On moderate deviations in Poisson approximation

Aihua Xia (夏爱华)

School of Mathematics and Statistics
The University of Melbourne, VIC 3010

24 June, 2019

Charles Stein’s Influence on Probability
(Based on a joint work with Qingwei Liu)
Why?

• Distributional approximation pays little attention to the tail probabilities.

• In statistical inference, the tail probabilities matter!

• The error estimates of distributional approximation are useless because the tail probabilities are often significantly smaller than the error estimates.
What’s the moderate deviation?

Petrov (1975), p. 228: let $X_i$, $1 \leq i \leq n$, be independent and identically distributed (i.i.d.) random variables with $E(X_1) = 0$ and $\text{Var}(X_1) = 1$, if for some $t_0 > 0$,

$$\mathbb{E} e^{t_0 |X_1|} \leq c_0 < \infty,$$

then there exist positive constants $c_1$ and $c_2$ depending on $c_0$ and $t_0$ such that

$$\mathbb{P} \left( \frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i \geq z \right) = 1 + O(1) \frac{1 + z^3}{\sqrt{n}}, \quad 0 \leq z \leq c_1 n^{1/6},$$

where $\Phi(z)$ is the distribution function of the standard normal, $|O(1)| \leq c_2$.

- $c_1$ and $c_2$?
- The range of values of $n$?
Why do we need Poisson?

• Since the pioneering work Chen (1975), it has been shown that, for the counts of rare events, Poisson distribution and its “relatives” provide a better approximation in terms of stronger metrics.

• BUT for the tail probabilities, we don’t need the stronger metrics, can’t we use normal?
Example

- Let \( \{X_i : 1 \leq i \leq n\} \) be iid with a continuous cumulative distribution function, we are interested in the distribution of records in \( \{X_i\} \).

- \( X_1 \) is always a record: ignore it.

- For \( 2 \leq i \leq n \), \( X_i \) is a record if \( X_i > \max_{1 \leq j \leq i-1} X_j \).

- \( I_i := 1[X_i > \max_{1 \leq j \leq i-1} X_j] \).

- \( S_n := \sum_{i=2}^{n} I_i \).
Approximations of $S_n$

- $\mathbb{E} I_i = 1/i$ and $\{I_i : 2 \leq i \leq n\}$ are independent.
- $\lambda_n := \mathbb{E} S_n = \sum_{i=2}^{n} \frac{1}{i}$; $\sigma_n^2 = \text{Var}(S_n) = \sum_{i=2}^{n} \frac{1}{i} (1 - \frac{1}{i})$.
- $\lambda_n - \sigma_n \in (0, 1)$.

Under the Kolmogorov distance, the error of
  - normal approximation is $O(\log^{-1/2} n)$,
  - Poisson approximation is $O(\log^{-1} n)$.  

[Slide 5]
How about the tail probabilities?

We consider the tail probabilities $\mathbb{P}(S_n \geq v_n)$ with $v_n := \lambda_n + 3 \cdot \sigma_n$ and compare $\mathbb{P}(S_n \geq v_n)$ with moderate deviations based on $\text{Pn}(\lambda_n)$, $\text{Pn}(\sigma_n^2)$, $N_n \sim N(\lambda_n, \sigma_n^2)$. 
\[ \frac{\mathbb{P}(S_n \geq v_n)}{\mathbb{P}(N_n \geq v_n)} \]
\( \frac{P(S_n \geq v_n)}{P(N_n \geq v_n)} \) with 0.5 correction
\[ \frac{\mathbb{P}(S_n \geq \nu_n)}{P_n(\lambda_n)([\nu_n, \infty))} \]
$\mathbb{P}(S_n \geq v_n) / \mathbb{P}_n(\sigma^2_n([v_n, \infty)))$
The winner is $P_n(\lambda_n)$

[Slide 11]
Literature?


Pn(1) tails vs Normal tails (with and without .5 correction)
One std to four std away from the mean

**Graphs:**
- $P_n(4)/N(4,4)$
- $P_n(16)/N(16,16)$
- $P_n(100)/N(100,100)$
- $P_n(10000)/N(10000,10000)$
One std to six std away from the mean
Conclusions

Poisson(λ) vs N(λ, λ):

- Poisson has a heavier tail than normal tail;
- there is an acceleration of the ratio of the tail probabilities beyond a few standard deviations when λ is not large enough;
- unlike normal, looking at the # of standard deviations away does not work for Pn when λ is not large enough;
- a small change of the value of λ has significant impact on its moderate deviations.
Pn: mean or var?

- $W_n \sim \text{Bi}(n, p)$ with $0 < p < 1$, $Y_n \sim \text{Pn}(np)$ and $Z \sim N(0, 1)$, then for a fixed $x > 0$, as $n \to \infty$,

\[
\frac{\mathbb{P}(W_n \geq np + x \sqrt{np(1-p)})}{\mathbb{P}(Y_n \geq np + x \sqrt{np(1-p)})} \sim \frac{\mathbb{P}(Z \geq x)}{\mathbb{P}(Z \geq x \sqrt{1-p})}.
\]

- The error systematically deviates from 1 as $x$ moves away from 0.
• Introducing an adjustment into the approximate models:
  for a fixed $x > 0$, as $n \to \infty$,
  \[
  \frac{\mathbb{P}(W_n \geq np + x\sqrt{np(1-p)})}{\mathbb{P}(Y_n \geq np + x\sqrt{np})} \sim 1
  \]
  or equivalently, with $Y'_n \sim \text{Pn}(np(1-p))$,
  \[
  \frac{\mathbb{P}(W_n \geq np + x\sqrt{np(1-p)})}{\mathbb{P}(Y'_n \geq np(1-p) + x\sqrt{np(1-p)})} \sim 1.
  \]

• **Conclusion:**
  - the variance of the Pn matters and it must be large to have good moderate deviation approximation;
  - either we have to shift the mean of the Pn or twist $K$ in $\mathbb{P}(\text{Pn} \geq K)$ to remove the systematic bias.
$\sigma_n^2 = 0.09n$, the plot of $\left( \frac{\text{Bi}(n, 0.1)([0.1n + 3\sigma_n, \infty))}{\text{Pn}(\sigma_n^2)([\sigma_n^2 + 3\sigma_n, \infty))} - 1 \right) \times \sqrt{n}$:
Zoom in

![Graph showing Mod Dev Error against n, with the formula (Bin tail/Pn tail-1) * root(n).]
Pn vs Bi\((n, 0.01)\)

\[
\sigma^2_n = 0.0099n, \quad \left( \frac{\text{Bi}(n, 0.01)([0.01n+3\sigma_n, \infty))}{\text{Pn}(\sigma^2_n([\sigma^2_n+3\sigma_n, \infty)))} - 1 \right) \times \sqrt{n}:
\]

![Graph showing Mod Dev Error (Bin tail/Pn tail-1)*root(n)]
Zoom in

Mod Dev Error

$\left(\frac{\text{Bin tail}}{\text{Pn tail}} - 1\right) \times \sqrt{n}$

[Slide 22]
Poisson moderate deviation
– Chen, Fang & Shao (2013)

• For a non-negative random variable $W$ with mean $\lambda > 0$, rv $W^s$ is said to have $W$-size biased distribution if

$$E[Wf(W)] = \lambda E[f(W^s)]$$

for all suitable functions $f$.

• If $W$ and $W^s$ are defined on the same probability space, then

$$d_{TV}(\mathcal{L}(W), Pn(\lambda)) \leq (1 - e^{-\lambda})E|W + 1 - W^s|.$$
• Assume that $\Delta := W + 1 - W^s \in \{-1, 0, 1\}$ and there are non-negative constants $\delta_1, \delta_2$ such that

$$P(\Delta = -1|W) \leq \delta_1, \quad P(\Delta = 1|W) \leq \delta_2 W.$$ 

For integers $k \geq \lambda$, let $\xi = (k - \lambda)/\sqrt{\lambda}$, there exists positive constants $c$ and $C$, such that for $(\delta_1 + \delta_2 \lambda)(1 + \xi^2) \leq c$, we have

$$\left| \frac{P(W \geq k)}{P_{n}(\lambda)([k, \infty))} - 1 \right| \leq C \frac{(\delta_1 + \delta_2 \lambda)(1 + \xi^2)}{\leq c}.$$
A toy example

- Let $X_i$'s be Bernoulli rvs with $\mathbb{P}(X_i = 1) = p_i$, $1 \leq i \leq n$.
- $W = \sum_{i=1}^{n} X_i$.
- For the size-biased distribution, we let $W_i = W - X_i$, $\mu = \mathbb{E}(W)$, then

$$\mathbb{E}[W f(W)] = \sum_{i=1}^{n} \mathbb{E}[X_i f(W)] = \mu \sum_{i=1}^{n} (p_i/\mu) \mathbb{E} f(W_i + 1).$$

- Let $I$ be a $\{1, \ldots, n\}$-valued rv independent of $\{X_i\}$ and having $\mathbb{P}(I = i) = p_i/\mu$, then $W^s := W_I + 1$. 
• \( \Delta = W + 1 - W_s = X_I \), so \( \mathbb{P}(\Delta = -1|W) = 0, \)

\[
\mathbb{P}(\Delta = 1|W) = \mathbb{E}(X_I|W) = \sum_{i=1}^{n} \left( \frac{p_i}{\mu} \right) \mathbb{P}(X_i = 1|W)
\]

\[
\leq \frac{\text{max } p_i}{\mu} \sum_{i=1}^{n} \mathbb{P}(X_i = 1|W)
\]

\[
= \frac{\text{max } p_i}{\mu} \mathbb{E}(W|W) = \left[ \frac{\text{max } p_i}{\mu} \right] W,
\]

so \( \delta_2 = \frac{\text{max } p_i}{\mu} \).

• The moderate deviation bound is

\[
\left| \frac{\mathbb{P}(W \geq k)}{\mathbb{P}_n(\mu)[[k, \infty))} - 1 \right| \leq C(\text{max } p_i)(1 + \xi^2).
\]

• What is \( C \)?
Matching: similar

- $n$: fixed;
- $\pi$: a uniform random permutation of $\{1, \ldots, n\}$;
- $X_i = 1_{\{i = \pi(i)\}}$;
- $W = \sum_{i=1}^{n} X_i$: # fixed points in the permutation.
- $\mathbb{P}(\{i = \pi(i)\}) = 1/n$ so $\mu = 1$.
- For $j \neq i$, $\mathbb{E}(X_j | X_i = 1) = 1/(n - 1)$, so $\sum_{j \neq i} \mathbb{E}(X_j | X_i = 1) = 1$, $\mathbb{E}(W^2) = 2/n$ and $\text{Var}(W) = 1$.
- For the size-biased distribution, let $W_i = W - X_i$, then

$$
\mathbb{E}[W f(W)] = \sum_{i=1}^{n} \mathbb{E}[X_i f(W)] = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[f(W_i+1) | i = \pi(i)].
$$
- $W^s$: equally likely, pick and fix an $i$, randomly permute the others.

- Chatterjee, Diaconis, and Meckes (2005): using $\pi$, we let $I$ be uniformly distributed on $\{1, \ldots, n\}$ and independent of $\pi$, when $I = i$:
  - if $i = \pi(i)$, do nothing and let $\pi^s = \pi$;
- if $i \neq \pi(i)$, move $i$ in $\pi$ back to $i$, $\pi^{-1}(i)$ to $\pi(i)$.
• We can write

\[
\pi^s(j) = \begin{cases} 
I & \text{for } j = I, \\
\pi(I) & \text{for } j = \pi^{-1}(I), \\
\pi(j) & \text{else.}
\end{cases}
\]

• Using this coupling, for all \(k\) with \(k^2/n \leq c\),

\[
\left| \frac{\mathbb{P}(W \geq k)}{\text{Pn}(1)([k, \infty))} - 1 \right| \leq C \frac{k^2}{n}. \quad (\leq c)
\]
We do both!

- $W$: a non-negative integer-valued random variable with mean $\mu$ and variance $\sigma^2$.

- We consider a $P_n$ approximation of $W - a$ for an $a < \mu$.

- Let $\lambda = \mu - a$, $Y \sim P_n(\lambda)$. Then for fixed integer $k$ with $x := \frac{k-\lambda}{\sqrt{\lambda}} \geq 1$, we have

$$ \left| \frac{\mathbb{P}(W - a \geq k)}{\mathbb{P}(Y \geq k)} - 1 \right| $$

$$ \leq 3\lambda^{-1}xe^{x^2+1} \left\{ \mu\mathbb{E}|W + 1 - W^s| + |\mu - \lambda| \right\} $$

$$ + \mathbb{P}(W - a < -1). $$
When $a = 0$

The bound is reduced to
\[
\left| \frac{\mathbb{P}(W \geq k)}{\mathbb{P}(Y \geq k)} - 1 \right| \leq 3xe^{x^2+1}\mathbb{E}|W + 1 - W^s|.
\]

vs Chen, Fang & Shao (2013):

\[
\left| \frac{\mathbb{P}(W \geq k)}{\mathbb{P}(Y \geq k)} - 1 \right| \leq C(\delta_1 + \delta_2 \lambda)(1 + x^2).
\]

- Our bound is easy to compute.
- It contains no unspecified constants.
Can we do better?

- \{X_i : i \in \mathcal{I}\}: a class of non-negative integer valued random variables.

- Chen and Shao (2004): the class satisfies

(LD2) For each \(i \in \mathcal{I}\), there exists an \(A_i \subset B_i \subset \mathcal{I}\) such that \(X_i\) is independent of
\[\{X_j : j \in A_i^c\}\] and \(\{X_i : i \in A_i\}\) is independent of \(\{X_j : j \in B_i^c\}\).

- We set \(W = \sum_{i \in \mathcal{I}} X_i\), \(\mu = \mathbb{E}W\), \(\sigma^2 = \text{Var}W\);
\(X_A = \{X_i : i \in A\}\).

- Let \(\theta_i := \text{ess sup} \max_j \mathbb{P}(W = j|X_{A_i}) \approx \max_j \mathbb{P}(W = j)\).
Let $\mu_i = \mathbb{E}X_i$, $\lambda = \mu - a$, $Y \sim \text{Pn}(\lambda)$. For fixed integer $k$ with $x := \frac{k-\lambda}{\sqrt{\lambda}} \geq 1$, we have

$$\left| \frac{\mathbb{P}(W - a \geq k)}{\mathbb{P}(Y \geq k)} - 1 \right|$$

$$\leq \lambda^{-1} \left[ 4xe^{x^2+1} + 1 \right] \sum_{i \in I} \theta_i \{ \mathbb{E}(X_i - \mu_i) \underbrace{Z_i}_{A_i} | \mathbb{E}(Z_i') \}$$

$$+ \mathbb{E} \left[ |X_i - \mu_i| Z_i (Z_i' - Z_i/2 - 1/2) \right]$$

$$+ 3|\lambda - \sigma^2| \lambda^{-1} xe^{x^2+1} + \mathbb{P}(W - a < -1).$$

[Slide 34]
Do we need MGF condition?

- In normal approximation: \( \mathbb{E} e^{t_0|X_1|} \leq c_0 < \infty \).

- In Pn: we don’t need it because the Pn tails are fatter.
  - In special cases, it is possible to prove that they meet the condition.
The Stein-Chen method: a key lemma

For $h = 1_{[k, \infty)}$, the solution of

$$\lambda f(j + 1) - j f(j) = h(j) - P_n(\lambda)\{h\}, \quad j \geq 0,$$

satisfies

(i) $\|f\| := \sup_{i \in \mathbb{Z}_+} |f(i)| \leq \lambda^{-1/2} e^{x^2} + 1 P_n(\lambda)\{h\}$;

(ii) $\Delta f(i)$ is negative and decreasing in $i \leq k - 1$; and positive and decreasing in $i \geq k$;

(iii) $\sup_{i \leq k-1} |\Delta f(i)| \leq \lambda^{-1} \left(1 + xe^{x^2} + 1\right) P_n(\lambda)\{h\}$ and

$$\sup_{i \geq k} |\Delta f(i)| \leq 3\lambda^{-1} xe^{x^2} + 1 P_n(\lambda)\{h\};$$

(iv) $\|\Delta f\| := \sup_{i \in \mathbb{Z}_+} |\Delta f(i)| \leq 3\lambda^{-1} xe^{x^2} + 1 P_n(\lambda)\{h\}$ and

$$\|\Delta^2 f\| := \sup_{i \in \mathbb{Z}_+} |\Delta^2 f(i)| \leq \lambda^{-1} \left[4xe^{x^2} + 1\right] P_n(\lambda)\{h\}.$$
Example

- \( \{X_i, \ 1 \leq i \leq n\} \): independent Bernoulli random variables with \( \mathbb{P}(X_i = 1) = p_i \in (0, 1) \), \( W = \sum_{i=1}^{n} X_i \).

- \( \lambda = \mathbb{E}W - a > 0 \), \( \sigma^2 = \text{Var}(W) \), \( Y \sim \text{Pn}(\lambda) \) and \( x := \frac{k-\lambda}{\sqrt{\lambda}} \geq 1 \),

\[
\left| \frac{\mathbb{P}(W - a \geq k)}{\mathbb{P}(Y \geq k)} - 1 \right|
\leq \left[ 4xe^{x^2+1} + 1 \right] \text{ something like } \max_i p_i / \sigma
\]

\[
+ 3|\lambda - \sigma^2| x \lambda^{-1} e^{x^2+1} + \exp \left\{ -\frac{(\mu - a + 2)^2}{2\mu} \right\}.
\]
Matching problem

For a fixed \( n \), let \( \pi \) be a uniform random permutation of \( \{1, \ldots, n\} \), \( W = \sum_{i=1}^{n} 1_{\{i=\pi(i)\}} \) be the number of fixed points in the permutation, then

\[
\left| \frac{\mathbb{P}(W \geq k)}{\text{Pn}(1)([k, \infty))} - 1 \right| \leq \frac{6}{n} xe^{x^2+1},
\]

where \( x := k - 1 \geq 1 \).
2-runs

- \{\xi_1, \ldots, \xi_n\}: i.i.d. \textit{Bernoulli}(p) random variables with \(n \geq 9\), \(p < 2/3\), \(\xi_{j+n} = \xi_j\) for \(-3 \leq j \leq n\).

- \(X_i = \xi_i \xi_{i+1}\), \(W = \sum_{i=1}^{n} X_i\).

- \(\mu = np^2\) and \(\sigma^2 = np^2(1 - p)(3p + 1)\).

- \(a := \lfloor np^3(3p - 2)\rfloor\), \(\lambda = \mu - a\), \(Y \sim \text{Pn}(\lambda)\), then

\[
\left| \frac{\mathbb{P}(W_n - a \geq k)}{\mathbb{P}(Y \geq k)} - 1 \right| \\
\leq \frac{9.2(4xe^{x^2+1} + 1)(1 + 5p)}{(1 + 2p - 3p^2)\sqrt{(n - 8)(1 - p)^3}} \\
+ \frac{3xe^{x^2+1}}{1 \vee [np^2(1 + 2p - 3p^2)]}. \tag{1}
\]
Take home messages

• For the counts of rare events, the tail probabilities can be approximated by the moderate deviations of $P_n$ with twists of the parameters.

• The robustness of the tail behaviour of the $P_n$ for large $\lambda$ has not been incorporated into the bound.

• We conjecture that bound can be sharpened by a factor possibly as much as $1/3$.

• We don’t have any idea about the lower bound.
Thank you!