Will Monroe
July 14, 2017

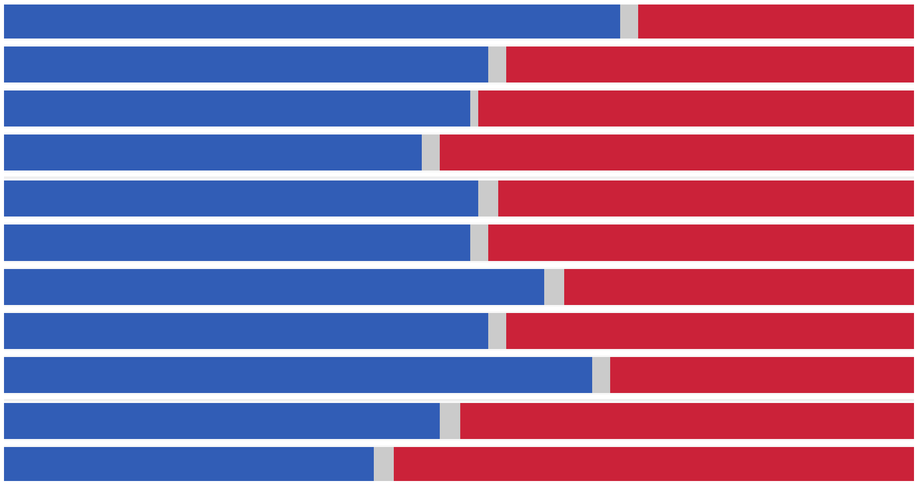
with materials by
Mehran Sahami
and Chris Piech

More discrete distributions

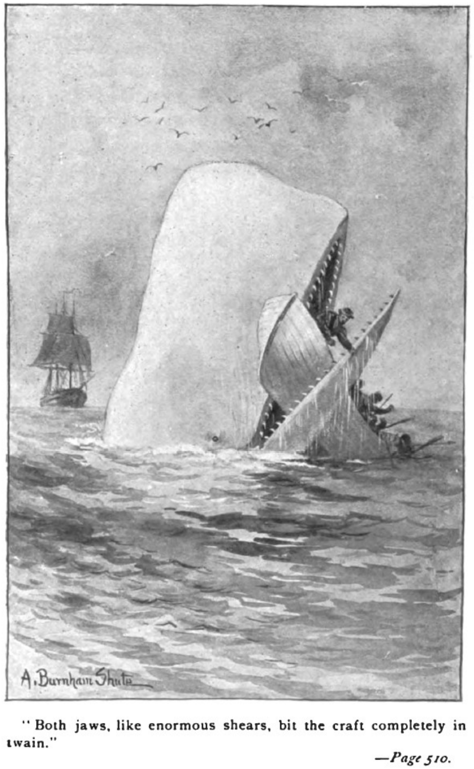
Announcements: Problem Set 3

Posted yesterday on the course website.

Due **next Wednesday**, 7/19, at 12:30pm (before class).



(election prediction)



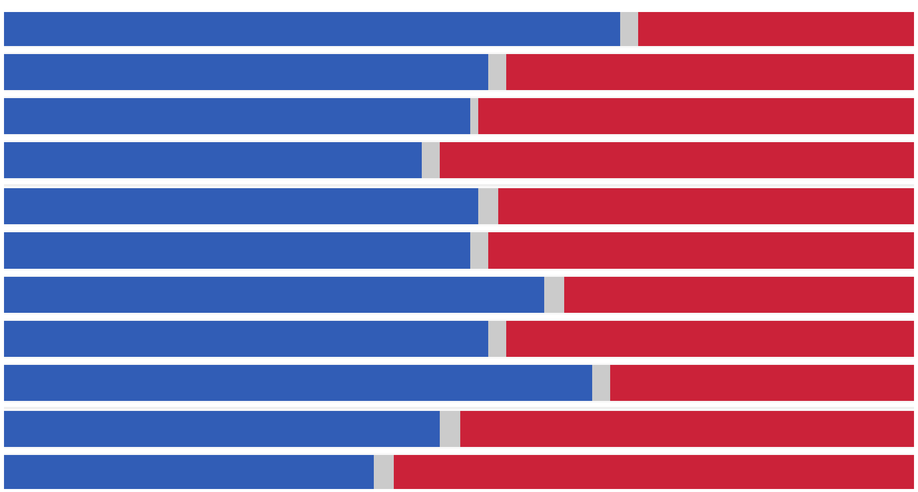
(Moby Dick)

Announcements: Problem Set 3

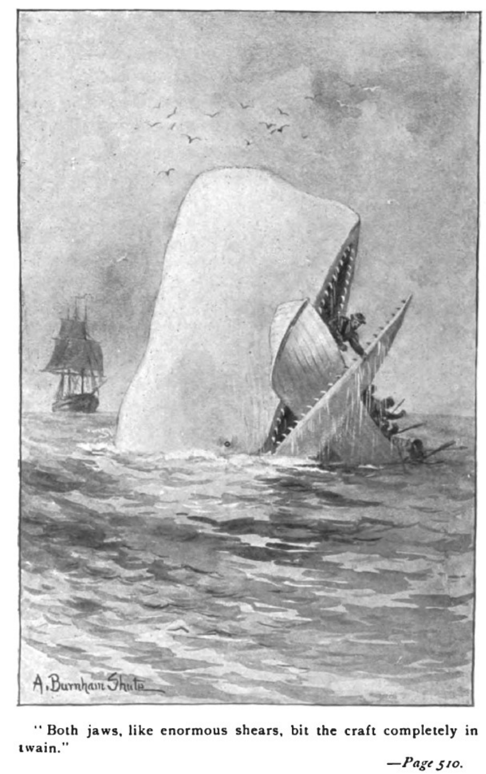
Posted yesterday on the course website.

Due **next Wednesday**, 7/19, at 12:30pm (before class).

Everybody gets an extra late day! (4 total)



(election prediction)



(Moby Dick)

Review: Bernoulli random variable

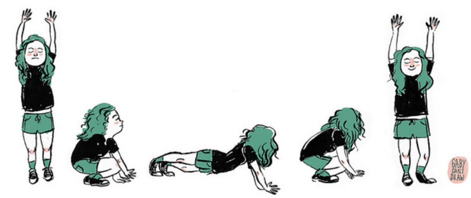
An indicator variable (a possibly biased coin flip) obeys a **Bernoulli distribution**. Bernoulli random variables can be 0 or 1.

$$X \sim \text{Ber}(p)$$

$$p_X(1) = p$$
$$p_X(0) = 1 - p \qquad \text{(0 elsewhere)}$$

Review: Bernoulli fact sheet

$$X \sim \text{Ber}(p)$$



?

↑ probability of "success" (heads, ad click, ...)

PMF:

$$p_X(1) = p$$

$$p_X(0) = 1 - p \qquad \textbf{(0 elsewhere)}$$

expectation: $E[X] = p$

variance: $\text{Var}(X) = p(1 - p)$

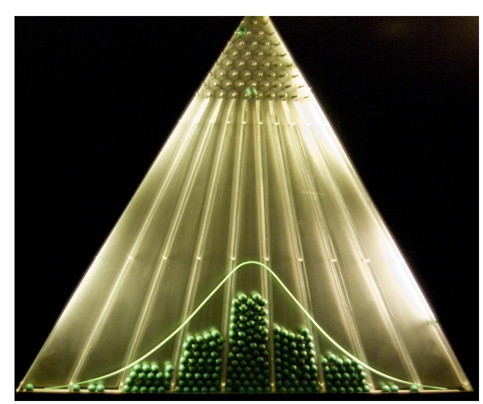
image (right): Gabriela Serrano

Review: Binomial random variable

The **number of heads** on *n* (possibly biased) coin flips obeys a **binomial distribution**.

$$X \sim \mathrm{Bin}(n, p)$$

$$p_X(k) = \begin{cases} \binom{n}{k} p^k (1-p)^{n-k} & \text{if } k \in \mathbb{N}, 0 \le k \le n \\ 0 & \text{otherwise} \end{cases}$$



Review: Binomial fact sheet

number of trials (flips, program runs, ...)

$$X \sim \text{Bin}(n, p)$$

probability of "success" (heads, crash, ...)

PMF: $$p_X(k) = \begin{cases} \binom{n}{k} p^k (1-p)^{n-k} & \text{if } k \in \mathbb{N}, 0 \le k \le n \\ 0 & \text{otherwise} \end{cases}$$

expectation: $E[X] = np$

variance: $\text{Var}(X) = np(1-p)$

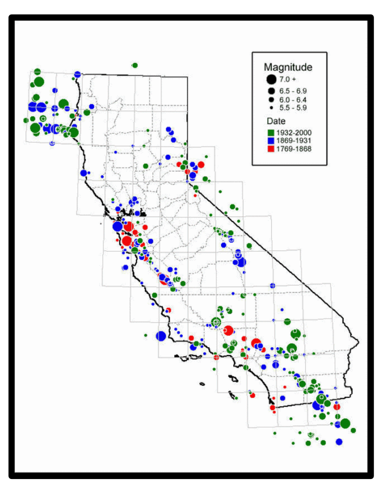
note: $\text{Ber}(p) = \text{Bin}(1, p)$

Review: Poisson random variable

The **number of occurrences** of an event that occurs with **constant rate** λ (per unit time), in 1 unit of time, obeys a **Poisson distribution**.

$$X \sim \text{Poi}(\lambda)$$

$$p_X(k) = \begin{cases} e^{-\lambda} \dfrac{\lambda^k}{k!} & \text{if } k \in \mathbb{Z}, k \geq 0 \\ 0 & \text{otherwise} \end{cases}$$



Review: Poisson fact sheet

$$X \sim \text{Poi}(\lambda)$$

rate of events (requests, earthquakes, chocolate chips, ...)
per unit time (hour, year, cookie, ...)

PMF: $$p_X(k) = \begin{cases} e^{-\lambda} \dfrac{\lambda^k}{k!} & \text{if } k \in \mathbb{Z}, k \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

expectation: $$E[X] = \lambda$$

variance: $$\text{Var}(X) = \lambda$$

Geometric random variable

The **number of trials** it takes to get **one success**, if successes occur independently with probability *p*, obeys a **geometric distribution**.

$$X \sim \text{Geo}(p)$$

$$p_X(k) = \begin{cases} (1-p)^{k-1} \cdot p & \text{if } k \in \mathbb{Z}, k \geq 1 \\ 0 & \text{otherwise} \end{cases}$$

Catching Pokémon

Wild Pokémon are captured by throwing Poké Balls at them.

Each ball has probability *p* of capturing the Pokémon.
How many are needed **on average** for a successful capture?

X: number of Poké Balls until (and including) capture

C_i: event that Pokémon is captured on the *i*-th throw

$$
\begin{aligned}
P(X=k) &= P\left(C_1^C C_2^C \ldots C_{k-1}^C C_k\right) \\
&= P\left(C_1^C\right) P\left(C_2^C\right) \ldots P\left(C_{k-1}^C\right) P\left(C_k\right) \\
&= (1-p)^{k-1} p
\end{aligned}
$$

Geometric: Fact sheet

$$X \sim \text{Geo}(p)$$

probability of "success" (catch, heads, crash, ...)

PMF:

$$p_X(k) = \begin{cases} (1-p)^{k-1} \cdot p & \text{if } k \in \mathbb{Z}, k \geq 1 \\ 0 & \text{otherwise} \end{cases}$$

Catching Pokémon

X: number of Poké Balls until (and including) capture

$$P(X=k)=(1-p)^{k-1}\cdot p$$

$$\begin{aligned}
E[X] &= \sum_{k=1}^{\infty} k\cdot(1-p)^{k-1}\cdot p \\
&= \sum_{k=1}^{\infty} (k-1+1)\cdot(1-p)^{k-1}\cdot p \\
&= \sum_{k=1}^{\infty} (k-1)\cdot(1-p)^{k-1}\cdot p + \sum_{k=1}^{\infty} (1-p)^{k-1}\cdot p \\
&= \sum_{j=0}^{\infty} j\cdot(1-p)^{j}\cdot p + \sum_{j=0}^{\infty} (1-p)^{j}\cdot p \\
&= (1-p)\boxed{\sum_{j=0}^{\infty} j\cdot(1-p)^{j-1}\cdot p} + p\cdot\sum_{j=0}^{\infty} (1-p)^{j} \\
&= (1-p)E[X] + \cancel{p}\cdot\frac{1}{\cancel{1-(1-p)}} \\
&= (1-p)E[X]+1
\end{aligned}$$

$$\sum_{j=0}^{\infty} x^{j} = \frac{1}{1-x}$$

$$\begin{aligned}
E[X] &= (1-p)E[X]+1 \\
(1-(1-p))E[X] &= 1 \\
pE[X] &= 1 \\
E[X] &= \frac{1}{p}
\end{aligned}$$

Geometric: Fact sheet

$$X \sim \mathrm{Geo}(p)$$

probability of "success" (catch, heads, crash, …)

PMF:

$$p_X(k) = \begin{cases} (1-p)^{k-1} \cdot p & \text{if } k \in \mathbb{Z}, k \geq 1 \\ 0 & \text{otherwise} \end{cases}$$

expectation:

$$E[X] = \frac{1}{p}$$

Catching Pokémon

Wild Pokémon are captured by throwing Poké Balls at them.

Each ball has probability p = 0.1 of capturing the Pokémon. How many are needed so that probability of successful capture is **at least** 99%?

X: number of Poké Balls until (and including) capture

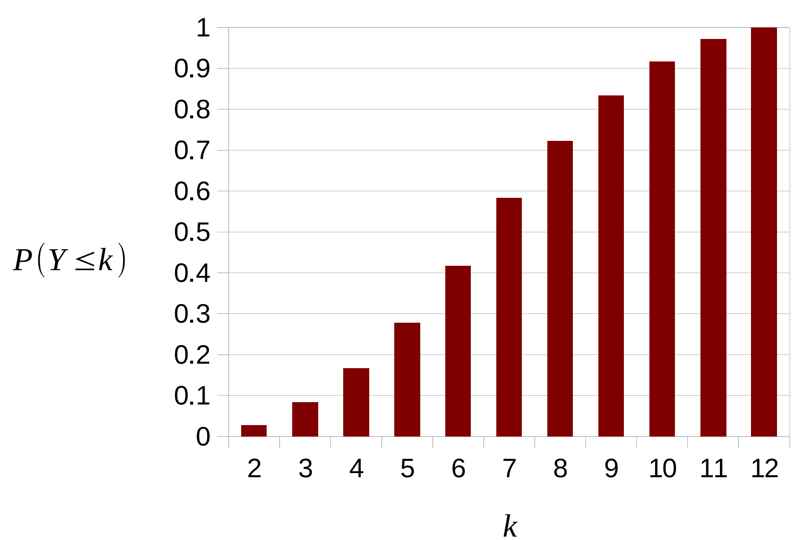
C_i: event that Pokémon is captured on the i-th throw

$$
\begin{aligned}
P(X \le k) &= 1 - P(X > k) \\
&= 1 - P\left(C_1{}^C C_2{}^C \ldots C_k{}^C\right) \\
&= 1 - P\left(C_1{}^C\right) P\left(C_2{}^C\right) \ldots P\left(C_k{}^C\right) \\
&= 1 - (1-p)^k
\end{aligned}
$$

Cumulative distribution function

The **cumulative distribution function** (CDF) of a random variable is a function giving the probability that the random variable is **less than or equal to** a value.

$$F_Y(k) = P(Y \leq k)$$



(CDF of the sum of two dice)

Geometric: Fact sheet

$$X \sim \text{Geo}(p)$$

probability of "success" (catch, heads, crash, ...)

PMF: $$p_X(k) = \begin{cases} (1-p)^{k-1} \cdot p & \text{if } k \in \mathbb{Z}, k \geq 1 \\ 0 & \text{otherwise} \end{cases}$$

CDF: $$F_X(k) = \begin{cases} 1-(1-p)^{k} & \text{if } k \in \mathbb{Z}, k \geq 1 \\ 0 & \text{otherwise} \end{cases}$$

expectation: $$E[X] = \frac{1}{p}$$

Catching Pokémon

Wild Pokémon are captured by throwing Poké Balls at them.

Each ball has probability p = 0.1 of capturing the Pokémon. How many are needed so that probability of successful capture is **at least** 99%?

X: number of Poké Balls until (and including) capture

$$P(X \leq k) = 1 - (1-p)^k \geq 0.99$$

$$0.01 \geq (1-p)^k$$

$$\log 0.01 \geq k \log(1-p)$$

$$43.7 \approx \frac{\log 0.01}{\log(1-p)} \leq k$$

Geometric: Fact sheet

$$X \sim \text{Geo}(p)$$

probability of "success" (catch, heads, crash, ...)

PMF: $$p_X(k) = \begin{cases} (1-p)^{k-1} \cdot p & \text{if } k \in \mathbb{Z}, k \geq 1 \\ 0 & \text{otherwise} \end{cases}$$

CDF: $$F_X(k) = \begin{cases} 1-(1-p)^{k} & \text{if } k \in \mathbb{Z}, k \geq 1 \\ 0 & \text{otherwise} \end{cases}$$

expectation: $$E[X] = \frac{1}{p}$$

variance: $$\text{Var}(X) = \frac{1-p}{p^2}$$

Break time!

Negative binomial random variable

The **number of trials** it takes to get *r* successes, if successes occur independently with probability *p*, obeys a **negative binomial distribution**.

$$X \sim \text{NegBin}(r, p)$$

$$p_X(n) = \begin{cases} \binom{n-1}{r-1} p^r (1-p)^{n-r} & \text{if } n \in \mathbb{Z}, n \geq r \\ 0 & \text{otherwise} \end{cases}$$

Getting that degree

A conference accepts papers (independently and randomly?) with probability p = 0.25.

A hypothetical grad student needs 3 accepted papers to graduate. What is the probability this **takes exactly 10 submissions**?

X: number of tries to get 3 accepts
Y: number of accepts in first 9 tries

accept on 10^{th} try

$$P(X=10)=P(Y=2)\cdot p$$

$$=\binom{9}{2}(1-p)^7 p^2\cdot p\approx 0.075$$

Getting that degree

A conference accepts papers (independently and randomly?) with probability p.

A hypothetical grad student needs r accepted papers to graduate. What is the probability this **takes exactly n submissions**?

X: number of tries to get r accepts
Y: number of accepts in first $n - 1$ tries

accept on n^{th} try

$$P(X = 10) = P(Y = r - 1) \cdot p$$

$$= \binom{n-1}{r-1} (1-p)^{n-r} p^{r-1} \cdot p$$

Negative binomial: Fact sheet

number of **sucesses** (heads, crash, ...)

$$X \sim \mathrm{NegBin}(r, p)$$

number of **trials** (flips, program runs, ...)

probability of "success"

PMF: $$p_X(n) = \begin{cases} \binom{n-1}{r-1} p^r (1-p)^{n-r} & \text{if } n \in \mathbb{Z}, n \geq r \\ 0 & \text{otherwise} \end{cases}$$

Getting that degree

A conference accepts papers (independently and randomly?) with probability p = 0.25.

A hypothetical grad student needs 3 accepted papers to graduate. How many submissions will be necessary on **average**?

X: number of tries to get 3 accepts

$$E[X] = ?$$

A) $3 \cdot 0.25$

B) $\frac{3}{0.25}$

C) $3^{0.25}$

D) 3^4

https://bit.ly/1a2ki4G → https://b.socrative.com/login/student/
Room: CS109SUMMER17

Getting that degree

A conference accepts papers (independently and randomly?) with probability ***p***.

A hypothetical grad student needs ***r*** accepted papers to graduate. How many submissions will be necessary on **average**?

X: number of tries to get ***r*** accepts

$$E[X] = \frac{r}{p}$$

https://bit.ly/1a2ki4G → https://b.socrative.com/login/student/
Room: CS109SUMMER17

Negative binomial: Fact sheet

number of **sucesses** (heads, crash, ...)

$$X \sim \text{NegBin}(r, p)$$

number of **trials** (flips, program runs, ...)

probability of "success"

PMF: $$p_X(n) = \begin{cases} \binom{n-1}{r-1} p^r (1-p)^{n-r} & \text{if } n \in \mathbb{Z}, n \geq r \\ 0 & \text{otherwise} \end{cases}$$

expectation: $$E[X] = \frac{r}{p}$$

variance: $$\text{Var}(X) = \frac{r(1-p)}{p^2}$$

note:

$$\text{Geo}(p) = \text{NegBin}(1, p)$$

A few optional (but hopefully interesting) distributions

(these won’t be on tests or problem sets)

Hypergeometric distribution

balls to draw → n; number of red balls → m

$$X \sim \mathrm{HypG}(n, N, m)$$

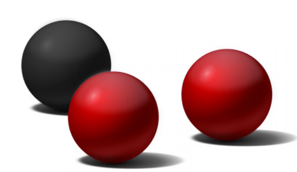
number of red balls drawn **without replacement** → X

total number of balls (black + red) → N

PMF: $$p_X(k) = \begin{cases} \dfrac{\binom{m}{k}\binom{N-m}{n-k}}{\binom{N}{n}} & \text{if } k \in \mathbb{Z}, 0 \le k \le \min(n, m) \\ 0 & \text{otherwise} \end{cases}$$

expectation: $$E[X] = n\frac{m}{N}$$

variance: $$\mathrm{Var}(X) = \frac{nm(N-n)(N-m)}{N^2(N-1)}$$



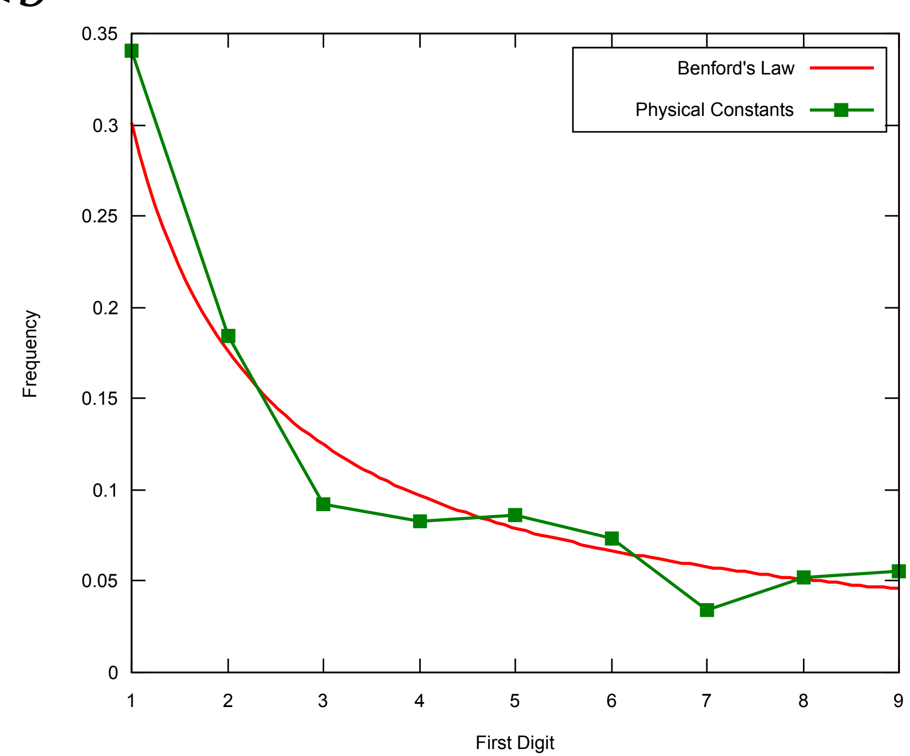
Benford distribution

base of number system (e.g. 10)

$$X \sim \text{Benford}(b)$$

first digit of naturally occurring number

PMF: $$p_X(d) = \begin{cases} \log_b\left(1 + \frac{1}{d}\right) & \text{if } d \in \mathbb{Z}, 0 \le d < b \\ 0 & \text{otherwise} \end{cases}$$



Zipf distribution

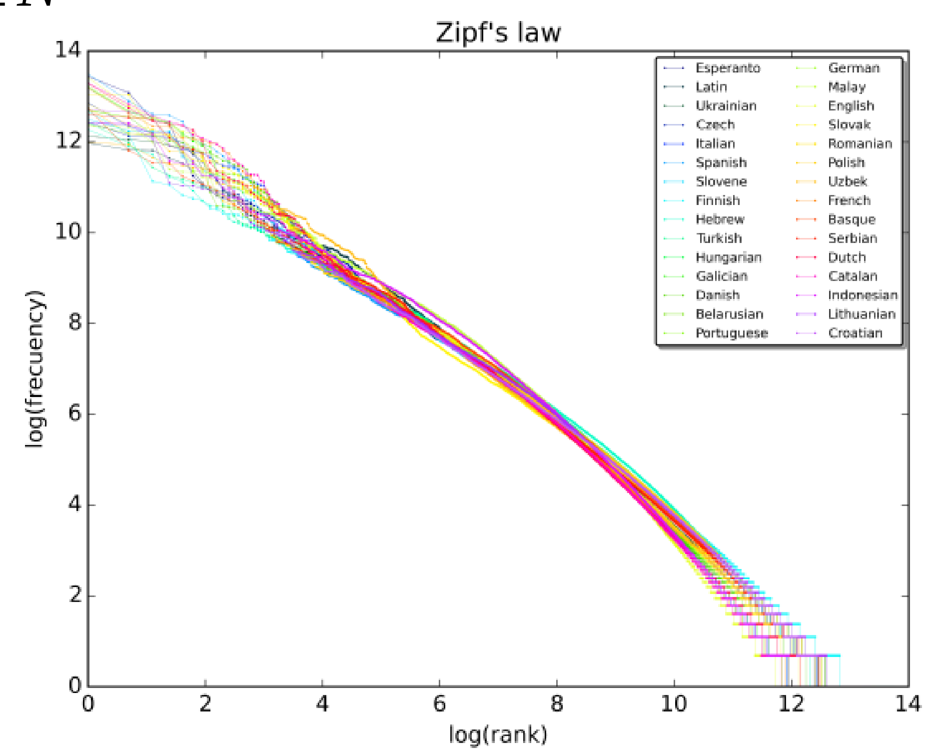
"power law" exponent (often close to 1)

$$X \sim \text{Zipf}(s, N)$$

rank of randomly chosen word

vocabulary size

PMF: $$p_X(k) = \begin{cases} \dfrac{1/k^s}{\sum_{n=1}^{N} (1/n^s)} & \text{if } k \in \mathbb{Z}, 0 \le k \le N \\ 0 & \text{otherwise} \end{cases}$$



A grid of random variables

	number of **successes**	**time** to get successes	
One trial	$X \sim \text{Ber}(p)$	$X \sim \text{Geo}(p)$	One success
	↑ $n = 1$	↑ $r = 1$	
Several trials	$X \sim \text{Bin}(n, p)$	$X \sim \text{NegBin}(r, p)$	Several successes
Interval of time	$X \sim \text{Poi}(\lambda)$	$X \sim \text{Exp}(\lambda)$ (coming soon!)	One success after interval of time

Rapid-fire random variables

number of Snapchats you receive today

A) $\text{Ber}(p)$

B) $\text{Bin}(n, p)$

C) $\text{Poi}(\lambda)$

D) $\text{Geo}(p)$

E) $\text{NegBin}(r, p)$

https://bit.ly/1a2ki4G → https://b.socrative.com/login/student/
Room: CS109SUMMER17

Rapid-fire random variables

number of children until the first one with brown eyes

A) $\text{Ber}(p)$

B) $\text{Bin}(n, p)$

C) $\text{Poi}(\lambda)$

D) $\text{Geo}(p)$

E) $\text{NegBin}(r, p)$

with $r = 1$

https://bit.ly/1a2ki4G → https://b.socrative.com/login/student/

Room: CS109SUMMER17

Rapid-fire random variables

whether the stock market went up today
(1 = up, 0 = down)

A) $\text{Ber}(p)$

B) $\text{Bin}(n, p)$

with $n = 1$

C) $\text{Poi}(\lambda)$

D) $\text{Geo}(p)$

E) $\text{NegBin}(r, p)$

https://bit.ly/1a2ki4G → https://b.socrative.com/login/student/
Room: CS109SUMMER17

Rapid-fire random variables

number of years in some decade
with more than 6 Atlantic hurricanes

A) $\text{Ber}(p)$

B) $\text{Bin}(n, p)$

C) $\text{Poi}(\lambda)$

D) $\text{Geo}(p)$

E) $\text{NegBin}(r, p)$

https://bit.ly/1a2ki4G → https://b.socrative.com/login/student/
Room: CS109SUMMER17