_{\geq 6} = N_7$

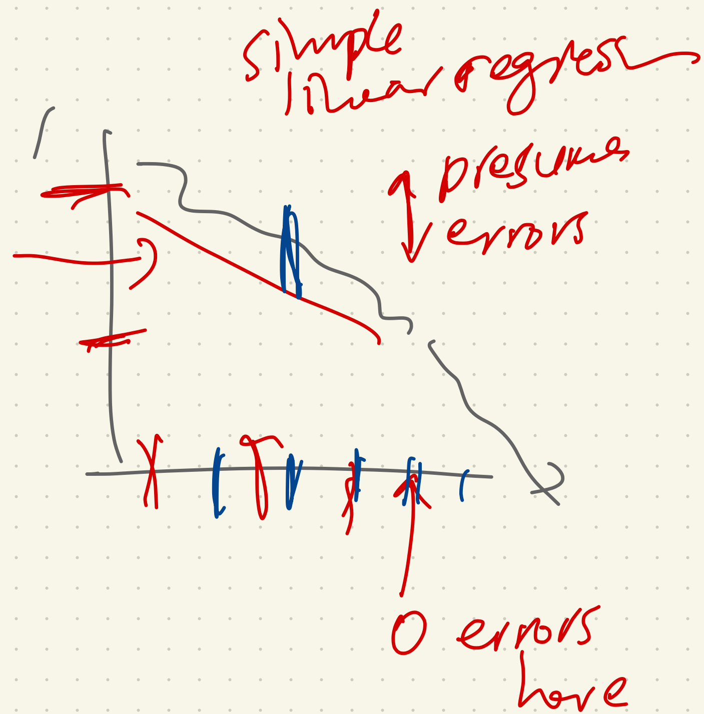
$N_{\geq 7} = N_7$

$N_{\geq 8} = 0.$

~~<s> text ... </s>~~

$$\alpha = \frac{1}{\gamma - 1}$$

$$\gamma > 1$$



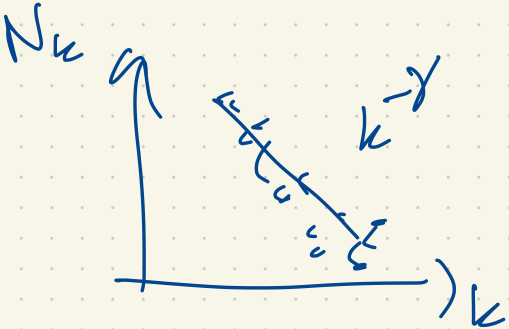
$$\begin{cases} y = mx + c = ax + b \\ x = \frac{1}{m} y - \frac{c}{m} = \alpha y + \beta \end{cases}$$

actual measurement: $a = \frac{1}{\alpha} r^2$ correl coeff.

$N_k \sim k^{-\gamma}$

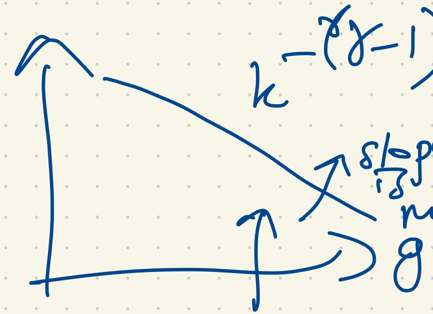
$N_{\geq k} \sim k^{-\underbrace{(\gamma-1)}_{\simeq \frac{2}{3}}}$

$\gamma = \frac{2}{3} + 1$



$\Longrightarrow$

Integral smooths + makes a heavier tail



slope is more gradual

$\alpha = \frac{1}{\gamma - 1}$

$\gamma > 1$ PDF slope

$\Updownarrow$

$\underline{\gamma - 1} > 0$ CCDF slope

Principles of Complex Systems, Vols. 1, 2, & 3D
CSYS/MATH 300, 303, & 394
University of Vermont, Fall 2022
Assignment 05
SQUEAK

Due: Friday, September 30, by 11:59 pm
https://pdodds.w3.uvm.edu/teaching/courses/2022-2023pocsverse/assignments/05/
Some useful reminders:
Deliverator: Prof. Peter Sheridan Dodds (contact through Teams)
Assistant Deliverator: Dylan Casey (contact through Teams)
Office: The Ether
Office hours: See Teams calendar
Course website: https://pdodds.w3.uvm.edu/teaching/courses/2022-2023pocsverse

---

All parts are worth 3 points unless marked otherwise. Please show all your workingses clearly and list the names of others with whom you ~~conspired~~ collaborated.

For coding, we recommend you improve your skills with Python, R, and/or Julia. The (evil) Deliverator uses (evil) Matlab.

Graduate students are requested to use LaTeX (or related TeX variant). If you are new to LaTeX, please endeavor to submit at least n questions per assignment in LaTeX, where n is the assignment number.

Assignment submission:

Via Blackboard.

Notes for Baby Name analysis:

- You will have the data sets on hand from the previous assignment.
- Unix systems will work (Linux, the Apple things, etc.). Windows may not (alternative given below).
- As is, you will need the command epstopdf. Please install if not on deck already.

1. Plot time series for the rank of these baby names in the US over all years in the census data:

    - Shirley.

- Desmond.
- Madison.
- Aiden.
- A name of your choice.

Note that if you plotted relative frequency rather than rank, you would need to know (or estimate) the overall number of babies born. Ranks are both easy simple to work with and easy to understand.

2. Only this question requires Matlab. See alternative below if you cannot get the allotaxonometer to work on your system.

   Generate allotaxonographs comparing the following four pairs:

   (a) Baby girl names in 1952 versus baby girl names in 2002.
   (b) Baby boy names in 1952 versus baby boy names in 2002.
   (c) Baby girl names in 1952 versus baby boy names in 1952.
   (d) Baby girl names in 2002 versus baby boy names in 2002.

   Use rank-turbulence divergence with $\alpha = 0$ and $\alpha = \infty$.

   Online appendices for main papers is here:
   http://compstorylab.org/allotaxonometry/.

   Contains overview, examples, links to papers, figure-making code, etc.

   Alternative:

   Using rank-turbulence divergence with $\alpha = 0$ and ∞, list the top 30 contributing baby names for the four comparisons listed above.

   Indicate which year each contributing baby name comes from in parentheses.

   For ordering, you do not need to compute RTD in full but rather just the core structure:

$$\left| \frac{1}{[r_{\tau,1}]^{\alpha}} - \frac{1}{[r_{\tau,2}]^{\alpha}} \right| . \tag{1}$$

   Recall that for $\alpha = 0$ and $\alpha = \infty$, the essential core structure becomes:

$$\left| \ln \frac{r_{\tau,1}}{r_{\tau,2}} \right| \text{ and } \max_{\tau} \left\{ \frac{1}{r_{\tau,1}}, \frac{1}{r_{\tau,2}} \right\} . \tag{2}$$

3. Everyday random walks and the Central Limit Theorem:

   Show that the observation that the number of discrete random walks of duration $t = 2n$ starting at $x_0 = 0$ and ending at displacement $x_{2n} = 2k$ where $k \in \{0, \pm 1, \pm 2, \ldots, \pm n\}$ is

   $$N(0, 2k, 2n) = \binom{2n}{n+k} = \binom{2n}{n-k}$$

   leads to a Gaussian distribution for large $t = 2n$:

   $$\mathbf{Pr}(x_t \equiv x) \simeq \frac{1}{\sqrt{2\pi t}} e^{-\frac{x^2}{2t}}.$$

   Please note that $k \ll n$.

   Stirling's sterling approximation will prove most helpful.

   benign and friendly

   dangerous stuff

   Hint: You should be able to reach this form:

   "$\frac{1}{\sqrt{2\pi}\sigma}$"

   $$\frac{\text{Some stuff not involving penguins}}{\text{Some other penguin-free stuff} \times (1 - k^2/n^2)^{n+1/2}(1 + k/n)^k(1 - k/n)^{-k}}.$$

   k is fixed
   n is large increasing

   "$e^{-\frac{x^2}{2\sigma^2}}$"

   Lots of sneakiness here. You'll want to examine the natural log of the piece shown above, and see how it behaves for large n.

   You may very well need to use the Taylor expansion $\ln(1 + z) \simeq z$.

   Exponentiate and carry on.

   Tip: If at any point penguins appear in your expression, you're in real trouble. Get some fresh air and start again.

4. From lectures, show that the number of distinct 1-d random walk that start at $x = i$ and end at $x = j$ after t time steps is

   $$N(i, j, t) = \binom{t}{(t + j - i)/2}.$$

   Assume that j is reachable from i after t time steps.

   Hint—Counting random walks:
   http://www.youtube.com/watch?v=daSIYz-0U3E

5. *Discrete random walks:*

   In class, we argued that the number of random walks returning to the origin for the first time after $2n$ time steps is given by

   $$N_{\text{first return}}(2n) = N_{\text{fr}}(2n) = N(1, 1, 2n - 2) - N(-1, 1, 2n - 2)$$

where

$$N(i, j, t) = \binom{t}{(t + j - i)/2}.$$

Find the leading order term for $N_{\mathrm{fr}}(2n)$ as $n \to \infty$.

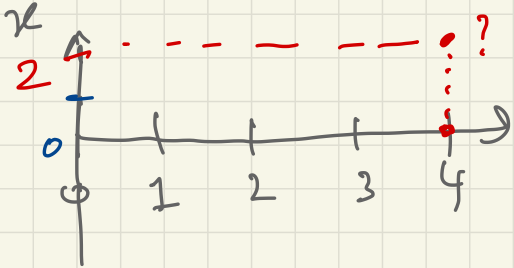
Two-step approach:

(a) Combine the terms to form a single fraction,

(b) and then again use Stirling's bonza approximation.

If you enjoy this sort of thing, you may like to explore the same problem for random walks in higher dimensions. Seek out George Pólya.

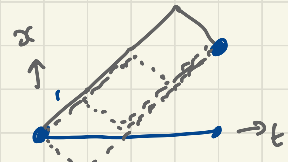
And we are connecting to much other good stuff in combinatorics; more to come in the solutions.



$N(i, j, t)$

$N(0, 2, 4)$

start finish time taken



1	2	3	4		1	2	3	4
+1	+1	+1	−1	= +2	R	R	R	L
−1	+1	+1	+1	= +2	L	R	R	R
+1	−1	+1	+1	= +2	R	L	R	R
+1	+1	−1	+1	= +2	R	R	L	R

$3\times(+1)+1(-1)$

$3\times(R)+1(L)=2R$

1,2,3 | 4
1 | 2,3,4
2 | 1,3,4
3 | 1,2,4

$$\binom{4}{3}\equiv\binom{4}{1}=4=\frac{4!}{3!\,1!}$$

$x = i$ and end at $x = j$ after t time steps is

$$N(i,j,t) = \binom{t}{(t+j-i)/2}.$$

Assume that j is reachable from i after t time steps.

Hint—Counting random walks:
http://www.youtube.com/watch?v=daSIYz-0U3E

5. *Discrete random walks:*

In class, we argued that the number of random walks returning to the origin for the first time after $2n$ time steps is given by

$$N_{\text{first return}}(2n) = N_{\text{fr}}(2n) = N(1,1,2n-2) - N(-1,1,2n-2)$$

$$N(1,1,2n-2) - N(-1,1,2n-2)$$

$$= \binom{2n-2}{\frac{2n-2+\cancel{1-1}}{2}} - \binom{2n-2}{\frac{\cancel{2}n\cancel{-2}+\cancel{1-(-1)}}{\cancel{2}}}$$

$$= \binom{2n-2}{n-1} - \binom{2n-2}{n} \qquad \binom{n}{k} = \frac{n!}{k!\,(n-k)!}$$

$$= \frac{(2n-2)!}{(n-1)!\,(n-1)!} - \frac{(2n-2)!}{n!\,(n-2)!} \qquad (n-2)! = \frac{(n-1)!}{(n-1)} \qquad n! = n\times(n-1)!$$

$$= \underbrace{\frac{(2n-2)!}{(n-1)!\,(n-1)!}}_{\text{big mess}} \underbrace{\left(1 - \frac{n-1}{n}\right)}_{\text{simple}}.$$

$$= \frac{(2n-2)!}{(n-1)!\,(n-1)!}\left(\frac{\cancel{n} - (\cancel{n}-1)}{n}\right) \qquad +1 \qquad \frac{1}{n}$$

$$= \frac{1}{n}\frac{(2n-2)!}{(n-1)!\,(n-1)!} \qquad \text{well organized}$$

useful way to write Stirling approx:

$$n! \sim \sqrt{2\pi}\, n^{n+\frac{1}{2}} e^{-n}$$

instead of ~~$\sqrt{2\pi n}\left(\frac{n}{e}\right)^n$~~ no.

$$\omega! \sim \sqrt{2\pi}\,\omega^{\omega+\frac{1}{2}} e^{-\omega}$$

$\omega = 2n-2$

$\omega = n-1$

$$\lim_{n\to\infty} \frac{1}{n}\frac{(2n-2)!}{(n-1)!\,(n-1)!}$$

$$\sim \frac{1}{n}\frac{\cancel{\sqrt{2\pi}}\,(2n-2)^{2n-2+\frac{1}{2}}\cancel{e^{-(2n-2)}}}{\cancel{\sqrt{2\pi}}\,(n-1)^{n-1+\frac{1}{2}}\cancel{e^{-(n-1)}}\sqrt{2\pi}\,(n-1)^{n-1+\frac{1}{2}}\cancel{e^{-(n-1)}}}$$

$$= \frac{1}{n}\frac{1}{\sqrt{2\pi}}\frac{(\overbrace{2n-2}^{2(n-1)})^{2n-2+\frac{1}{2}}}{(n-1)^{2n-2+1}} \qquad (ab)^\phi = a^\phi b^\phi$$

$$= \frac{1}{n}\frac{1}{\sqrt{2\pi}}\frac{2^{\overbrace{2n-2+\frac{1}{2}}^{-3/2}}\,(\cancel{n-1})^{\cancel{2n-2+\frac{1}{2}}}}{(n-1)^{\cancel{2n-2+1}\,\frac{1}{2}}}$$

$$= \frac{1}{\sqrt{2\pi}}\frac{1}{n}\frac{2^{2n-3/2}}{(n-1)^{1/2}} \quad n \text{ is very large...?}$$

$$P_{fr}(t) = \frac{1}{2^t}N_{fr}(t)$$

$$= \frac{1}{\sqrt{2\pi}}\frac{1}{n}\frac{2^{2n-3/2}}{n^{1/2}\left(1-\frac{1}{n}\right)^{1/2}} \qquad \to 1 \text{ as } n\to\infty$$

good form: $1+\epsilon+\epsilon^2+\ldots$

c.f. $\lim_{n\to\infty}\left(1+\frac{a}{n}\right)^n = e^a$

$$= \frac{1}{\sqrt{2\pi}}\frac{2^{2n-3/2}}{n^{3/2}}$$

back to lecture slides.

$$= \frac{1}{\sqrt{2\pi}}\frac{2^{2n}}{(2n)^{3/2}} \leftarrow \text{\# of possible walks}$$

$\sim t$

$$P(t) \sim \frac{1}{\sqrt{2\pi}}t^{-3/2}$$

for t large.

$\boxed{1 + a_1 x} + a_2 x^2 + \cdots$

not involving penguins

$\times\ (1 - k^2/n^2)^{n+1/2}(1 + k/n)^k(1 - k/n)^{-k}.$

nt to examine the natural log of the piece shown

k is fixe

n is

more than one way to do this!!

$n \gg k.$

notes

- expand as a product
- Use $(1+\epsilon)^k \simeq (1 + k\epsilon + \binom{k}{2}\epsilon^2 + \cdots)$

Monk

* Very helpful: Consider ln of expression

undo log

$\sim e^{+k^2/n}$

$$\ln\left(\left(1-\frac{k^2}{n^2}\right)^{n+\frac{1}{2}}\left(1+\frac{k}{n}\right)^k\left(1-\frac{k}{n}\right)^{-k}\right)$$

$$= \left(n+\tfrac{1}{2}\right)\ln\left(1-\frac{k^2}{n^2}\right) + k\ln\left(1+\frac{k}{n}\right) - k\ln\left(1-\frac{k}{n}\right)$$

$$\left(n+\tfrac{1}{2}\right)\left(-\frac{k^2}{n^2}\right) + k\cdot\frac{k}{n} - k\left(-\frac{k}{n}\right) = +\frac{k^2}{n} - \cancel{\frac{1}{2}\frac{k^2}{n^2}}$$

small

$$\frac{1}{1-\omega} \simeq 1+\omega+\omega^2+\omega^3+\cdots$$

$$\frac{1}{1+\omega} = 1-\omega+\omega^2-\omega^3+\cdots$$

$$\int \frac{1}{1+x}\,dx = \ln(1+x) = \frac{x}{1} - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \cdots$$

$$\ln(1+z) \simeq z.$$

$$\ln(1+\omega) \simeq \omega.$$