Short thesis for the degree
of Doctor of Philosophy (PhD)

Limit theorems and convergence rate for longest contaminated runs of heads

written by SUJA MICHAEL OCHIENG
and supervised by PROF. DR. ISTVÁN FAZEKAS

1949

UNIVERSITY OF DEBRECEN

Doctoral School of Mathematical and Computational Sciences
Debrecen, 2024

Notations and Symbols

T	Number of tails interrupting (contaminating) head run sequence.
τ_m	The first hitting time of the longest contaminated run of length m.
N	The number of coin tossing or the length of the experiment
$\tilde{\xi}^T(n, N)$	Number of precisely T-contaminated head runs of length n.
$\xi^T(n, N)$	Number of at most T-contaminated head runs of length n.
$\mu(N)$	Length of the longest contaminated run in a single experiment.
$[x]$	Represents the largest integer less than or equal to x.
$f = O(g)$	Growth rate of a function, that is $f(x)/g(x)$ remains bounded as $x \to \infty$.
$f \sim g$	Asymptotic equality, that is $f(x)/g(x) \to 1$ as $x \to \infty$.
$\mathbb{I}\{A\}$	Indicator function of a subset A assuming values 0 or 1.

In this short booklet, we summarize the most important results of this dissertation. We mention some of the most fascinating lemmas, propositions, theorems and simulation results based on our research which consist of three published papers; [**Fazekas and Suja (2021)**, **Fazekas, Fazekas, and Suja (2024)** and (**Fazekas, Fazekas, and Suja (2023)**)].

The introduction contains several historical facts on limit theorems in probability theory and their applications to the case of coin tossing experiments. In particular, the study of success runs in Bernoulli trials which has received indubitable attention of several researchers due to its inherent theoretical interest and intriguing applications. The problem of the length of the longest pure head run for n Bernoulli random variables was first raised by T. Varga in his classroom experiment and the findings herald overwhelming research interest, variations and extensions to other situations.

The results of **Erdős and Rényi (1970)** and **Földes (1979)** had immense influence on the trajectory of our study. A lot of insights regarding proofs of theorems were drawn from the main Lemma of **Csáki et al. (1987)**. This powerful Lemma provided a good approximation to the probabilities which offered limiting distribution of the random variable τ_m, the first occurrence time of the event of interest.

In Chapter 1, we defined a T-contaminated run of heads and study the limiting distributions of their numbers together with the first hitting time and the asymptotic behaviour of the length of the longest T-contaminated head run. More emphasis was devoted to approximation of the numbers of contaminated runs to both Poisson and compound Poisson limit laws.

In Chapter 2, we dealt with T-contaminated head run but more emphasis was now shifted to the asymptotic distribution of the length of the longest T-contaminated head run. Here we investigated the rate of convergence to an accompanying distribution and also obtained results for the first hitting time for the same.

In Chapter 3, we defined a two type contaminated run and studied the limiting distribution of the first hitting time and the accompanying distribution of the longest at most two type contaminated runs with trinary outcomes. Our approach mirrored the one used in Chapter 2.

At the end of the dissertation, possible further research based on the results obtained is given a long with the appendix which contains the main Lemma nonstationary finite form of **Csáki et al. (1987)** where other than providing the elegant proof, we precisely fixed the condition of the lemma.

Chapter 1

Limit theorems of T-contaminated run of heads

Now, we begin with the problem setting for our research. Consider the classical coin tossing experiment. Let $p \in (0,1)$ be the probability of heads and $q = 1-p$ the probability of tails. Here, p is fixed while we toss a coin N times independently. We write 1 when the result is head and 0 when the result is tail. Therefore we consider independent identically distributed random variables $X_1, X_2, \ldots, X_N$ with $\mathbb{P}(X_i = 1) = p$ and $\mathbb{P}(X_i = 0) = q$, $i = 1, 2, \ldots, N$. Let $T \geq 0$ be fixed integer.

In this chapter we list the results of Fazekas and Suja (2021). These are extensions of the results of Földes (1975) for the case of a fair coin in which $p = 1/2$ to an arbitrary case, $p \in (0,1)$.

Number of precisely those T-contaminated run of heads

Let

$$\tilde{\eta}_i = \tilde{\eta}_i{}^T(n) = \begin{cases} 1, & \text{if there are precisely } T \text{ 0 values among} \\ & X_i, \ldots, X_{i+n-1} \text{ and } X_{i-1} = 0, \\ 0, & \text{otherwise.} \end{cases}$$

Here we let X_0 be defined as $X_0 = 0$ and let

$$\tilde{\xi} = \tilde{\xi}^T(n, N) = \sum_{i=1}^{N-n+1} \tilde{\eta_i}^T(n).$$

Now $\tilde{\xi} = \tilde{\xi}^T(n, N)$ denote the number of those precisely T-contaminated n length runs of heads for which the preceding element is a tail.

Our main condition in this first chapter is the following. If we let $p \in (0, 1)$ be fixed and T be a fixed non-negative integer. Now if we let $N \to \infty$ and $n \to \infty$ such that

$$\frac{Nq^{T+1}p^{n-T}n^T}{T!} \to \lambda > 0, \tag{1.1}$$

where if λ is fixed, then we remark that above condition implies that $N/n \to \infty$. Now we intend to show that the distribution of $\tilde{\xi}$ converges to the λ parameter Poisson distribution.

Theorem 1. *Let T be fixed. Let $N \to \infty$ and $n \to \infty$ so that the above condition (1.1) is satisfied. Then*

$$\lim_{N\to\infty} \mathbb{P}(\tilde{\xi}^T(n, N) = k) = \frac{e^{-\lambda}\lambda^k}{k!}, k = 0, 1, 2, \ldots$$

In proving the theorem, we consider $l_m = N - n + 1$ and $Y_i = \tilde{\eta_i}$ $i = 1, 2, \ldots, l_m$ and checked fulfilment of the conditions of Proposition below due to Sevast'yanov.

Proposition 1. *(Sevast'yanov (1972)) Let ${Y_i}^{(m)}$, $i = 1, 2, \ldots, l_m$, $m = 1, 2, \ldots$, be a triangular array of Bernoulli random variables, i.e. the values of ${Y_i}^{(m)}$ are 0 or 1. Let*

$$\mathbb{Z}_m = {Y_1}^{(m)} + {Y_2}^{(m)} + \cdots + {Y_{l_m}}^{(m)}, m = 1, 2, \ldots$$

be the row sums and

$$b^{(m)}_{i_1,i_2,\ldots,i_r} = \mathbb{P}({Y_{i_1}}^{(m)} = {Y_{i_2}}^{(m)} = \cdots = {Y_{i_r}}^{(m)} = 1),$$

where $(i_1, i_2, \ldots, i_r)$ denotes an r dimensional vector such that integers $i_1, i_2, \ldots, i_r$ are pairwise different with $1 \le i_t \le l_m$, $t = 1, 2, \ldots, r$, $r = 1, 2, \ldots$.
Assume that for each $r = 2, 3, \ldots$, $m = 1, 2, \ldots$ there exists an exceptional set $I_r(m)$ consisting of certain vectors $\alpha_r = (i_1, i_2, \ldots, i_r)$ such that the numbers $i_1, i_2, \ldots, i_r$ are pairwise different with $1 \le i_t \le l_m$, $t = 1, 2, \ldots, r$.

In addition, we assume the following that

$$\lim_{m\to\infty} \max_{1\le i\le l_m} {b_i}^{(m)} = 0,$$

$$\lim_{m\to\infty} \sum_{i=1}^{l_m} {b_i}^{(m)} = \lambda > 0,$$

$$\lim_{m\to\infty} \sum_{\alpha_r \in I_r(m)} b^{(m)}_{i_1,i_2,\ldots,i_r} = 0,$$

$$\lim_{m\to\infty} \sum_{\alpha_r \in I_r(m)} b^{(m)}_{i_1} \cdots b^{(m)}_{i_r} = 0,$$

and uniformly for all $\alpha_r \notin I_r(m)$

$$\lim_{m\to\infty} \frac{b^{(m)}_{i_1,i_2,\ldots,i_r}}{b^{(m)}_{i_1} \ldots b^{(m)}_{i_r}} = 1.$$

Then

$$\lim_{m\to\infty} \mathbb{P}(\mathbb{Z}_m = k) = \frac{e^{-\lambda}\lambda^k}{k!}, k = 0, 1, 2, \ldots$$

Number of at most T-contaminated run of heads

Now we turn to the problem of the number of at most T-contaminated runs of heads and let

$$\eta_i = \eta_i^T(n) = \begin{cases} 1, & \text{if there are at most T 0 values among} \\ & X_i, \ldots, X_{i+n-1} \\ 0, & \text{otherwise} \end{cases}$$

Now we let

$$\xi = \xi^T(n, N) = \sum_{i=1}^{N-n+1} \eta_i^T(n).$$

Therefore ξ is considered as the number of head runs being at most T-contaminated and having length n. Now we want to prove that the distribution of ξ converges to a compound Poisson distribution in the limit.

Theorem 2. *Let T be fixed. We let $N \to \infty$ and $n \to \infty$ so that condition (1.1) is satisfied. Then, for the generator functions we have*

$$\lim_{N\to\infty} \mathbb{E}\left(z^{\xi^T(n,N)}\right) = \exp\left[\lambda\left(\frac{qz}{1-pz} - 1\right)\right].$$

Remark 1. *More specifically, in our case we need a particular version of compound Poisson distribution, that is the so called geometric Poisson distribution. Let γ have Poisson distribution $\mathbb{P}(\gamma = k) = \lambda^k e^{-\lambda}/k!$, $k = 0, 1, 2, \ldots$. Let $\varrho_1, \varrho_2, \ldots$,be random variables independent of each other and of γ having q parameter geometric random distribution:*
$\mathbb{P}(\varrho_i = l) = p^{l-1}q$, $l = 1, 2, \ldots$, $q \in (0, 1)$, $p = 1 - q$.
When $\gamma = k$, we let the distribution of ϱ to be the same as that of $\varrho_1 + \ldots + \varrho_k$. (Here, an empty sum is defined as 0, i.e $\varrho = 0$ when $\gamma = 0$).
Then ϱ has generator function

$$\mathbb{E}(z^{\varrho}) = \exp\left[\lambda\left(\frac{qz}{1-pz} - 1\right)\right] \text{ for } \quad |zp| < 1.$$

To give a formal explanation of this fact, let

$$\eta'_i = 1 = \eta'^T_i(n) = \begin{cases} \tilde{\eta_i}^T(n).X_{i+n-1}, & \text{if} \quad i > 1 \\ \eta_i^T(n), & \text{if} \quad i = 1. \end{cases}$$

To be more precise, we considered the following representation of $\xi = \xi^T(n, N)$,

$$\xi = \xi^T(n, N) = \sum_{i=1}^{N-n+1} \gamma_i^T(n) = \sum_{i=1}^{N-n+1} \gamma_i,$$

where

$$\gamma_i = \gamma_i^T(n) = \eta'_i\,[\min\{k > 0 : \text{either} \quad \eta_{i+k} = 0 \quad \text{or} \quad i + k + n - 1 > N\}]$$

First hitting time of T-contaminated runs of heads

Now, we are going to briefly consider the first hitting time of T-contaminated runs of heads. This is τ, the number of tosses needed in a coin tossing experiment for a T-contaminated head run of length n to appear for the very first time i.e. its the first observation time when the number of tails among the last n outcomes is at most T.

Let

$$\tau = \tau^T(n) = \min\{N : \xi^T(n, N) > 0\}.$$

If $T = 0$, τ is the usual waiting time for a pure head run of length n. We show that the appropriately normalized version of τ has exponential limiting distribution.

Theorem 3. *Let T be fixed. Then, for any $0 < x < \infty$,*

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{\tau^T(n)n^T}{T!}q^{T+1}p^{n-T} \le x\right) = 1 - e^{-x}.$$

Length of the longest T-contaminated runs of heads

Next, we now consider the Length of the longest T-contaminated runs of heads. We describes the accompanying distribution of $\mu^T(N)$. Let

$$\mu = \mu^T(N) = \max\{n : \xi^T(n, N) > 0\}.$$

Considering the result of tossing a coin N times, μ is the length of the longest run of heads containing at most T tails. We offer a two parameter family of distributions to approximate the distribution of μ. By letting B be a fixed positive number, then for any positive x, we have that

$$x = kB + r,$$

where k is integer and r is the residual for which $0 \le r < B$. Here k and r are uniquely determined. We define $[x]_B$ and $\{x\}_B$ as $[x]_B = kB$ and $\{x\}_B = r$.

Theorem 4. *Let T be fixed. Let B be a fixed positive number and let S be a fixed number. Then, for any integer k we have*

$$\mathbb{P}(\mu^T(N) - [\log N + T\log(\log N + S\log\log N)]_B < k) =$$
$$= \exp\left(-q^{T+1}p^{(k-T-\{\log N+T\log(\log N+S\log\log N)\}_B)}/T!\right) + o(1).$$

Here log *denotes logarithm to base* $1/p$.

We also give a new proof based on the above theorem contrary to the extreme value theory approach where $[x]$ denote the usual integer part of x and {x} is the fractional part.

Remark 2. *The limiting distribution of the length of the longest head run containing T tails is the same as the limiting distribution of the length of the longest head*

run containing at most T tails. To prove it, let A be the event that the length of the longest head run containing at most T tails is greater than n. Then, $A = B \cup C$ where B is the event that the length of the longest head run containing precisely T tails is greater than n and C is the event that the length of a head run containing less than T tails is greater than n and it is not possible to add some tails to it. But

$$P(C) \leq \sum_{i=0}^{T-1} \binom{N}{i} p^{N-i} q^i \leq c p^N N^{T-1} \to 0$$

as $N \to \infty$.

In **Gordon et al. (1986)**, the original proof was based on extreme value theory, but here we give a new proof using the method of our Theorem. Let $[x]$ denote the usual integer part of x and $\{\mathrm{x}\}$ is the fractional part.

Proposition 2. *Let $\mu(N)$ denote the length of the longest T-contaminated run of heads during the coin tossing experiment of length N, then*

$$\mathbb{P}(\mu^T(N) - \mu_T(qN) \leq t) = \mathbb{P}\left(\left[\frac{W}{ln(\frac{1}{p})} + \{\mu_T(qN)\}\right] - \{\mu_T(qN)\} \leq t\right) + o(1)$$

for all t, where

$$\mu_T(qN) = \log(qN) + T \log\log(qN) + T \log(q/p) - \log(T!)$$

and W has an extreme value distribution $\mathbb{P}(W \leq t) = exp(-e^{-t})$.

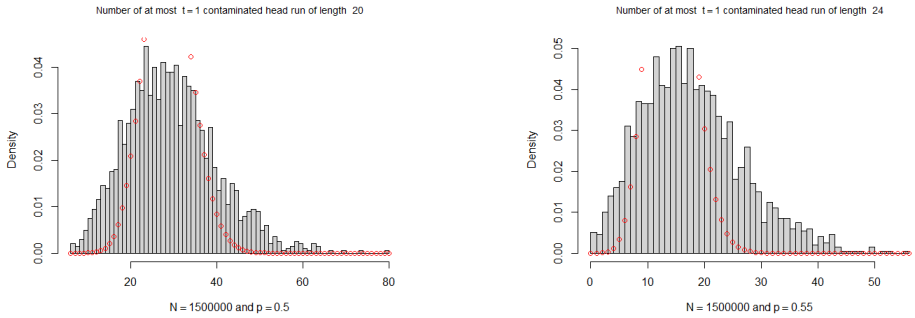
Remark 3. *We emphasize that the above proposition does not offer a limiting law for $\mu^T(N) - \mu_T(qN)$ but it gives a sequence of accompanying laws. The distances of the laws between the two sequences converge to 0 (as $n \to \infty$).*

Simulation Results

We chose sufficiently large lengths of the sequence N after which contaminated head runs of specified lengths n are investigated under varying probability values. We evaluate the contaminated runs and present results for the case of $T = 1$. We performed our simulations in R package.

Example 1 (Number of at most T-contaminated runs of heads.)**.** *The figures below show the empirical distribution of the at most T-contaminated head run and its approximation suggested by theorem 2 and denoted by the red dots.*

For 2000 simulations, $N = 1.5 \times 10^6, p = \{0.5, 0.55\}$ *and* $T = 1$ *we try out different run lengths* n *to generate our results. Analysis of the above figures reveal*



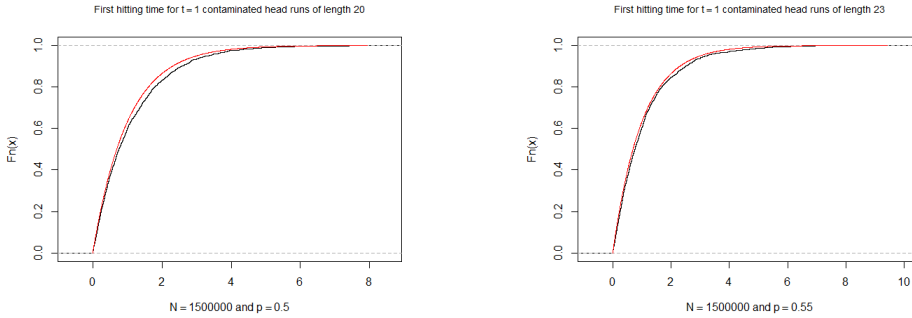
(a) (b)

Figure 1.1: Distribution of the length of at most $T = 1$ contaminated head run

a reasonable fit indicating convergence to the suggested compound Poisson distribution.

Example 2 (First hitting time for at most T-contaminated head runs of any specified length)**.** *The figures below show the empirical distribution of the first hitting time of the at most* T*-contaminated head run and its approximation suggested by theorem 3 and denoted by the red dotted line.*

For 2000 simulations, $N = 1.5 \times 10^6, p = \{0.5, 0.55\}$ *and* $T = 1$ *with various run lengths* n*, we obtain the results.*



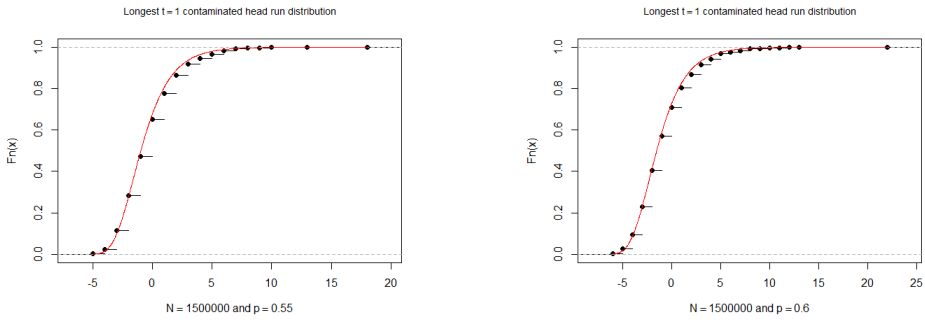
(a) (b)

Figure 1.2: Distribution of first hitting times for $T = 1$ contaminated head runs

The figures reveal near perfect fit between the empirical and theoretical distributions for $T = 1$ *even with higher values of* p*.*

Example 3 (Length of longest at most T-contaminated head runs)**.** *The figures below show the empirical distribution of the length of the longest T-contaminated head run and its approximation suggested by theorem 4 and denoted by the red doted line.*

This variable is independent of the length n and to investigate its properties, we consider $N = 1.5 \times 10^6, p = \{0.55, 0.6\}$ and $T = 1$.



(a) (b)

Figure 1.3: Distribution of the length of longest $T = 1$ contaminated head run

The figures reveal perfect fit between the empirical and theoretical distributions for $T = 1$ even with higher values of p. However, some slight left skewness is observed which generally tend to diminish as p increases.

Chapter 2

Convergence rate for the longest T-contaminated head runs

In this part of the summary, we list the results of our paper Fazekas, Fazekas, and Suja (2024). We describe the background of our problem, rate of convergence for the longest T-contaminated head runs. We consider the previous approximation provided by Theorem 1 of **Gordon et al. (1986)** and after performing some manipulations to the approximation of the length of the longest run, we state the following;

Proposition 3. *(**Gordon et al. (1986)**) Let $\mu^T(N)$ denote the length of the longest T-contaminated run of heads during the coin tossing experiment of length N. Let*

$$m_0(N) = \log(qN) + T\log(\log(qN)) + T\log(q/p) - \log(T!),$$

where log denotes the logarithm to base 1/p. Let $[m_0(N)]$ denote the integer part of $m_0(N)$ and $\{m_0(N)\}$ denote the fractional part of $m_(N)$. Then

$$\mathbb{P}(\mu^T(N) - [m_0(N)] < k) = \exp\Big(-p^{k-\{m_0(N)\}}\Big) + o(1).$$

where o(1) denotes a quantity converging to 0 as $N \to \infty$.

However, numerical experiments show that the above offered approximation is quite weak and we therefore aim at improving the result for the quite simple but

most important cases of $T = 1$ and $T = 2$.

Let us consider a set of independent and identically distributed random variables, denoted as $X_1, X_2, \cdots, X_N$ with $\mathbb{P}(X_i = 1) = p$ and $\mathbb{P}(X_i = 0) = q$, $i = 1, 2, \ldots, N$. Let $T \geq 0$ be a fixed non - negative integer.

We study the T-interrupted runs of heads which means that there are T zeros in an m length sequence of ones and zeros. So if we let m be a positive integer and $A_n = A_{n,m}$ to denote the occurrence of the event at the n^{th} step ,that is, there are precisely T zeros in the block of sequence $X_n, X_{n+1}, \ldots X_{n+m-1}$. Here, we clarify that the condition $X_{n-1} = 0$ is not assumed. Therefore, $\mathbb{P}(\bar{A}_1\bar{A}_2\cdots\bar{A}_N)$ is the probability that no event $A_1 = A_{1,m}$ occurred in any of the first N blocks of length m i.e the waiting time for the T-contaminated run of heads of length m described by A_1 is longer than N.

We let τ_m be the first hitting time of the T-contaminated run of heads having length m. We wish to find the asymptotic distribution of τ_m as $m \to \infty$.

Theorem 5. *Let $T = 1$ or $T = 2$, $0 < p < 1$. Let τ_m be the first hitting time for the T contaminated run of heads having length m. Then, for $x > 0$,*

$$\mathbb{P}(\tau_m \alpha P(A_1) > x) \sim e^{-x}$$

as $m \to \infty$. Here if $T = 1$, then $\alpha = q + \frac{2p^{m-1}-1}{m}$ and $P(A_1) = mp^{m-1}q$. However, when $T = 2$, then $\alpha = q - \frac{2}{m}$ and $P(A_1) = \binom{m}{2}p^{m-2}q^2$.

Remark 4. *One can show that the above Theorem is valid for $T = 2$ with $\alpha = q - \frac{2}{m} + \frac{2(m-2)}{m}p^{m-2} - \frac{2(m-4)}{m}p^{m-1}$.*

Now, we turn to the case of the length of the longest T-contaminated run of heads, provide the approximation of its length and the accompanying distribution from which the rate of convergence is evaluated.

Theorem 6. *Let $T = 1$ or $T = 2$, and let $0 < p < 1$ be fixed. Let $\mu^T(N)$ be the length of the longest T-contaminated run of heads during N times of coin tossing. Let*

$$\begin{aligned} m(N) =& \log(qN) + T\log(\log(qN)) + \\ &+ T^2\frac{\log(\log(qN))}{c\log(qN)} - \frac{T}{cq_0\log(qN)} - \frac{T^3}{2c}\left(\frac{\log(\log(qN))}{\log(qN)}\right)^2 + \\ &+ T^2\frac{\log(\log(qN))}{cq_0(\log(qN))^2} + T^3\frac{\log(\log(qN))}{(c\log(qN))^2} + \\ &+ \left(T\log(\frac{q}{p}) - \log(T!)\right)\left(1 + \frac{T}{c\log(qN)} - T^2\frac{\log(\log(qN))}{c(\log(qN))^2}\right), \end{aligned}$$

where log *denotes the logarithm to base* $1/p$ *and* $c = ln(1/p)$*, where ln denotes the natural logarithm to base e. Let* $[m(N)]$ *denotes the integer part of* $m(N)$ *while* $\{m(N)\}$ *denotes the fractional part of* $m(N)$*, i.e.* $\{m(N)\} = m(N) - [m(N)]$*. Then,*

$$\mathbb{P}(\mu^T(N) - [m(N)] < k) = e^{-p^{(k-\{m(N)\})\left(1-\frac{T}{c\log(qN)}+T^2\frac{\log(\log(qN))}{c(\log(qN))^2}\right)}}\left(1+O\left(\frac{1}{(\log N)^2}\right)\right)$$

for any integer k*, where* $f(N) = O(h(N))$ *means that* $f(N)/h(N)$ *is bounded as* $N \to \infty$*.*

Remark 5. *Using our method for* $T = 1$ *and* $T = 2$ *and for* $m_0(N)$ *from the above proposition* (3)*, we obtain that the rate of convergence is* $O(\log(\log(N))/\log(N))$*, that is;*

$$\mathbb{P}(\mu^T(N) - [m_0(N)] < k) = \exp\left(-p^{k-\{m_0(N)\}}\right)(1 + O(\log(\log(N)/\log(N)))).$$

By doing a comparison of the two approximations, it can be seen that our Theorem 6 *considerably improves Theorem 1 of* ***Gordon et al. (1986)*** *in the cases of* $T = 1$ *and* $T = 2$*.*

We now present preliminary proofs to some Lemmas in **Csáki et al. (1987)** which plays a fundamental role in the proofs of our theorems.

Lemma 1. *(main lemma, stationary case, finite form). Let* m *be fixed. Assume that* A_n *is stationary. Assume that there is a fixed number* p*,* $0 < p \leq 1$*, such that the following three conditions hold for some fixed* k *with* $2 \leq k \leq m$*, and fixed* ε *with* $0 < \varepsilon < \min\{p/10, 1/42\}$

(SI)

$$|\mathbb{P}(\bar{A}_2 \cdots \bar{A}_k | A_1) - p| < \varepsilon,$$

(SII)

$$\sum_{k+1 \leq i \leq 2m} \mathbb{P}(A_i | A_1) < \varepsilon,$$

(SIII)

$$P(A_1) < \varepsilon/m.$$

Then, for all $N > 1$*,*

$$\left|\frac{\mathbb{P}(\bar{A}_2 \cdots \bar{A}_N | A_1)}{\mathbb{P}(\bar{A}_2 \cdots \bar{A}_N)} - p\right| < 7\varepsilon$$

and

$$e^{-(p+10\varepsilon)NP(A_1)-2mP(A_1)} < \mathbb{P}(\bar{A}_1 \cdots \bar{A}_N) < e^{-(p-10\varepsilon)NP(A_1)+2mP(A_1)}. \quad (2.1)$$

We check conditions (SI) - (SIII) of the Lemma, for the case $k = m$ and try verifying them with appropriate choices of ε. This made it possible to determine the limiting distribution of the waiting time $\tau_m = \{\text{first } n; \text{ such that } A_n \text{ occurs}\}$.

Remark 6. *We first considered condition (SIII) and show that it is true for any T if m is large enough. We have*

$$P(A_1) = \binom{m}{T} p^{m-T} q^T \leq \frac{m^T}{T!} p^{m-T} q^T < \frac{\varepsilon}{m},$$

if

$$m^{T+1} p^m < \varepsilon \left(\frac{p}{q}\right)^T T!,$$

and the last inequality is satisfied for any positive ε if m is large enough.

Remark 7. *Consider condition (SII).*

$$\mathbb{P}(A_i|A_1) = P(A_i) = \binom{m}{T} p^{m-T} q^T \leq \frac{m^T}{T!} p^{m-T} q^T,$$

if $i > m$ because of independence. So

$$\sum_{i=m+1}^{2m} \mathbb{P}(A_i|A_1) = mP(A_1) = m\binom{m}{T} p^{m-T} q^T \leq m\frac{m^T}{T!} p^{m-T} q^T < \varepsilon,$$

therefore we obtain again condition (SIII) hence condition (SII) is true if m is large enough.

To check condition (SI) of the Lemma, we separately evaluated the joint probabilities $\mathbb{P}(A_1 \bar{A}_2 \cdots \bar{A}_k)$ taking into account different values of T. First, we fixed $T = 1$.

Lemma 2. *Condition (SI) of the Lemma, stationary case finite form is satisfied for $T = 1$ and $k = m$ in the following form*

$$|\mathbb{P}(\bar{A}_2 \cdots \bar{A}_m|A_1) - \alpha| < \varepsilon,$$

with $\alpha = q + \frac{2p^{m-1}-1}{m}$.

Lemma 3. *Condition (SI) of the Lemma, stationary case finite form is satisfied for $T = 2$ and $k = m$ in the following form*

$$|\mathbb{P}(\bar{A}_2\bar{A}_3\cdots\bar{A}_m|A_1) - \alpha| < \varepsilon,$$

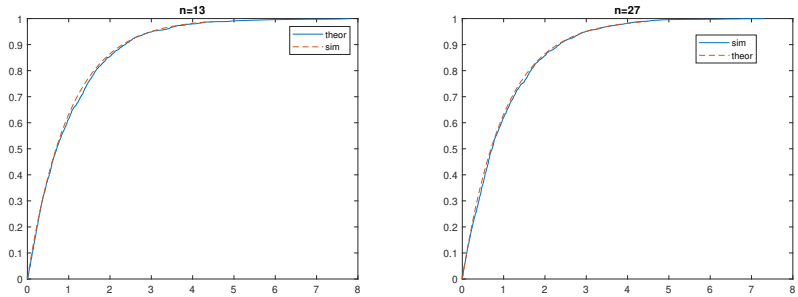
with $\alpha = q - \frac{2}{m} + O(p^m)$ as as $m \to \infty$

Remark 8. *A more careful calculation shows that the Lemma, stationary case finite form is valid for $T = 2$ with $\alpha = q - \frac{2}{m} + \frac{2(m-2)}{m}p^{m-2} - \frac{2(m-4)}{m}p^{m-1}$, too.*

Simulation Results

In this section, we begin by presenting simulation results that demonstrate the numerical behaviour of the first hitting time τ_m for a T-contaminated head run. The obtained findings provide empirical evidence in favor of Theorem 5.

Example 4. *Let $p = 0.5$, length of the coin tossing experiment denoted as $N = 10^6$, while the number of the repetitions of the experiment $s = 2000$.*



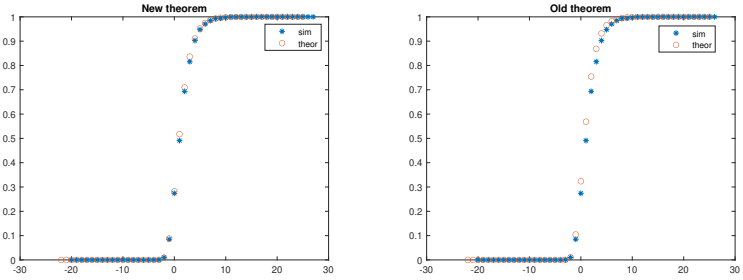
(a)First hitting time, $T = 1$ (b) First hitting time, $T = 2$

Figure 2.1: Comparison of empirical and asymptotic distribution

Figure 2.1 shows the first hitting time of the T-contaminated run with lengths of $n = 13$ and 27, corresponding to T values of 1 and 2, respectively. The empirical distribution, represented by the solid line in the simulation, is contrasted with the asymptotic theoretical distribution, depicted by the dashed line as described in Theorem 5. The fit of the item is satisfactory.

We now present simulation results for $\mu(N)$, i.e. for the length of the longest T-contaminated run. They show that our new approximation in Theorem 6 is better than the former one quoted in Proposition 3. We implemented the simulation in Matlab.

Example 5. *Let $p = 0.5$, $T = 1$, length of the coin tossing experiment $N = 10^6$ and the number of the repetitions of the experiment $s = 2000$. On parts (a) and (b) of Figure 2.2 sign $\circ$ shows the theoretical asymptotic probability and $*$ shows the relative frequency of those experiments when $\mu^T(N)$, that is the longest T-contaminated run is shorter than the given value on the horizontal axis.*



(a) Longest run (new) (b) Longest run (old)

Figure 2.2: Comparison of empirical and asymptotic distribution

Part (a) of Figure 2.2 shows the fit of the empirical distribution of $\mu(N)$ to the asymptotic distribution given by our Theorem 6. This shows a nice fit.
Part (b) of Figure 2.2 shows the fit of the empirical distribution of $\mu(N)$ to the asymptotic distribution given by the old result quoted in Proposition 3. The distribution does not fit nicely.

Convergence rate

In order to assess the numerical accuracy of the approximation to the limit distribution, the uniform distance measure, also known as Kolmogorov's distance measure, is employed. This measure is defined as

$$d_k(X,Y) \equiv d_k(F_X;F_Y) = \sup_x |F_X(x) - F_Y(x)|$$

, where X and Y represent random variables with distribution functions F_X and F_Y, respectively.The determination of the rate of convergence will be deduced through the utilization of Kolmogorov's distance measures.

Example 6 (Convergence rate)**.** *We performed the coin tossing experiment of length $N = 10^6$, with 2000 repetitions and calculated the Kolmogorov's distance. In the table below, T is the number of contaminations, p is the probability of heads. K_{old} is the Kolmogorov's distance between the empirical distribution of $\mu(N)$ and*

the asymptotic distribution given by the old result quoted in Proposition 3. The values are high hence indicating a poor fit. K_{new} is the Kolmogorov's distance between the empirical distribution of $\mu(N)$ and the asymptotic distribution given by our Theorem 6. The values are low hence indicating a relatively good fit.

T	p	K_{old}	K_{new}
1	0.5	0.0778	0.0264
2	0.4	0.1948	0.0172
2	0.5	0.2129	0.0148
2	0.6	0.1953	0.0250

Table 2.1: Kolmogorov's distance measure

Chapter 3

Limit theorems for runs containing two types of contaminations

In this chapter we list the results of our paper Fazekas, Fazekas, and Suja (2023). We defined and investigated the at most two-type contaminated sequence of runs with trinary trials. Let $X_1, X_2, \ldots, X_N$ be a sequence of independent random variables with three possible outcomes; 0, +1 and -1 labeled as success, failure of type I and failure of type II.

$\mathbb{P}(X_i = 0) = p$, $\mathbb{P}(X_i = +1) = q_1$ and $\mathbb{P}(X_i = -1) = q_2$ where $p + q_1 + q_2 = 1$ and $p > 0, \quad q_1 > 0, \quad q_2 > 0$.

An m length sequence is called a pure run if it contains only 0 values. It is called a one-type contaminated run if it contains precisely one non-zero element either a +1 or a -1. On the other hand, it is called a two-type contaminated run if contains precisely one +1, and one -1 while the rest of the elements are 0's.

A run is called at most two-type contaminated if it is either pure, or one-type contaminated, or two-type contaminated. So for an arbitrary fixed m, let $A_n = A_{n,m}$ denote the occurrence of the event at the n^{th} step ,that is, there is at most a two-type contaminated run in the sequence $X_n, X_{n+1}, \ldots X_{n+m-1}$ and let $\bar{A}_n$ be its non-occurrence.

Let τ_m be the first hitting time of the at most two-type contaminated run of heads having length m. We shall be interested in finding the limiting distribution of τ_m as $m \to \infty$ for the case of a sequence containing at most two types of contamination but no two of the same type.

Theorem 7. *Let* $\mathbb{P}(X_i = 0) = p$, $\mathbb{P}(X_i = +1) = q_1$ *and* $\mathbb{P}(X_i = -1) = q_2$ *be probabilities of success, failure of type I and failure of type II, respectively where* $p + q_1 + q_2 = 1$ *and* $p > 0, \quad q_1 > 0, \quad q_2 > 0$. *Let* τ_m *be the first hitting time of the at most two-type contaminated run of heads having length* m. *Then, for* $x > 0$,

$$\mathbb{P}(\tau_m \alpha P(A_1) > x) \sim e^{-x}$$

as $m \to \infty$. *Here*

$$\alpha = \frac{C_0 + \frac{1}{m}c_1 + \frac{1}{m(m-1)}C_2}{1 + \frac{p(1-p)}{(m-1)q_1q_2} + \frac{p^2}{m(m-1)q_1q_2}}$$

where; $C_0 = (q_1 + q_2)$, $C_1 = \frac{p(q_1^2+q_2^2)}{q_1q_2} - 1$, $C_2 = \frac{(q_1^2+q_2^2)p^2}{q_1q_2(p-1)} + \frac{p}{p-1} + \frac{2(2p+1)q_1q_2}{(p-1)^3}$ *and* $P(A_1) = p^m + m(1-p)p^{m-1} + m(m-1)p^{m-2}q_1q_2$

We again check the fulfilment of the conditions given in the main Lemma of **Csáki et al. (1987)** for the case of $k = m$ (for fixed m) and $0 < p \leq 1$, such that for $\varepsilon > 0$:

Remark 9. *First, we shall consider condition (SIII) and show that it is true for any large enough* m.

$$\begin{aligned} P(A_1) =& p^m + m(1-p)p^{m-1} + m(m-1)p^{m-2}q_1q_2 \\ =& m(m-1)p^{m-2}q_1q_2\left\{1 + \frac{p(1-p)}{(m-1)q_1q_2} + \frac{p^2}{m(m-1)q_1q_2}\right\} \leq \frac{\varepsilon}{m} \end{aligned}$$

This inequality is true for any positive ε *if* m *is large enough.*
If $m \approx \log N$, *then* $p^m \approx p^{\log N} \approx \frac{1}{N}$ *and then,* $\varepsilon \approx \frac{(\log N)^3}{N}$. *(Here,* log *denotes logarithm to base* $\frac{1}{p}$*)*

Remark 10. *Now, considering condition (SII), if* $i > m$, *then* A_i *and* A_1 *are independent, therefore*

$$\sum_{i=m+1}^{2m} \mathbb{P}(A_i|A_1) = mP(A_1) < \varepsilon$$

which gives precisely the previous assumption on satisfaction of condition (SIII).

Lemma 4. *Condition (SI) is satisfied for* $k = m$ *in the following form*

$$|\mathbb{P}(\bar{A_2}, \bar{A_3}, \cdots, \bar{A_m}|A_1) - \alpha| < \varepsilon$$

with

$$\alpha = \frac{C_0 + \frac{1}{m}c_1 + \frac{1}{m(m-1)}C_2}{1 + \frac{p(1-p)}{(m-1)q_1q_2} + \frac{p^2}{m(m-1)q_1q_2}}.$$

Let $\mu(N)$ be the length of the longest at most two-type contaminated run in $X_1, X_2, \ldots, X_N$. Then,
$\{\mu(N) < m\} \iff$ Any m length run in $X_1, X_2, \ldots, X_N$ is neither two-type contaminated nor one-type contaminated nor pure.

Theorem 8. *Let $0 < p < 1$ be fixed. Let $\mu(N)$ be the length of the longest at most two-type contaminated run in $X_1, X_2, \ldots, X_N$. Then for $k > 0$,*

$$\begin{aligned}\mathbb{P}(\mu(N) - [m(N)] < k) =& e^{-p^{H(k-\{m(N)\})+O\left(\frac{1}{(\log N)^3}\right)}} \\ =& e^{-p^{-(k-\{m(N)\})+O\left(\frac{1}{(\log N)^3}\right)}}\end{aligned}$$

Here,

$$m(N) = \log(C_0 p^{-2} q_1 q_2) + H(k - \{m(N)\}) + O\left(\frac{1}{(\log N)^3}\right)$$

, $C_0 = (q_1 + q_2)$ and

$$\begin{aligned}H(X) =& -X + \frac{2X}{C\log N} - \frac{4}{C}\frac{\log\log N}{(\log N)^2}X - \frac{2}{C}\frac{4\log\log N}{C(\log N)^3}X - \frac{1}{C}\frac{1}{(\log N)^2}X^2 + \\ &+\frac{8(\log\log N)^2}{C(\log N)^3}X + \frac{4(\log\log N)}{C(\log N)^3}X^2 + \frac{C_1 - C_0}{CC_0}\frac{1}{(\log N)^2}X + \\ &+\frac{4(C_1 - C_0)}{CC_0}\frac{\log\log N}{(\log N)^3}X \\ =& -X + \frac{2}{C\log N}X - \frac{4}{C}\frac{\log\log N}{(\log N)^2}X + \frac{C_1 - C_0}{CC_0}\frac{1}{(\log N)^2}X + \\ &+\left(\frac{4(C_1 - C_0)}{CC_0} - \frac{8}{C^2}\right)\frac{\log\log N}{(\log N)^3}X + \frac{8}{C}\frac{(\log\log N)^2}{(\log N)^3}X - \\ &-\frac{1}{C}\frac{1}{(\log N)^2}X^2 + \frac{4}{C}\frac{\log\log N}{(\log N)^3}X^2\end{aligned}$$

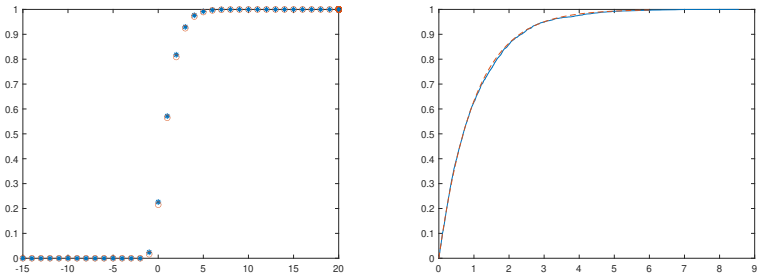
Simulation results

By considering analysis at the beginning of the proof of Theorem 8, we can see that the lemma of **Csáki et al. (1987)** offers good approximation if p is small, but it does not offer good approximation if p is close to 1. However, our simulation study show that the approximation for the longest run is very good for small values of p, but it is still appropriate if p is close to 1.

We performed several computer simulations for certain fixed values of p, q_1 and q_2. The left hand side part of each figure shows the empirical distribution of the longest at most two-type contaminated run and its approximation suggested by Theorem 8. Asterisk (i.e. $*$) denotes the result of the simulation, i.e. the empirical distribution of the longest at most two-type contaminated run and circle ($\circ$) denotes approximation offered by Theorem 8.

The right hand side of each figure shows the first hitting time of the m-length at most two-type contaminated run. Solid line shows the result of the simulation for the distribution function and dashed line shows the distribution function $1 - e^{-x}$ suggested by our Theorem 7.

Example 7. *Let $p = 1/3$, $q_1 = 1/3$, $q_2 = 1/3$. The length of the coin tossing experiment is $N = 3 \times 10^6$, the number of the repetitions of the experiment is $s = 3000$.*



(a) Longest run (b) First hitting time, $m = 16$

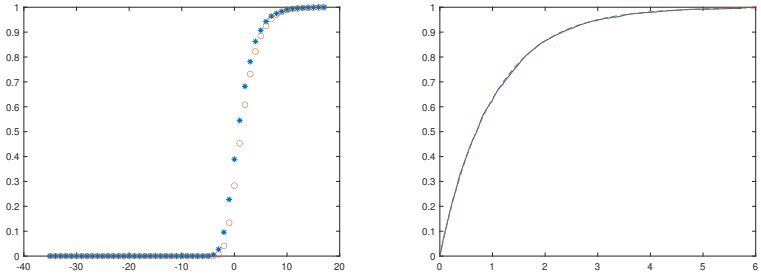
Figure 3.1: Longest at most two-type contaminated run and the first hitting time

.

The figure 3.1 shows the fit of the empirical distribution of $\mu(N)$ to the asymptotic distribution given by our Theorem 8. The fit is good.

Example 8. *Let $p = 0.6$, $q_1 = 0.2$, $q_2 = 0.2$. The length of the coin tossing experiment is $N = 4 \times 10^6$, the number of the repetitions of the experiment is $s = 3000$ in both cases.*

The figure 3.2 shows the fit of the empirical distribution of $\mu(N)$ to the asymptotic distribution given by our Theorem 8. The fit is not good.



(a) Longest run (b) First hitting time, $m = 34$

Figure 3.2: Longest at most two-type contaminated run and the first hitting time

Research Conference Participation

1. *Asymptotic theorems for contaminated runs of heads in the coin tossing experiment*, 27^{th} Conference of Young Statistician Meeting (YSM 2023), $29-1^{st}$ October 2023, Osijek, Croatia.

2. *Limit theorems for runs containing two types of contamination*, 9^{th} International Conference on Mathematics and Informatics (MATHINFO 2023), September $7-8^{th}$, 2023 Târgu Mureş/Marosvásárhely, Romania.

3. The 20^{th} Conference of the Applied Stochastic Models and Data Analysis International Society (ASMDA 2023) and Demographics2023 Workshop. A Hybrid Conference, $6-9^{th}$ June 2023, Heraklion, Crete, Greece.

4. *Convergence rate for the longest T contaminated runs of heads*, 12^{th} International Conference on Applied Informatics (ICAI 2023), March $2-4^{th}$ Eger, Hungary, 2023.

5. Workshop of Writing manuscripts for Official Statistics journals: Guidelines for practitioners and researchers, International Statistical Institute, Online, $23-25^{th}$ February, 2022.

6. *Asymptotic results for contaminated runs of heads*, 13^{th} Joint Conference on Mathematics and Computer Science (MaCS 2020), October $1-3^{rd}$, ELTE, Budapest 2020,(Virtual Conference).

Bibliography

E. Csáki, A. Földes, and J. Komlós. Limit theorems for Erdős-Rényi type problems. *Studia Sci. Math. Hungar*, 22:321–332, 1987.

P. Erdős and A. Rényi. On a new law of large numbers. *Analyse Math.*, 23:103–111, 1970.

I. Fazekas and M. Suja. Limit theorems for contaminated runs of heads. *Annales Univ. Sci*, 52:131–146, 2021.

I. Fazekas, B. Fazekas, and M. O. Suja. Limit theorems for runs containing two types of contaminations. Paper with detailed proofs. *arXiv preprint arXiv:2309.11602*, 2023.

I. Fazekas, B. Fazekas, and M. O. Suja. Convergence rate for the longest T-contaminated runs of heads. *Statistics & Probability Letters*, 208:110059, 2024. ISSN 0167-7152. doi: https://doi.org/10.1016/j.spl.2024.110059.

A. Földes. On the limit distribution of the longest head-run (in Hungarian). *Mat. Lapok*, 26(1–2):105–116, 1975.

A. Földes. The limit distribution of the length of the longest head-run. *Periodica Mathematica Hungarica*, 10(4):301–310, 1979.

L. Gordon, M. F. Schilling, and M. S. Waterman. An extreme value theory for long head runs. *Probability Theory and Related Fields*, 72(2):279–287, 1986.

B. A. Sevast'yanov. Limit Poisson law in a scheme of dependent random variables. *Teoriya Veroyatnostei i ee Primeneniya*, 17(4):733–738, 1972.

UNIVERSITY AND NATIONAL LIBRARY
UNIVERSITY OF DEBRECEN
H-4002 Egyetem tér 1, Debrecen
Phone: +3652/410-443, email: publikaciok@lib.unideb.hu

Registry number: DEENK/77/2024.PL
Subject: PhD Publication List

Candidate: Michael Ochieng Suja
Doctoral School: Doctoral School of Mathematical and Computational Sciences
MTMT ID: 10083960

List of publications related to the dissertation

Foreign language scientific articles in Hungarian journals (2)

1. Fazekas, I., Fazekas, B., **Suja, M. O.**: Limit theorems for runs containing two types of contaminations.
   Period. Math. Hung. [Accepted by publisher] (-), 1-25, 2024. ISSN: 0031-5303.
   IF: 0.8 (2022)
2. Fazekas, I., **Suja, M. O.**: Limit theorems for contaminated runs of heads.
   Ann. Univ. Sci. Budapest, Sect. Comp. 52, 131-146, 2021. ISSN: 0138-9491.

Foreign language scientific articles in international journals (1)

3. Fazekas, I., Fazekas, B., **Suja, M. O.**: Convergence rate for the longest T-contaminated runs of heads.
   Stat. Probab. Lett. 208, 1-8, 2024. ISSN: 0167-7152.
   DOI: http://dx.doi.org/10.1016/j.spl.2024.110059
   IF: 0.8 (2022)

Total IF of journals (all publications): 1,6
Total IF of journals (publications related to the dissertation): 1,6

The Candidate's publication data submitted to the iDEa Tudóstér have been validated by DEENK on the basis of the Journal Citation Report (Impact Factor) database.

08 March, 2024