

Poisson Approximation and the Chen-Stein Method

Marcin Mider Ella Kaye Beniamino Hadr-Amar

University of Warwick

November 6th, 2015

Introduction

Poisson Law of Small Numbers

Let $W \sim \text{Bin}(n, \lambda/n)$, $\lambda > 0$, and let $Z \sim \text{Poi}(\lambda)$.

Then, as $n \rightarrow \infty$:

$$\mathbb{P}(W = k) \xrightarrow{d} e^{-\lambda} \frac{\lambda^k}{k!} = \mathbb{P}(Z = k), \quad k \in \mathbb{Z}^+.$$

In other words,

$$d_{TV}(W, Z) \rightarrow 0, \quad \text{where} \quad d_{TV}(W, Z) = \sup_{A \subseteq \mathbb{Z}^+} |W(A) - Z(A)|$$

Introduction

Questions

- ▶ Relax assumption of *independence*?
- ▶ Relax assumption of *identically distributed*?
- ▶ How good is the Poisson approximation?

Chen-Stein operator

$$A_\lambda g(x) := \lambda g(x+1) - xg(x),$$

for every bounded function $g : \mathbb{Z}^+ \rightarrow \mathbb{R}$

Introduction

- ▶ W is distributed as $Z \sim \text{Poisson}(\lambda)$, if and only if

$$\mathbb{E}[A_\lambda g(W)] = 0.$$

Introduction

- ▶ W is distributed as $Z \sim \text{Poisson}(\lambda)$, if and only if

$$\mathbb{E}[A_\lambda g(W)] = 0.$$

- ▶ We have that

$$d_{TV}(W, Z) \leq |\mathbb{E}[A_\lambda g(W)]|$$

Introduction

- ▶ W is distributed as $Z \sim \text{Poisson}(\lambda)$, if and only if

$$\mathbb{E}[A_\lambda g(W)] = 0.$$

- ▶ We have that

$$d_{TV}(W, Z) \leq |\mathbb{E}[A_\lambda g(W)]|$$

- ▶ To show that W is close to Z , we have to check

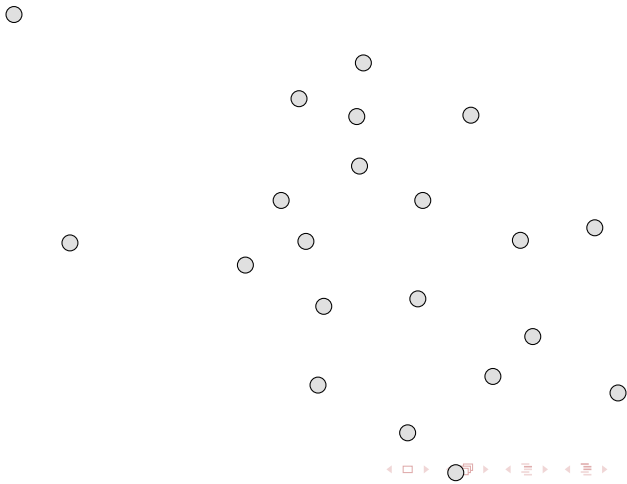
$$\mathbb{E}[A_\lambda g(W)] \rightarrow 0.$$

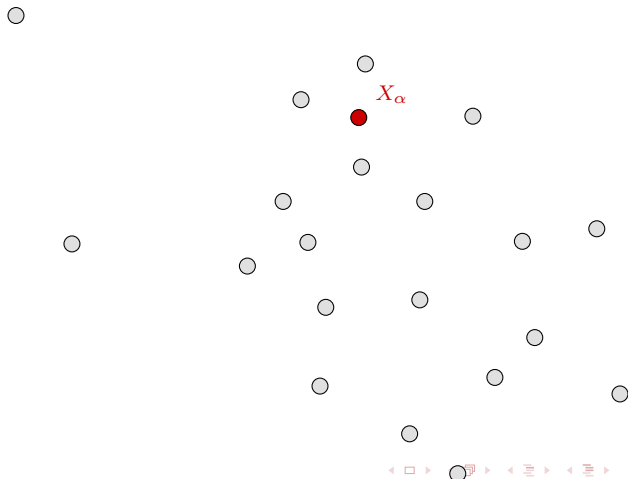
General setting

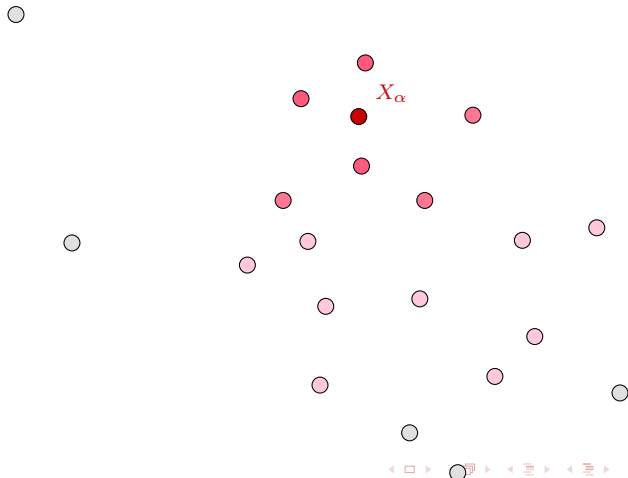
Let $X_\alpha, \alpha \in I$, with I a countable index set and:

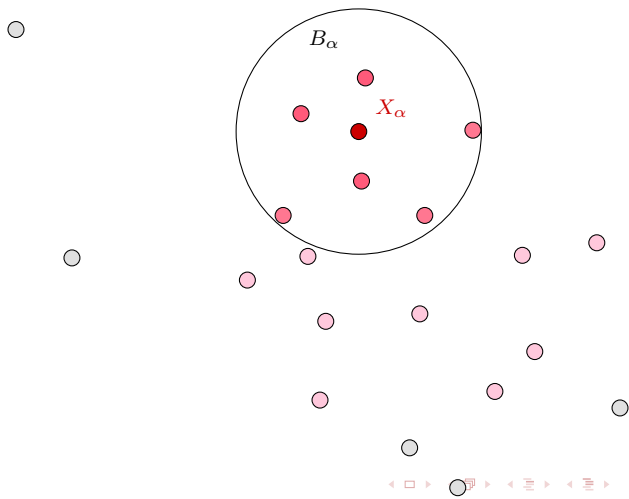
$$\mathbb{P}(X_\alpha = 1) = 1 - \mathbb{P}(X_\alpha = 0) = p_\alpha.$$

Define $W := \sum_{\alpha \in I} X_\alpha$, with $\lambda := \mathbb{E}[W]$,
and a neighbourhood B_α

Neighbourhood B_α 

Neighbourhood B_α 

Neighbourhood B_α 

Neighbourhood B_α 

General setting

$$b_1 = \sum_{\alpha \in I} \sum_{\beta \in B_\alpha} p_\alpha p_\beta, \quad (1)$$

$$b_2 = \sum_{\alpha \in I} \sum_{\beta \in B_\alpha \setminus \{\alpha\}} \mathbb{E}[X_\alpha X_\beta], \quad (2)$$

$$b_3 = \sum_{\alpha \in I} \mathbb{E} \left[\mathbb{E}[X_\alpha - p_\alpha | \sigma(X_\beta : \beta \notin B_\alpha)] \right]. \quad (3)$$

Chen-Stein bound

Theorem

Let $W = \sum_{\alpha} X_{\alpha}$, with $\lambda = \mathbb{E}[W] < \infty$ and let $Z \sim \text{Pois}(\lambda)$.

Then:

$$\|\mathcal{L}(W) - \mathcal{L}(Z)\|_{TV} \leq 2 \left[(b_1 + b_2) \frac{1 - e^{-\lambda}}{\lambda} + b_3 \left(1 \wedge \frac{1.4}{\sqrt{\lambda}} \right) \right],$$

and

$$|\mathbb{P}(W = 0) - e^{-\lambda}| \leq (b_1 + b_2 + b_3) \frac{1 - e^{-\lambda}}{\lambda}.$$



We have n people in the room and we are looking for a k -way coincidence.

Assume d days in the year, and a uniform distribution for birthdays throughout the year (i.e. the probability of being born on any given day is $\frac{1}{d}$).

- ▶ Let $\{1, \dots, n\}$ denote the group of individuals.
- ▶ Let the index set $I \equiv \{\alpha \subset \{1, \dots, n\} : |\alpha| = k\}$.

- ▶ Let $\{1, \dots, n\}$ denote the group of individuals.
- ▶ Let the index set $I \equiv \{\alpha \subset \{1, \dots, n\} : |\alpha| = k\}$.
- ▶ Let X_α be a Bernoulli random variable, which takes the value 1 when all the k people indexed by α share a birthday.
- ▶ This happens with probability $p_\alpha = (\frac{1}{d})^{k-1} = d^{1-k}, \forall \alpha$.

- ▶ Let $\{1, \dots, n\}$ denote the group of individuals.
- ▶ Let the index set $I \equiv \{\alpha \subset \{1, \dots, n\} : |\alpha| = k\}$.
- ▶ Let X_α be a Bernoulli random variable, which takes the value 1 when all the k people indexed by α share a birthday.
- ▶ This happens with probability $p_\alpha = (\frac{1}{d})^{k-1} = d^{1-k}$, $\forall \alpha$.
- ▶ Then W , the total number of coincidences, is given by $W = \sum_{\alpha \in I} X_\alpha$,
- ▶ And $\lambda = \mathbb{E}[W] = \sum_{\alpha \in I} d^{1-k} = \binom{n}{k} d^{1-k}$.

- ▶ Let $\{1, \dots, n\}$ denote the group of individuals.
- ▶ Let the index set $I \equiv \{\alpha \subset \{1, \dots, n\} : |\alpha| = k\}$.
- ▶ Let X_α be a Bernoulli random variable, which takes the value 1 when all the k people indexed by α share a birthday.
- ▶ This happens with probability $p_\alpha = (\frac{1}{d})^{k-1} = d^{1-k}$, $\forall \alpha$.
- ▶ Then W , the total number of coincidences, is given by $W = \sum_{\alpha \in I} X_\alpha$,
- ▶ And $\lambda = \mathbb{E}[W] = \sum_{\alpha \in I} d^{1-k} = \binom{n}{k} d^{1-k}$.
- ▶ Approximate W with a Poisson random variable, Z , with mean λ .

Classic case: $n = 23, k = 2, d = 365$

Classic case: $n = 23, k = 2, d = 365$

$$\mathbb{P}(W = 0) = \prod_{i=1}^{n-1} \left(1 - \frac{i}{d}\right) = 0.4927$$

Classic case: $n = 23, k = 2, d = 365$

$$\mathbb{P}(W = 0) = \prod_{i=1}^{n-1} \left(1 - \frac{i}{d}\right) = 0.4927$$

$$\mathbb{P}(Z = 0) = e^{-\lambda} = \exp \left\{ -\binom{n}{k} d^{1-k} \right\} = 0.4999982$$

Classic case: $n = 23, k = 2, d = 365$

$$\mathbb{P}(W = 0) = \prod_{i=1}^{n-1} \left(1 - \frac{i}{d}\right) = 0.4927$$

$$\mathbb{P}(Z = 0) = e^{-\lambda} = \exp \left\{ -\binom{n}{k} d^{1-k} \right\} = 0.4999982$$

The approximation is always conservative when birthdays are uniform.

- ▶ If $\alpha \cap \beta = \emptyset$, then X_α and X_β are independent.

- ▶ If $\alpha \cap \beta = \emptyset$, then X_α and X_β are independent.
- ▶ Therefore, take the neighbourhood of dependence to be

$$B_\alpha = \{\beta \in I : \alpha \cap \beta \neq \emptyset\}.$$

- ▶ If $\alpha \cap \beta = \emptyset$, then X_α and X_β are independent.
- ▶ Therefore, take the neighbourhood of dependence to be

$$B_\alpha = \{\beta \in I : \alpha \cap \beta \neq \emptyset\}.$$

- ▶ With this choice,

$$b_3 = \sum_{\alpha \in I} \mathbb{E}[|\mathbb{E}[X_\alpha - p_\alpha] | \sigma(X_\beta : \beta \notin B_\alpha)|] = 0.$$

$$\begin{aligned} b_1 &= \sum_{\alpha \in I} \sum_{\beta \in B_\alpha} p_\alpha p_\beta \\ &= \binom{n}{k} \left\{ \binom{n}{k} - \binom{n-k}{k} \right\} d^{2-2k} \end{aligned}$$

$$\begin{aligned} b_2 &= \sum_{\alpha \in I} \sum_{\beta \in B_\alpha \setminus \{\alpha\}} \mathbb{E}[X_\alpha X_\beta] \\ &= \sum_{j=1}^{k-1} \binom{n}{k} \binom{k}{j} \binom{n-k}{k-j} d^{1+j-2k} \end{aligned}$$



Bounds as n increases, for fixed d

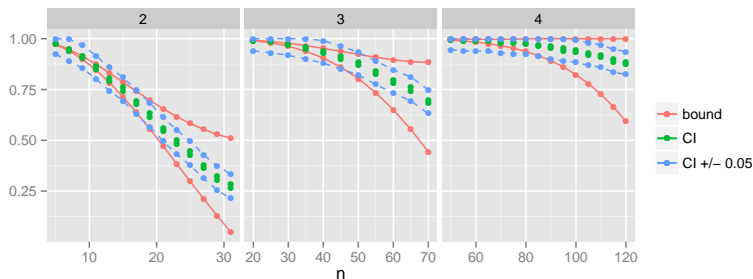


Figure: Simulations for $\mathbb{P}(W = 0)$, compared to the bounds given by the Chen-Stein method. The bounds are good when they (the red lines) are inside the blue lines, i.e. no more that 0.05 away from the simulated values. The bounds widen as n increases, for fixed $d = 365$, for each of $k = 2, 3, 4$.

Bounds as d increases, for fixed n

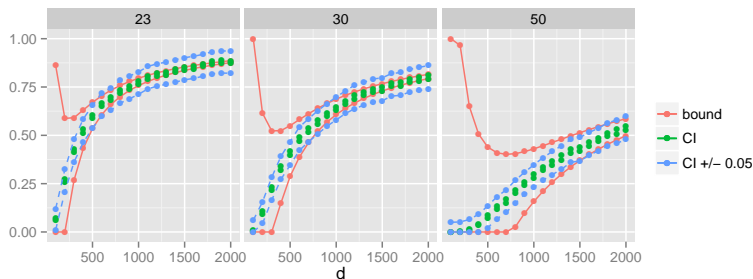


Figure: Simulations for $\mathbb{P}(W = 0)$, compared to the bounds given by the Chen-Stein method. The bounds are good when they (the red lines) are inside the blue lines, i.e. no more that 0.05 away from the simulated values. The bounds widen as d increases, for fixed $k = 2$, for each of $n = 23, 30, 50$.

- Take both $n, d \rightarrow \infty$. We do this in such a way that λ/d stays bounded away from zero and ∞ , denoted $\lambda \asymp d$.

- ▶ Take both $n, d \rightarrow \infty$. We do this in such a way that $\lambda/1$ stays bounded away from zero and ∞ , denoted $\lambda \asymp 1$.
- ▶ This implies that $n^k \asymp d^{k-1}$ (since $\lambda = \binom{n}{k} d^{1-k}$).

- ▶ Take both $n, d \rightarrow \infty$. We do this in such a way that $\lambda/1$ stays bounded away from zero and ∞ , denoted $\lambda \asymp 1$.
- ▶ This implies that $n^k \asymp d^{k-1}$ (since $\lambda = \binom{n}{k} d^{1-k}$).
- ▶ We fixed the ratio $\frac{n^k}{d^{1-k}}$ at 1.45 (the value it takes in the classic case).

- ▶ Take both $n, d \rightarrow \infty$. We do this in such a way that $\lambda/1$ stays bounded away from zero and ∞ , denoted $\lambda \asymp 1$.
- ▶ This implies that $n^k \asymp d^{k-1}$ (since $\lambda = \binom{n}{k} d^{1-k}$).
- ▶ We fixed the ratio $\frac{n^k}{d^{1-k}}$ at 1.45 (the value it takes in the classic case).
- ▶ The order of the Chen-Stein bound here is the same as the order of b_2 , which is

$$n^{k+1} d^{-k} \asymp n^{-1/(k-1)}.$$

Thus the Chen-Stein method yields that the total variation distance decays at a rate no slower than $O(n^{-1/(k-1)})$

Bounds as $n, d \rightarrow \infty$

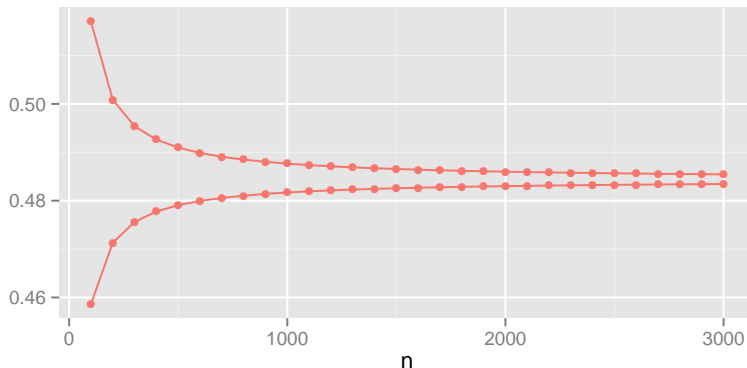
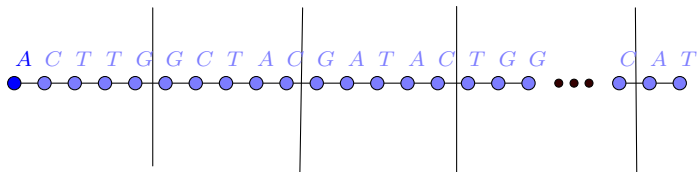


Figure: The Chen-Stein bounds on the distance between $\mathbb{P}(W = 0)$ and $\mathbb{P}(Z = 0)$ as n increases, keeping the ratio $\frac{n^k}{d^{1-k}}$ constant at 1.45

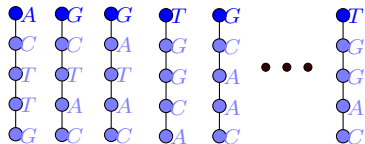
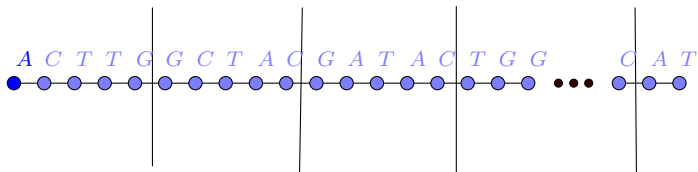
DNA example



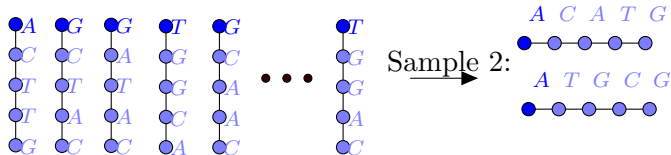
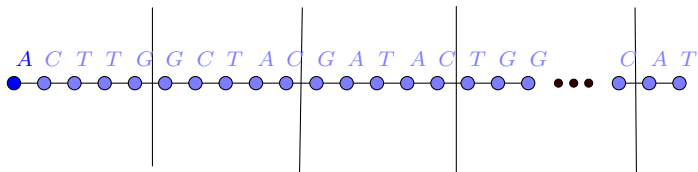
DNA example



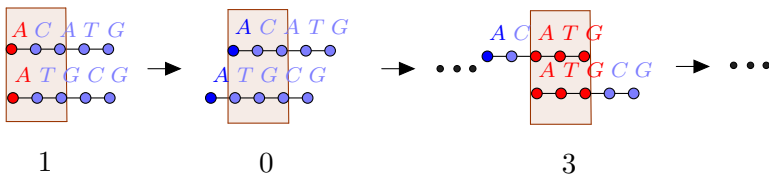
DNA example



DNA example



DNA example



DNA example

We used a complete chloroplast genome of *Marchantia Polymorpha* (Liverwort), downloaded from GenBank. It consists of one sequence of 121,024 letters.



DNA example - algorithm

Data: Cut the sequence into 236 disjoint *stripes*, each consisting of 512 letters. Discard remaining letters.

w.size = 21; max.matches = vector[200];

for $i = 1 : 200$ **do**

 draw 2 *stripes* at random, store as str.A and str.B;

 current.max = 0;

for each possible placement of window of length w.size on str.A and str.B **do**

 current.count = number of matches within window;

if $\text{current.count} > \text{current.max}$ **then**

 update current.max;

end

end

 max.matches[i] = current.max

end

DNA example

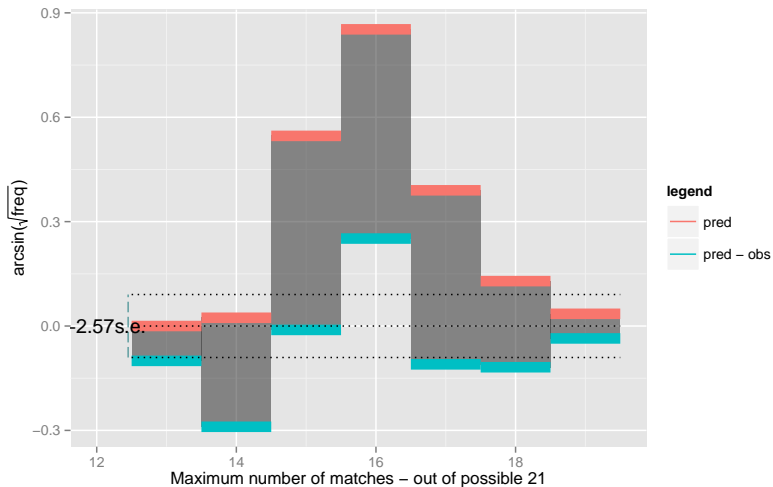
Suppose that A_1, A_2, \dots, A_n and B_1, B_2, \dots, B_n are two *stripes*, where $A_i, B_i \in \{a, c, t, g\}$, chosen at random according to common distribution μ . Define:

$$M_n(t) = \max_{1 \leq i, j \leq n-t+1} \sum_{k=0}^{t-1} \mathbb{1}_{A_{i+k} = B_{j+k}}, \quad (4)$$

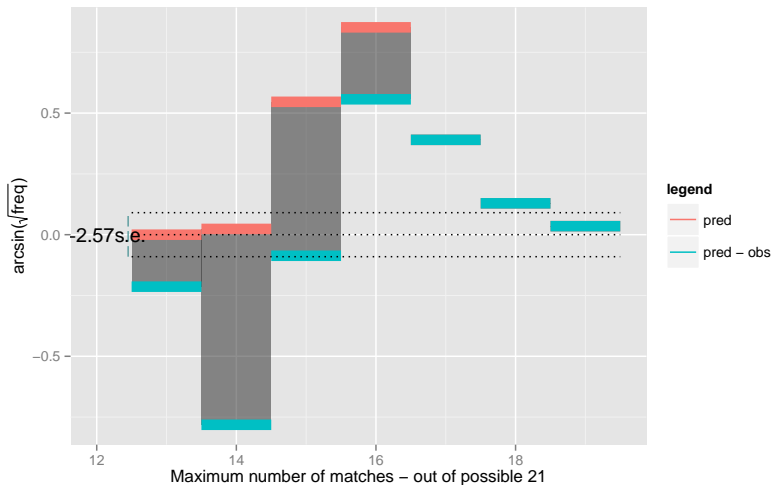
Then, under some regularity conditions:

$$\mathbb{P}[M_n(t) < s] - e^{-n(\frac{s}{t}-p)} \mathbb{P}[\text{Bin}(t,p) \geq s] \rightarrow 0. \quad (5)$$

DNA example



DNA example



The Length of the Longest Head Run

Consider

- ▶ n independent coin tosses $C_i \sim \text{Ber}(p)$
- ▶ test length t

Goal: What's the probability of having a run of *at least* t heads in the sequence of tosses?

Runs could occur in **clumps**:

- $[0, 1, 0, \textcolor{red}{1}, \textcolor{red}{1}, \textcolor{red}{1}, 1, 0, 1]$
- $[0, 1, 0, 1, \textcolor{red}{1}, \textcolor{red}{1}, \textcolor{red}{1}, 0, 1]$

We count only the first sequence of t heads.

The Length of the Longest Head Run

Let's define

$$Y_\alpha = \prod_{i=\alpha}^{\alpha+t-1} C_i.$$

The Length of the Longest Head Run

Let's define

$$Y_\alpha = \prod_{i=\alpha}^{\alpha+t-1} C_i.$$

and de-clump

$$X_\alpha = (1 - C_{\alpha-1})Y_\alpha, \quad \text{where } X_1 = Y_1.$$

The Length of the Longest Head Run

Let's define

$$Y_\alpha = \prod_{i=\alpha}^{\alpha+t-1} C_i.$$

and de-clump

$$X_\alpha = (1 - C_{\alpha-1})Y_\alpha, \quad \text{where } X_1 = Y_1.$$

Therefore

$$W = \sum_{\alpha \in I} X_\alpha \approx \text{Poi}(\lambda) \quad \text{where } \lambda = \mathbb{E}(W).$$

$$\lambda = p^t[(1-p)(n-1) + 1].$$

The Length of the Longest Head Run

- Neighbourhood of dependence

$$B_\alpha = \{\beta \in I : |\alpha - \beta| \leq t\}$$

The Length of the Longest Head Run

- ▶ Neighbourhood of dependence

$$B_\alpha = \{\beta \in I : |\alpha - \beta| \leq t\}$$

- ▶ $b_3 = 0$

$$\mathbb{E}[X_\alpha - p_\alpha \mid \sigma(X_\beta : \beta \notin B_\alpha)] = \mathbb{E}(X_\alpha - p) = 0$$

The Length of the Longest Head Run

- ▶ Neighbourhood of dependence

$$B_\alpha = \{\beta \in I : |\alpha - \beta| \leq t\}$$

- ▶ $b_3 = 0$

$$\mathbb{E}[X_\alpha - p_\alpha \mid \sigma(X_\beta : \beta \notin B_\alpha)] = \mathbb{E}(X_\alpha - p) = 0$$

- ▶ $b_2 = 0$

$$X_\alpha = 1 \iff X_\beta = 0 \quad \forall \quad \beta \in B_\alpha, \beta \neq \alpha$$

$$\mathbb{E}(X_\alpha X_\beta) = 0.$$

The Length of the Longest Head Run

- ▶ Neighbourhood of dependence

$$B_\alpha = \{\beta \in I : |\alpha - \beta| \leq t\}$$

- ▶ $b_3 = 0$

$$\mathbb{E}[X_\alpha - p_\alpha \mid \sigma(X_\beta : \beta \notin B_\alpha)] = \mathbb{E}(X_\alpha - p) = 0$$

- ▶ $b_2 = 0$

$$X_\alpha = 1 \iff X_\beta = 0 \quad \forall \quad \beta \in B_\alpha, \beta \neq \alpha$$

$$\mathbb{E}(X_\alpha X_\beta) = 0.$$

- ▶ $b_1 < \lambda^2(2t+1)/n + 2\lambda p^t$

The Length of the Longest Head Run

Example: Consider a sequence $n = 2047$ coin tosses, $p = 0.5$.

“What’s the probability of obtaining a run of at least $t = 14$ heads?”

