Large Deviations on Sets of Words

ALEA'10

Mireille Régnier

INRIA-AMIB
web page : www.lix.polytechnique.fr/ regnier

March, 23-rd – 2010

Motivation

Find exceptional words, assess the significance

- Problem statement
  Compute $P(X_n \geq k)$, where X_n is the r.v. that counts occurrences of a set of words in a random text of size n
- Method Use generating functions and combinatorial properties of words to compute LD results.

Motivation

Find exceptional words, assess the significance

- Problem statement
  Compute $P(X_n \geq k)$, where X_n is the r.v. that counts occurrences of a set of words in a random text of size n
- Method Use generating functions and combinatorial properties of words to compute LD results.

Central Limit Theorem

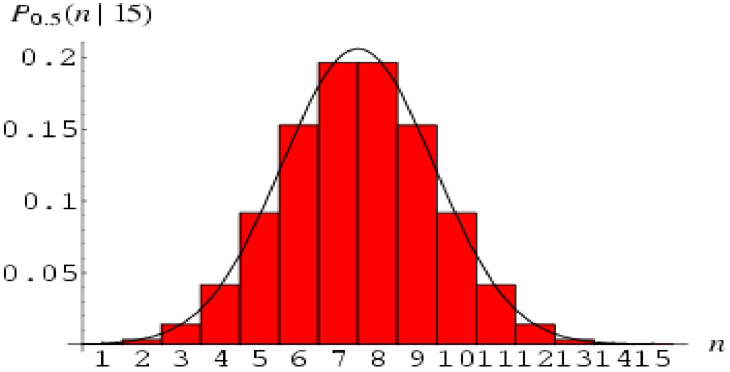
Theorem

Let $X_1, \cdots, X_n$ be i.i.d. random variables with mean μ and variance σ^2, with $0 < \sigma^2 < +\infty$. Then

$$\frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n}\sigma} \to \mathcal{N}(0, 1)$$

when $n \to \infty$.

Illustration



Zscore = comparison with normal law

$$Z(H) = \frac{O(H) - E(H)}{V(H)}$$

$$Z(H) \to \mathcal{N}(0,1)$$

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE | INRIA

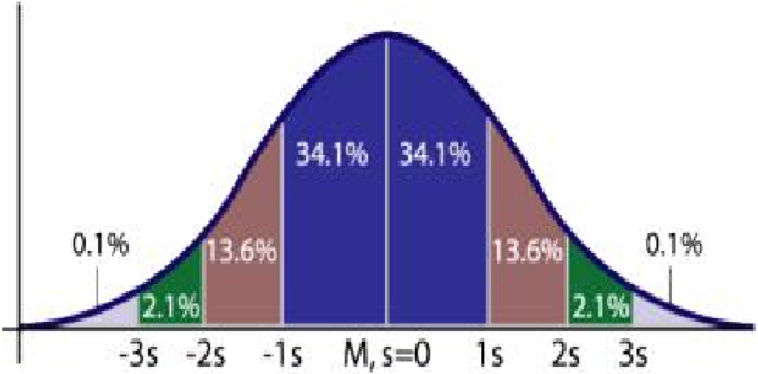
Zscore = comparison with normal law

$$Z(H) = \frac{O(H) - E(H)}{V(H)}$$

$$Z(H) \to \mathcal{N}(0,1)$$

Zscore = comparison with normal law

$$Z(H) = \frac{O(H) - E(H)}{V(H)}$$



INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE | INRIA

Exceptional events

(Discrete) Probability generating functions

Def.

$$\phi_X(t) = \sum_{k=0}^{\infty} e^{tk} P(X = k) \ ,$$

$$\psi_X(u) = \sum_{k=0}^{\infty} u^k P(X = k) = \phi(t = \log u) \ .$$

Levy's Th.

Let (X_n) be a sequence of r.v. and X be a r.v. If

$$\phi_{X_n}(t) \to \phi_X(t) \ ,$$

when t is in a neighbourhood of 0, then $X_n \to X$ (cv. in law).

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

INRIA

Large deviations: basics

Law of Large numbers

$$S_n = \sum_i X_i,\ X_i i.i.d.$$

$$\forall \epsilon > 0 : \lim Prob(|\frac{S_n}{n} - \frac{E(S_n)}{n}| > \epsilon) \to 0\ .$$

Large deviation definitions

$$\frac{1}{\phi(n)} \log Prob(X_n \geq a > E(X_n)) \to I(a)$$

$$\begin{aligned} \phi(n) &= \sqrt{n}, n, \cdots : \textit{speed} \\ I(a) &: \textit{rate} \end{aligned}$$

Words: Combinatorial properties $\to$ explicit expression.

Large deviations: basics

A direct computation: Bernoulli

$$P(X_n = na) = \binom{n}{na} p^{na}(1-p)^{n(1-a)}$$

Stirling formula

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

$$\to \cdots$$

$$P(X_n = na) \sim \cdots e^{-n(a \log \frac{a}{p} + (1-a)\log(\frac{1-a}{1-p}))}$$

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE | INRIA

Large deviations: basics

A direct computation: Bernoulli

$$P(X_n = na) = \binom{n}{na} p^{na}(1-p)^{n(1-a)}$$

Stirling formula

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

$$\to \cdots$$

$$P(X_n = na) \sim \cdots e^{-n(a \log \frac{a}{p} + (1-a)\log(\frac{1-a}{1-p}))}$$

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE | INRIA

LD: generating functions scheme

Generating functions

$$\begin{aligned} P(X_n = na) &= [u^{na}]\psi_{X_n}(u) \\ (Cauchy's\ formula) &= \frac{1}{2i\pi}\int \frac{\psi_{X_n}(u)}{u^{na+1}}du \end{aligned}$$

Integrand $e^{\log\phi_{X_n}(t) - nat} = e^{n[\frac{1}{n}\log\phi_{X_n}(t) - at]}$

$$h_a(t) = \frac{1}{n}\log\phi_{X_n}(t) - at.$$

Hint for the scheme

$$h_a(t) = h_a(t_a) + h_a'(t_a)(t - t_a) + O((t - t_a)^2)$$

$$\frac{1}{n}\log P(X_n = na) \to I(a) = h_a(t_a)$$

with: $h_a'(t_a) = 0, n(t - t_a)^2 \to 0.$

Large deviations: simple examples

Bernoulli

$$\begin{aligned} \phi_{X_n}(t) &= (1 + p(e^t - 1))^n \ , \\ h_a(t) &= \log[1 + p(e^t - 1)] - at \\ h'_a(t) &= \frac{pe^t}{1 + p(e^t - 1)} - a \end{aligned}$$

$$\begin{aligned} t_a &= \log[\frac{a}{p} \cdot \frac{1-p}{1-a}] \\ h_a(t_a) &= a \log \frac{a}{p} + (1-a) \log(\frac{1-a}{1-p}) \end{aligned}$$

Poisson

$$
\begin{aligned}
\phi_{X_n}(t) &= e^{\lambda(e^t-1)} \quad , \\
h_a(t) &= p(e^t - 1) - at \\
h'_a(t) &= pe^t - a
\end{aligned}
$$

$$
\begin{aligned}
t_a &= \log \frac{a}{p} \\
h_a(t_a) &= a - p - a \log(\frac{a}{p})
\end{aligned}
$$

Normal distribution

Mean np, variance $n\sigma^2$.

$$
\begin{aligned}
\phi_{X_n}(t) &= e^{pt-\frac{\sigma^2 t^2}{2}} \quad , \\
h_a(t) &= -\sigma^2\frac{t^2}{2} + (p-a)t \\
h_a'(t) &= -\sigma^2 t + (p-a)
\end{aligned}
$$

$$
\begin{aligned}
t_a &= \frac{p-a}{2\sigma^2} \\
h_a(t_a) &= \frac{(p-a)^2}{2\sigma^2})
\end{aligned}
$$

Remark: B, P $\rightarrow \mathcal{N}(0,1)$
large deviations are different...

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE | INRIA

Large deviations: a single word

$$
\begin{aligned}
P(X_n \geq na) &= P(X \geq k) \\
L_k(z) &= \sum_{n \geq 0} P(X_n \geq k) z^n \\
&= R(z) \cdot M^{na-1}(z) \cdot \frac{1}{1-z} \ .
\end{aligned}
$$

- R: s.g. 1st occurrence;
- M: s.g. other occurrences (possibly overlapping);
- $\frac{1}{1-z}$: s.g. all words.

Large deviations: a single word (2)

$$
\begin{aligned}
P(X_n \geq na) &= \cdots \int \frac{L_k(z)}{z^{na+1}} dz \\
&\to \log R(z) \cdot M^{na-1}(z) \cdot \frac{1}{1-z} - (na+1) \log z \\
h_a(z) &= \log M(z) - a \log z \quad .
\end{aligned}
$$

where

$$
M(z) = 1 - \frac{A_H^{-1}(z)}{1 + \frac{P(H) z^{|H|} A_H^{-1}(z)}{1-z}} \quad .
$$

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE INRIA

Theorem

Let $a > P(H)$*. Then:*

$$Prob(N(H) \geq na) \sim \frac{1}{\sigma_a \sqrt{n}} e^{-nI(a)+\delta_a}$$

where

$$I(a) = a \ln \left(1 - \frac{1 - z_a}{(1 - z_a)(1 - A(z_a)) - z^m P(H)}\right) + \ln z_a$$

and z_a *is a solution of*

$$\begin{aligned} 0 &= (1 - z)^2 [A(z)^2 - A(z) - azA'(z)] \\ &+ z^m P(H) \\ &\times [(1 - z)(2A(z) - am - 1) + z^m P(H) - az] \ . \end{aligned}$$

Complexity: Solve a polynomial equation
vs exponential algorithm [Nuel01].

Several words

$$M(z) \to \mathbb{M}(z)$$

Matrix Dimension $|\mathcal{H}|$
$\mathbb{M}$ diagonal: $\lambda_1, \cdots, \lambda_{|\mathcal{H}|}$.

$$\begin{aligned} L_k(z) &= (\alpha_1, \cdots)\mathbb{M}^{k-1}\begin{pmatrix}\beta_1 \\ \cdot\end{pmatrix} \\ &= \alpha_1 \lambda_1(z)^{k-1}\beta_1 + \cdots \quad . \end{aligned}$$

$$h_a(z) = a \log \lambda_1(z) - \log z$$

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE
INRIA

Large deviations: two words

Counting on two strands

- $\mathbb{M}(z)$ is a 2×2 matrix;
- $d(z) = determinant(\mathbb{M}(z))$;
- $t(z) = Trace(\mathbb{M}(z))$.

Dependency to overlaps

Definition

Given a real a, the equation:

$$\begin{aligned}(azt'(z) - t(z))^2 d(z) + t(z)(azd'(z) - 2d(z))(t(z) - azt'(z)) \\ +(azd'(z) - 2d(z))^2 \quad &= \quad 0\end{aligned}$$

is called the *fundamental equation* .

Theorem

The rate function is :

$$I(a) = -a \log \lambda(z_a) + \log z_a \tag{1}$$

where

$$\lambda(z) = \frac{az(-\psi'(z)\theta(z) - (1 - \psi(z))\theta(z)) - 2(1 - \psi(z))\theta(z) - 2\theta(z)^2}{az(-\psi'(z)\theta(z) + \psi(z)\theta'(z)) - 2\theta(z)^2 + \psi(z)\theta(z)} \tag{2}$$

and z_a is the fundamental root of the fundamental equation.

Hint for the proof

$$\lambda^2(z) - t(z)\lambda + d(z) = 0 \ .$$

Rational expression for λ:

$$\lambda(z) = \frac{azd'(z) - 2d(z)}{azt'(z) - t(z)} \ .$$

$$\lambda_1(z) = \frac{t(z) + \sqrt{\Delta(z)}}{2}$$

$$\lambda_2(z) = \frac{t(z) - \sqrt{\Delta(z)}}{2}$$

$$\lambda_1'(z) = \frac{1}{2}(t'(z) + \frac{1}{2}\frac{\Delta'(z)}{\sqrt{\Delta(z)}}) \ .$$

Equation $h_a'(z) = 0$ rewrites : $az\lambda_1'(z) - \lambda_1(z) = 0$

$$\Rightarrow \ (azt'(z) - t(z))\sqrt{\Delta(z)} = \frac{2\Delta(z) - az\Delta'(z)}{2}$$

Hint for the proof (2)

$$\lambda^2(z) - t(z)\lambda + d(z) = 0 \ .$$

$$(azt'(z) - t(z))\sqrt{\Delta(z)} = \frac{2\Delta(z) - az\Delta'(z)}{2}$$

$$\Delta(z) = t^2(z) - 4d(z) \quad \to \quad \Delta'(z) = \cdots$$

$$\Delta(z) - \frac{az\Delta'(z)}{2} \quad = \quad (t(z) - azt'(z))t(z) + 2azd'(z) - 4d(z) \ .$$

Squaring $\sqrt{\Delta}$, substituting in expression for λ:

$$\lambda(z) = \frac{azd'(z) - 2d(z)}{azt'(z) - t(z)} \ .$$

Comparison with normal approximation

Zscore theory

LD rates

$$
\begin{aligned}
I_N(a) &= \frac{(a-p)^2}{2p} \\
I(a) &\sim a \log \frac{a}{p}
\end{aligned}
$$

Therefore

$$
\begin{aligned}
I(a) &< I_N(a) : \textit{underestimation} \\
I(a) &\sim I_N(a) \Leftrightarrow a - p << p : \textit{central domain}
\end{aligned}
$$

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

INRIA

Comparison with normal approximation

Zscore theory

LD rates

$$I_N(a) = \frac{(a-p)^2}{2p}$$
$$I(a) \sim a \log \frac{a}{p}$$

Therefore

$$I(a) < I_N(a) : \textit{underestimation}$$
$$I(a) \sim I_N(a) \Leftrightarrow a - p << p : \textit{central domain}$$

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE INRIA

Convergence paradox

Ratio and relative errors decrease exponentially

$$n \uparrow, e^{-nI(a)} \downarrow$$

Total variation distance

- evaluates the area between two different distributions
- returns the largest probability.

- The smaller is the maximum distribution
- The bigger is the relative error
- The worse is the approximation

The normal approximation is never correct

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE | INRIA

Convergence paradox

Ratio and relative errors decrease exponentially

$$n \uparrow, e^{-nI(a)} \downarrow$$

Total variation distance

- evaluates the area between two different distributions
- returns the largest probability.

- The smaller is the maximum distribution
- The bigger is the relative error
- The worse is the approximation

The normal approximation is never correct

Spouge's results

Homo sapiens promoter sequences: 30 most significant words sorted by p-value.

Word	ref.	occ.	p-val (rate)	Z- rank	Z-score	Normal-rate
TTTTTTTT	1	22589	0.000638191892	2	+264.507244	2.460797e-03
AAAAAAAA	2	20828	0.0006320334298	1	+271.099279	2.584981e-03
GATTACAG	3	3149	0.0004108880766	3	+213.754122	1.607051e-03
GGATTACA	3b	3039	0.0004008417233	4	+212.430286	1.587207e-03
GGGATTAC	3c	2837	0.0003594976322	5	+196.306517	1.355408e-03
TGGGATTA	3d	3098	0.0003556183659	6	+188.531536	1.250169e-03
ATTACAGG	3e	3153	0.0003397024728	8	+174.786684	1.074527e-03
TGTAATCC	4	2560	0.0003092471632	7	+177.115531	1.103352e-03
CTGTAATC	4b	2579	0.0003038123568	9	+173.212685	1.055261e-03
TAATCCCA	4c	2610	0.000291167097	12	+164.545919	9.523023e-04
GTAATCCC	4e	2381	0.0002799502047	11	+165.967080	9.688232e-04
GCTGGGAT	5	3041	0.0002776774186	15	+146.473114	7.545994e-04
GTGTGTGT	6	3747	0.0002771092421	10	+171.091543	1.029574e-03
CAGGCTGG	7	4086	0.0002665394041	32	+127.732346	5.738553e-04
CCAGGCTG	7b	3931	0.0002512786464	33	+123.609400	5.374074e-04
CCTGTAAT	4b	2629	0.0002502132673	20	+141.622084	7.054441e-04
TGTGTGTG	6b	3866	0.0002437044783	16	+146.040997	7.501536e-04
CTGGGATT	8	3119	0.0002410460071	31	+128.161604	5.777188e-04
ACACACAC	9	3164	0.0002408138032	14	+160.598938	9.071643e-04

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE — INRIA

B *Homo sapiens* promoter sequences: next significant words sorted by p-value.

Word	ref.	occ.	p-val (rate)	Z- rank	Z-score	Normal-rate
ATCCCAGC	10	2623	0.0002316849672	27	+131.961279	6.124825e-04
CCAGCCTG	11	3743	0.0002302858573	39	+116.774661	4.796206e-04
CCCAGCTA	10b	2428	0.0002296464312	26	+135.202979	6.429440e-04
AGTAGCTG	13	2048	0.000227794136	18	+145.360749	7.431816e-04
TTAGTAGA	14	1672	0.0002273984643	13	+163.174975	9.364999e-04
TAGCTGGG	13b	2498	0.0002260538932	29	+131.682643	6.098987e-04
CAGCCTGG	11b	3741	0.0002258388149	42	+115.030518	4.654004e-04
CACACACA	9b	3242	0.0002097672137	23	+135.956731	6.501328e-04
CAGCTACT	11'd	1862	0.0002022195337	25	+135.290088	6.437727e-04
GTAGCTGG	13b	1968	0.0002021605972	28	+131.709404	6.101466e-04
TCAGCCTC	11c	2951	0.0002021588294	44	+113.055714	4.495579e-04

Table: The last column displays the rate function for normal approximation. $n = 14215737$.

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE INRIA