

František Zítek

O odhadu pravděpodobností přechodu

*Aplikace matematiky*, Vol. 2 (1957), No. 4, 251--257

Persistent URL: <http://dml.cz/dmlcz/102575>

## Terms of use:

© Institute of Mathematics AS CR, 1957

Institute of Mathematics of the Academy of Sciences of the Czech Republic provides access to digitized documents strictly for personal use. Each copy of any part of this document must contain these *Terms of use*.



This paper has been digitized, optimized for electronic delivery and stamped with digital signature within the project *DML-CZ: The Czech Digital Mathematics Library* <http://dml.cz>

## O ODHADU PRAVDĚPODOBNOSTÍ PŘECHODU

FRANTIŠEK ZÍTEK

(Došlo dne 4. prosince 1956.)

DT: 519.217

Práce se týká odhadů pravděpodobností přechodu v homogenním Markovově řetězci o dvou stavech pomocí různých metod odhadu.

## 1. Úvod

V práci [6] se O. LANGE zabývá m. j. problémem odhadu pravděpodobností přechodu v homogenním Markovově řetězci (viz [6], § 8). Jeho výsledek lze stručně shrnouti takto:

Průběh procesu sledujeme po dobu  $N$  přechodů. Necht  $z$  celkového počtu  $N$  přechodů je  $m_{ij}$  přechodů ze stavu  $i$  do stavu  $j$ . Označme  $n_i = \sum_j m_{ij}$ ,  $N = \sum_i n_i$ . Nestrannými maximálně věrohodnými odhady pravděpodobností přechodu  $p_{ij}$  jsou pak výrazy

$$\hat{p}_{ij} = \frac{m_{ij}}{n_i}. \quad (1.1)$$

Pro varianci odhadu (1.1) uvádí Lange výraz

$$D^2(\hat{p}_{ij}) = \frac{1}{n_i} p_{ij}(1 - p_{ij}). \quad (1.2)$$

## 2. Vlastnosti Langeova odhadu

Podstatnou skutečností ovlivňující charakter odhadu (1.1) je ten fakt, že jak  $m_{ij}$ , tak také  $n_i$  jsou náhodné proměnné; jde tedy zde o t. zv. odhad poměrový (viz na př. [3] a [4]). Variance (1.2) je tudíž jen *podmíněnou* variancí odhadu (1.1) za předpokladu, že  $n_i$  je pevně dáno. Při praktickém uspořádání experimentu však máme pevné jen číslo  $N$  a matici pravděpodobností přechodu  $\mathbf{P} = (p_{ij})$ . Na tuto důležitou skutečnost však patrně v Langeově práci nebyl vzat zřetel.

Nepodmíněnou varianci odhadu (1.1) dostaneme, vezmeme-li střední hodnotu variance (1.2) vzhledem k rozložení náhodné proměnné  $n_i$ . Toto rozložení je v obecném případě značně složité. I když se omezíme jen na nejjednodušší případ homogenního Markovova řetězce o dvou stavech, budou formule pro pravděpodobnosti  $\mathbf{P}(n_i = k)$  velmi komplikované. Spokojíme se proto s asymptotickým řešením.

Budiž

$$\mathbf{P} = \begin{pmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{pmatrix}, p_{11}p_{12}p_{21}p_{22} \neq 0, \quad (2.1)$$

matice pravděpodobnosti přechodu daného řetězce  $\{X_k\}_{k=0}^{\infty}$ ; jako počáteční rozložení přijmeme

$$\mathbf{P}(X_0 = 1) = \bar{p}, \quad (2.2)$$

kde  $\bar{p}$  je řešením rovnice

$$\bar{p} = \bar{p} \cdot p_{11} + (1 - \bar{p}) \cdot p_{21}, \quad (2.3)$$

t. j.

$$\bar{p} = \frac{p_{21}}{p_{12} + p_{21}}. \quad (2.4)$$

Potom ovšem platí  $\mathbf{P}(X_k = 1) = \bar{p}$  pro všechna  $k \geq 0$ .

Je-li  $N$  dostatečně veliké a neliší-li se  $\bar{p}$  příliš od jedné poloviny, pak  $n_1$  má přibližně normální rozložení s průměrem  $N\bar{p}$  a variancí  $N\bar{p}(1 - \bar{p})$ , (viz na př. [7], § 14, věta 2), a  $m_{12}$  má (při daném  $n_1$ ) také přibližně normální rozložení s průměrem  $n_1 p_{12}$  a variancí  $n_1 p_{12} p_{11}$ .

V dalším postupu užitíme metody navržené E. C. FIELLEREM [2]. Označme  $y_1 = m_{12} - p_{12}n_1$  a  $y_2 = m_{21} - p_{21}n_2$ . Zřejmě  $\mathbf{E}(y_1) = 0 = \mathbf{E}(y_2)$ . Kromě toho platí

$$\mathbf{D}^2(y_1) = N\bar{p}p_{12}p_{11} \quad (2.5)$$

a

$$\mathbf{D}^2(y_2) = N(1 - \bar{p})p_{21}p_{22}. \quad (2.6)$$

Využijeme-li normality náhodných proměnných  $y_1$  a  $y_2$  můžeme psát intervalové odhady pro pravděpodobnosti přechodu  $p_{12}$  a  $p_{21}$  ve tvaru

$$p_{12}^* = \frac{m_{12}}{n_1} \pm \frac{t\sigma_1}{n_1}; \quad p_{21}^* = \frac{m_{21}}{n_2} \pm \frac{t\sigma_2}{n_2}, \quad (2.7)$$

kde  $\sigma_i^2 = \mathbf{D}^2(y_i)$ ,  $i = 1, 2$ , a  $t$  určíme z tabulek normálního rozložení.

Variance  $\sigma_i^2$  jsou ovšem, vyjádřeny ve tvaru (2.5) resp. (2.6), funkcemi odhadovaných pravděpodobností  $p_{12}$  a  $p_{21}$ . Fieller zde navrhuje použít bodových odhadů (1.1) dosazených do (2.5) a (2.6) místo skutečných hodnot pří-

slušných pravděpodobností. Tímto obratem dostaneme z (2.5) s použitím (2.4) výraz

$$\hat{\sigma}_1^2 = N \frac{m_{11}m_{21}m_{12}}{n_1(m_{12}n_2 + m_{21}n_1)} \quad (2.8)$$

a obdobně můžeme přepsati též (2.6). Získali jsme tak současně oba intervalové odhady založené jen na výběrových charakteristikách.

### 3. Haldaneův odhad

Složitým úvahám a výpočtům paragrafu 2 se můžeme vyhnouti jestliže místo (1.1) použijeme jiného odhadu pravděpodobností  $p_{ij}$ . Jednou z možností, jež se tu naskýtají, je použití metody, kterou pro nezávislé pokusy propracoval J. B. S. HALDANE v práci [5] a kterou lze bezprostředně přenést i na náš případ odhadu pravděpodobností přechodu v Markovově řetězci.

Pro začátek se opět omezíme na řetězec o dvou stavech s maticí pravděpodobností přechodu (2.1). Podstatný rozdíl Haldaneovy metody odhadu oproti odhadu (1.1) spočívá v tom, že není předem pevně určena délka pozorovaného úseku řetězce, nýbrž počet přechodů ze stavu 1 do 2 a naopak. Co do počátečního rozložení předpokládáme tentokrát

$$\mathbf{P}(X_0 = 1) = 1 \quad (3.1)$$

a rozhodneme, že pozorování skončí, jakmile dojde k  $m$ -tému přechodu z 2 do 1. Nastane tedy v průběhu pozorování také právě  $m$  přechodů z 1 do 2, neboť vzhledem k (3.1) začal proces ve stavu 1. Dále označme  $S_m$  úhrnnou „dobu pobytu“ ve stavu 1 a obdobně  $T_m$  úhrnnou dobu pobytu ve stavu 2. (Dobou pobytu ve stavu 1 rozumíme počet jedniček v posloupnosti  $\{X_i\}$ ,  $i = 0, 1, 2, \dots$ ; podobně pro stav 2. Poslední jednička — po  $m$ -tém přechodu z 2 zpět do 1 — se do této doby nezapočítává.) Snadnou úvahou pak zjistíme, že  $S_m$  resp.  $T_m$  jsou součty  $m$  nezávislých náhodných proměnných s Pascalovým rozložením, takže

$$\mathbf{P}(S_m = k) = \binom{k-1}{k-m} p_{12}^m p_{11}^{k-m} \quad (3.2)$$

a

$$\mathbf{P}(T_m = k) = \binom{k-1}{k-m} p_{21}^m p_{22}^{k-m} \quad (3.3)$$

Haldaneovým odhadem pravděpodobnosti  $p_{12}$  je pak výraz

$$p_{12}^* = \frac{m-1}{S_m-1}; \quad (3.5)$$

podobně pro pravděpodobnost  $p_{21}$  dostaneme

$$p_{21}^* = \frac{m-1}{T_m-1}. \quad (3.5)$$

Poznámka. Odhady (3.4) a (3.5) jsou — na rozdíl od odpovídajících odhadů typu (1.1) — stochasticky nezávislé a také „soběstačné“ v tom smyslu, že k určení vlastností jednoho z nich není třeba znát hodnotu pravděpodobnosti odhadované druhým odhadem. Tato „soběstačnost“ je určitou výhodou Haldaneovy metody, neboť umožňuje provést samostatně odhad i jen jedné z pravděpodobností  $p_{12}$  a  $p_{21}$ .

Vzhledem k právě uvedené vlastnosti se můžeme v dalším omezit jen na odhad (3.4), vlastnosti odhadu (3.5) jsou zcela analogické. Přitom pro jednoduchost píšeme  $p$  místo  $p_{12}$ .

Nestrannost odhadu (3.4) lze snadno ověřit přímým výpočtem (viz [5]), jako varianci odhadu (3.4) pak Haldane uvádí

$$D^2(p^*) = \frac{p^2(1-p)}{m} \cdot \sum_{k=0}^{\infty} \frac{(k+1)! m!}{(k+m)!} (1-p)^k. \quad (3.6)$$

Dosadíme-li do (3.6) za  $p$  odhad  $p^*$ , dostaneme přibližný výraz (viz opět [5])

$$D^2(p^*) \approx \frac{m(S_m - m)}{S_m^2(S_m - 1)} \cdot \left[ 1 + \frac{2(S_m - m)(S_m - 3m)}{m^2 S_m^2} + O\left(\frac{1}{m^3}\right) \right]. \quad (3.7)$$

V obecném případě Markovova řetězce o více stavech lze Haldaneovy metody použít na př. k odhadu diagonálních prvků matice pravděpodobností přechodu. Pozorování sloužící k odhadu  $p_{ii}$  probíhá tak dlouho, pokud nedojde k  $m$ -tému výstupu ze stavu  $i$ ; doba pobytu ve stavu  $i$  má opět rozložení typu (3.2). O jiných možných aplikacích Haldaneovy metody viz na př. [8], str. 114 ssq.

Místo Haldaneova odhadu (3.4) by bylo možno použít též bodového odhadu, sestrojeného pomocí metody maximální věrohodnosti. Dostaneme tak odhad

$$\hat{p} = \frac{m}{S_m}, \quad (3.8)$$

který však není nestranný a má dosti složité charakteristiky. Kromě toho by ovšem bylo možné uvažovat i jiné typy odhadů, na př. odhady pomocí maximální doby pobytu (místo jejich součtu  $S_m$ ), odhady mediánové atp. Jednoduchým ale málo ekonomickým odhadem je odhad

$$\hat{p}_{12} = \frac{m_{12}}{n_1} \quad (3.9)$$

kde  $n_1$  je předem pevně zvoleno, t. j. kdy pozorování konáme tak dlouho, až nastane  $n_1$ -tý přechod ( $\neq$  výstup) ze stavu 1. (Pro  $\hat{p}_{21}$  pak potřebujeme novou serii pozorování.) Odhady  $\hat{p}_{12}$  a  $\hat{p}_{21}$  jsou stochasticky nezávislé, binomické, s variancí (1.2).

Zajímavé by rovněž byly problémy testování hypotéz týkajících se pravděpodobností  $p_{ij}$ .

Poznámka. Numericky je hodnota odhadu (3.8) táž jako hodnota odhadu (1.1) při stejném  $m$ , avšak v (3.8) je  $m$  konstanta a nikoliv náhodná proměnná. Proto také odhad (1.1) může být nestranný, i když (3.8) nestranný není.

#### 4. Intervalový odhad

Pro parametr Pascalova rozložení lze snadno sestrojiti intervalový odhad pomocí intervalových odhadů parametru binomického rozložení; postup si ukážeme na příkladě odhadu pravděpodobnosti  $p = p_{12}$ . Z (3.2) vyplývá přímo rovnost

$$\mathbf{P}(S_m > k) = B(m - 1; p, k), \quad (4.1)$$

kde jsme symbolem  $B(m; p, k)$  označili hodnotu binomické distribuční funkce s parametry  $p$  a  $k$  v bodě  $m$ , tedy

$$B(m; p, k) = \sum_{j=0}^m \binom{k}{j} p^j (1-p)^{k-j}. \quad (4.2)$$

Funkce  $B$  je klesající funkcí argumentu  $p$ .

Budiž dáno reálné číslo  $\alpha$ ,  $0 < \alpha < 1$ . Pro celá čísla  $s \geq m$  definujme si dvě funkce  $P_1(s)$  a  $P_2(s)$  takto:  $P_1(s)$  bude největší  $p$ ,  $0 \leq p \leq 1$ , takové, že

$$B(m - 1; p, s) \geq 1 - \frac{\alpha}{2}, \quad (4.3)$$

kdežto  $P_2(s)$  bude nejmenší  $p$ ,  $0 \leq p \leq 1$ , splňující nerovnost

$$B(m - 1; p, s - 1) \leq \frac{\alpha}{2}. \quad (4.4)$$

Obě funkce jsou zřejmě klesající v  $s$ , platí  $P_1(s) \leq P_2(s)$  a kromě toho

$$\mathbf{P}\{P_1(S_m) \leq p \leq P_2(S_m)\} \geq 1 - \alpha. \quad (4.5)$$

Nerovnost (4.5), jež je přímým důsledkem definice funkcí  $P_1$  a  $P_2$  a rovnosti (4.1), neznamená pak vlastně nic jiného, než že náhodný interval

$$\langle P_1(S_m), P_2(S_m) \rangle \quad (4.6)$$

je konfidenční interval pro  $p$  s koeficientem spolehlivosti  $1 - \alpha$ . Zbývá jen určit funkce  $P_1$  a  $P_2$  explicitně. To je však již snadné, neboť hodnoty těchto funkcí najdeme přímo v tabulkách konfidenčních intervalů parametru binomického rozložení, na příklad v [1] (tabulka XIX, str. 520–529). Musíme ovšem poněkud upravit označení. Tak  $P_1(s)$  nalezneme jako dolní mez konfidenčního intervalu pro  $x = m - 1$  a  $n - x = s - m + 1$ , kdežto  $P_2(s)$  najdeme jako horní mez konfidenčního intervalu pro  $x = m - 1$  a  $n - x = s - m$ .

#### 5. Příklad

Pro ilustraci uvedených method odhadu byl pomocí vrhů mincí realizován Markovův řetězec s maticí pravděpodobností přechodu

$$\mathbf{P} = \begin{pmatrix} \frac{7}{8} & \frac{1}{8} \\ \frac{1}{4} & \frac{3}{4} \end{pmatrix}, \quad (5.1)$$

пři čemž za počáteční rozložení bylo vzato (3.1). Zvoleno  $m = 8$ . Výsledkem pokusu byly hodnoty  $S_8 = 50$ ,  $T_8 = 36$ , takže pro skutečné hodnoty  $p_{12} = 0,125$  a  $p_{21} = 0,25$  nám dávají

a) odhady (3.4) a (3.5) hodnoty

$$\hat{p}_{12}^* = 0,143 \quad \text{a} \quad \hat{p}_{21}^* = 0,200; \quad (5.2)$$

b) odhady (3.8) hodnoty

$$\hat{p}_{12} = 0,160 \quad \text{a} \quad \hat{p}_{21} = 0,222; \quad (5.3)$$

c) intervalové odhady (4.6) jsou při  $\alpha = 0,05$

$$p_{12} \in \langle 0,051; 0,241 \rangle \quad \text{a} \quad p_{21} \in \langle 0,068; 0,314 \rangle. \quad (5.4)$$

*V závěru chtěl bych ještě poděkovat s. dr. JAROSLAVU HÁJKOVI za cenné rady i za možnost seznámit se s rukopisem dosud neuveřejněné práce [4].*

#### LITERATURA

- [1] И. В. Душин-Барковский, Н. В. Смирнов: Теория вероятностей и математическая статистика в технике (общая часть); Москва 1955.
- [2] E. C. Fieller: A fundamental formula in the statistics of biological assay, and some applications; Quart. J. Pharm., 17 (1944), 117—123.
- [3] J. Hájek: Oblastní výběr; Aplikace matematiky, 1 (1956), 149—161.
- [4] J. Hájek: Contributions to the theory of random sampling from finite populations (dosud neotištěno).
- [5] J. B. S. Haldane: On a method of estimating frequencies; Biometrika, 33 (1943—46), 222—225.
- [6] O. Lange: Statistical estimation of parameters in Markov processes; Colloquium Math., 3 (1955), 147—160.
- [7] Т. А. Сарымсаков: Основы теории процессов Маркова; Москва 1954.
- [8] L. Truksa: Statistická dynamika (skripta), Praha 1956.

#### Резюме

#### ОБ ОЦЕНКАХ ВЕРОЯТНОСТЕЙ ПЕРЕХОДА

ФРАТИШЕК ЗИТЭК (František Zítek)

(Поступило в редакцию 4/XII 1956 г.)

В работе дан обзор нескольких методов нахождения оценки вероятностей перехода в однородной цепи Маркова с двумя состояниями. Указана ошибочность формулы (1.2) как дисперсии оценки (1.1) (см. О. Ланге [6]), и построены соответствующие доверительные интервалы. В § 3 рассматри-

вается оценка Холдейна [5]. В § 4 построены доверительные интервалы (4.6) для этих вероятностей перехода при помощи доверительных интервалов для параметра  $p$  биномиального распределения. В качестве примера в § 5 вычислены рассматриваемые оценки в одном частном случае.

### Summary

## ON ESTIMATING TRANSITION PROBABILITIES

FRANTIŠEK ZÍTEK

(Received December 4, 1956.)

The paper contains a review of several methods of estimation of transition probabilities in a homogeneous two-state Markov chain. The error of Lange [6] taking (1.2) for the variance of (1.1) is corrected and corresponding confidence intervals are given. In § 3 the method of J. B. S. Haldane [5] is considered. In § 4 confidence intervals (4.6) for the transition probabilities are constructed using confidence intervals for the parameter  $p$  of the binomial distribution. A numerical example is calculated in § 5 in order to illustrate the different methods of estimation.