TWO VARIANTS OF LEMPEL-ZIV CRITERION AND THEIR REASONING

V. G. Mikhailov, V. I. Kruglov

Russian Federation Cryptography Academy, Moscow

Moscow, 2021

- Let $X_1, X_2, \ldots,$ be a random binary sequence.

- $H_0 : X_i$ are independent and

$$\mathbf{P}\{X_i = 0\} = \mathbf{P}\{X_i = 1\} = \frac{1}{2}.$$

- Lempel-Ziv statistic: a sequence $X_1, \ldots, X_T$ is divided words (subsequences) in such a way that any next word is the least word that is not equal to any of previous words in the dictionary.

- Lempel-Ziv statistic is the amount $W(T)$ of such words.

- The first word of the dictionary is the empty word.

Example.

Sequence of 12 digits

$$011101101011$$

is divided into

$$0\ 1\ 11\ 01\ 10\ 101\ 1$$

so there are 7 words

$$\varnothing, (0), (1), (11), (01), (10), (101)$$

and remainder 1 that is not considered because it is equal to the third word.

Example.

Sequence of 12 digits

$$010101101010$$

is divided into

$$0\ 1\ 01\ 011\ 010\ 10$$

so there are 7 words

$$\varnothing, (0), (1), (01), (011), (010), (10).$$

The Lempel-Ziv criterion is a goodness-of-fit criterion for the hypothesis H_0.

It may be defined as

$$\{|W(T) - \mathbf{E}W(T)| < C\sqrt{\mathbf{D}W(T)}\} \Rightarrow H_0,$$
$$\{|W(T) - \mathbf{E}W(T)| \geq C\sqrt{\mathbf{D}W(T)}\} \Rightarrow \overline{H}_0,$$

where C is a critical level.

NIST Special Publication 800-22:

$$\lim_{T\to\infty} \frac{\mathbf{E}W(T)}{T/\log_2 T} = 1$$

and

$$\mathbf{D}W(T) \sim \frac{T(C_D + \delta\log_2(T))}{\log_2^3 T} \text{ for } T \to \infty,$$

where $C_D = 0.26600\ldots$ is a constant and $\delta(\cdot)$ is a slowly varying function with zero mean value, $|\delta(\cdot)| < 10^{-6}$.

NIST Special Publication 800-22.

For $T = 10^6$

$$\mathbf{E}W(T) = 69586.25,$$

$$\mathbf{D}W(T) = 70.448718.$$

Let $S(n)$ be equal to cumulative length of all words in a dictionary of n words, then

$$\{W(T) < n\} = \{S(n) > T\}.$$

Theorem 1. Let $X_1, X_2, \ldots$ be a sequence of independent random variables that are distributed on the set $\{0, 1\}$ with probabilities

$$\mathbf{P}\{X_t = 1\} = p, \quad \mathbf{P}\{X_t = 0\} = 1 - p,$$

where $p \in (0, 1)$.

Then for $r = 0, 1, \ldots, n(n-1)/2$

$$\mathbf{P}\{S(n+1) = n + r\} = \sum_{m=0}^{n} C_n^m p^m (1-p)^{n-m} \times$$

$$\times \sum_{l=0}^{r} \mathbf{P}\{S(m) = l\} \mathbf{P}\{S(n-m) = r - l\}.$$

Initial values for $\mathbf{P}\{S(n+1)=r\}$:

$$\mathbf{P}\{S(0)=0\}=\mathbf{P}\{S(1)=0\}=1,$$

$$\mathbf{P}\{S(2)=1\}=1, \quad \mathbf{P}\{S(2)=1+r\}=0, \text{ for } r\geq 1,$$

$$\mathbf{P}\{S(n+1)=r\}=0, \quad r=0,\ldots,n-1.$$

$$\mathbf{P}\{W(T) = n\}$$

n	$T = 1000$	n	$T = 2000$
169	0.0007	300	0.0012
170	0.0089	301	0.0103
171	0.0648	302	0.0564
172	0.2457	303	0.1848
173	0.4098	304	0.3317
174	0.2361	305	0.2915
175	0.0330	306	0.1088
176	0.0006	307	0.0143
		308	0.0005

$$\mathbf{P}\{W(T) = n\}$$

n	$T = 3000$	n	$T = 4000$
420	0.0001	536	0.0005
421	0.0011	537	0.0036
422	0.0081	538	0.0193
423	0.0406	539	0.0710
424	0.1321	540	0.1747
425	0.2647	541	0.2753
426	0.3050	542	0.2633
427	0.1863	543	0.1439
428	0.0545	544	0.0417
429	0.0067	545	0.0058
430	0.0003	546	0.0003

NIST Special Publication 800-22:

$$\lim_{T\to\infty} \frac{\mathbf{E}W(T)}{T/\log_2 T} = 1$$

and

$$\mathbf{D}W(T) \sim \frac{T(C_D + \delta \log_2(T))}{\log_2^3 T} \text{ for } T \to \infty.$$

T	$\mathbf{E}W_T$	$\frac{T}{\log_2 T}$	$\mathbf{D}W_T$	$\frac{0.266T}{\log_2^3 T}$
1000	172.899	100.343	0.96268	0.26874
2000	304.220	182.385	1.34154	0.40345
3000	425.627	259.723	1.65301	0.51781
4000	541.309	334.286	1.92859	0.62103
5000	653.046	406.910	2.18096	0.71686
6000	761.811	478.059	2.41656	0.80727
7000	868.213	548.025	2.63918	0.89348
8000	972.665	617.008	2.85136	0.97628

Criterion with summation.

Divide a sample $X = (X_1, X_2, \ldots, X_{2mT})$ into $2m$ samples, compute $W_1(T), W_2(T), \ldots, W_{2m}(T)$.

Compute

$$\tilde{W}(2mT) \stackrel{def}{=} (W_1 + W_2 + \ldots + W_m) -$$

$$-(W_{m+1} + W_{m+2} + \ldots + W_{2m}).$$

The distribution of $\tilde{W}(2mT)$ is simmetric about zero and

$$\mathbf{E}\tilde{W}(2mT) = 0, \mathbf{D}\tilde{W}(2mT) = 2m\mathbf{D}W(T).$$

Theorem 2. The following inequality is valid:

$$\sup_{-\infty<x<\infty}\left|\mathbf{P}\left\{\frac{\tilde{W}(2mT)}{\sqrt{\mathbf{D}\tilde{W}(2mT)}}<x\right\}-\Phi(x)\right|\leq$$

$$\leq\frac{C_1\mathbf{E}|V_1(2T)|^3}{(2m\mathbf{D}W(T)))^{3/2}},$$

where $C_1\leq 0.4774$.

Corollary. If $m\to\infty$, then for any $-\infty<x<\infty$

$$\mathbf{P}\left\{\frac{\tilde{W}(2mT)}{\sqrt{\mathbf{D}\tilde{W}(2mT)}}<x\right\}\to\Phi(x).$$

Goodness-of-fit criterion for hypothesis H_0:

$$\left\{ |\tilde{W}(2mT)| < C\sqrt{\mathbf{D}\tilde{W}(2mT)} \right\} \Rightarrow H_0,$$

$$\left\{ |\tilde{W}(2mT)| \geq C\sqrt{\mathbf{D}\tilde{W}(2mT)} \right\} \Rightarrow \overline{H}_0.$$

For a given significance level $\alpha > 0$ the critical level C is defined by

$$2(1 - \Phi(C)) = \alpha.$$

To compute the distribution of $\tilde{W}(2mT)$ we rewrite

$$\tilde{W}(2mT) = \sum_{i=1}^{m} V_i(2T) = \sum_{i=1}^{m} (W_i(T) - W_{i+m}(T)),$$

then $\mathbf{E}V_i(2T) = 0$ and distribution of $V_i(2T)$ may be computed by formula

$$\mathbf{P}\{V_i(2T) = k\} = \sum_{l} \mathbf{P}\{W(T) = l\}\mathbf{P}\{W(T) = l-k\}.$$

Distribution of $\tilde{W}(2mT)$ may be computed as the m-fold convolution of the distribution $V_i(2T)$.

Theorem 2. The following inequality is valid:

$$\sup_{-\infty<x<\infty}\left|\mathbf{P}\left\{\frac{\tilde{W}(2mT)}{\sqrt{\mathbf{D}\tilde{W}(2mT)}}<x\right\}-\Phi(x)\right|\leq$$

$$\leq\frac{C_1\mathbf{E}|V_1(2T)|^3}{(2m\mathbf{D}W(T)))^{3/2}},$$

where $C_1\leq 0.4774$.

Values of the right part of the inequality for $\tilde{W}(2mT)$.

T	$m = 1000$	$m = 2000$
1000	2.42924e-005	8.58868e-006
2000	2.42123e-005	8.56035e-006
3000	2.41834e-005	8.55011e-006
4000	2.41720e-005	8.54608e-006
5000	2.41702e-005	8.54547e-006
6000	2.41755e-005	8.54733e-006
7000	2.41869e-005	8.55136e-006
8000	2.42042e-005	8.55748e-006

Criterion of chi-square type.

- Divide a sample $X_1, \ldots, X_n, n = mrT$, into mr subsamples of size T.
- For each subsample of size T compute $W(T)$.
- We obtain mr values $W(T)$:

$$W_{1,1}(T), W_{1,2}(T), \ldots, W_{1,r}(T),$$
$$W_{2,1}(T), W_{2,2}(T), \ldots, W_{2,r}(T),$$
$$\ldots$$
$$W_{m,1}(T), W_{m,2}(T), \ldots, W_{m,r}(T).$$

$$T = 1000$$

n	$\mathbf{P}\{W(T) = n\}$	n	$\mathbf{P}\{W(T) = n\}$
169	0.0007	173	0.4098
170	0.0089	174	0.2361
171	0.0648	175	0.0330
172	0.2457	176	0.0006

$$\Delta_1 = \{0, \ldots, 171\}, \Delta_2 = \{172\}, \Delta_3 = \{173\},$$
$$\Delta_4 = \{174\}, \Delta_5 = \{175, 176, \ldots\}.$$

$$p_j^0 = \mathbf{P}\{W(T) \in \Delta_j\}, j = 1, \ldots, N, N = 5.$$
$$p_1^0 = 0.0746, p_2^0 = 0.2457, p_3^0 = 0.4098,$$
$$p_4^0 = 0.2361, p_5^0 = 0.03368.$$

For any $k = 1, \ldots, m$ and

$$W_{k,1}(T), \ldots, W_{k,r}(T)$$

we compute

$$v_{k,1}(T), v_{k,2}(T), v_{k,3}(T), v_{k,4}(T), v_{k,5}(T),$$

where $v_{k,j}$ is the amount of $W_{k,i}(T)$ in Δ_j.

Compute

$$\chi^2_k(rT) = \sum_{j=1}^{5} \frac{(v_{k,j} - np^0_j)^2}{np^0_j}$$

and

$$\tilde{\chi}^2(mrT) = \max_{1 \le k \le m} \chi^2_k(T).$$

Let $\chi^2_{N-1}(x)$ be the distribution function of χ^2-distribution with $N-1$ degrees of freedom and for a given significance level $\alpha \in (0,1)$ define $C(N-1,\alpha)$ by

$$\chi^2_{N-1}(C(N-1,\alpha)) = \alpha.$$

We have $N = 5$ so we calculate the quantile $C(4,\alpha^{1/m})$ and define the criterion by the rules

$$\{\tilde{\chi}^2(mrT) < C(N-1,\alpha^{1/m})\} \Rightarrow H_0,$$

$$\{\tilde{\chi}^2(mrT) \geq C(N-1,\alpha^{1/m})\} \Rightarrow \overline{H}_0.$$

Theorem 3. Let the hypothesis H_0 be true, so random variables $X_1, \ldots, X_{mrT}$ are independent and equiprobably distributed on $\{0, 1\}$.
If parameters m, T and N are fixed and $r \to \infty$, then for any $x \in (-\infty, +\infty)$

$$\mathbf{P}\{\tilde{\chi}^2(mrT) < x\} \to 1 - \left(1 - \chi^2_{N-1}(x)\right)^m$$

and

$$\mathbf{P}\{\tilde{\chi}^2(mrT) \geq C(N-1, \alpha^{1/m})\} \to \alpha.$$

So, if the hypothesis H_0 is true and the size mrT of sample increases, then the probability to reject H_0 tends to α.

If value T increases, then the number of values such that random variable $W(T)$ is equal to this value with significant probability also increases.

For example, for $T = 8000$ the set of possible values of $W(T)$ may be divided into $N = 7$ intervals, so the distribution of $\tilde{\chi}^2(mrT)$ converges to χ^2-distribution with $N - 1 = 6$ degrees of freedom.

Thank you.