Impact of dependency on the distribution of p-value

Marie Ernst* and Yvik Swan

University of Liege, Belgium

Baltimore, August 1st, 2017

Multiple testing

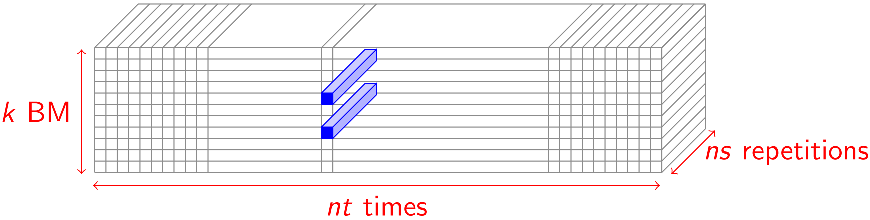
- Multiple null hypotheses $(H_{0,i})_{i=1,\ldots,m}$
- Each test is based on a dataset $\mathbf{x}_i = (x_{i1}, \ldots, x_{in_s})$

If the datasets are not independent, how can we detect deviations from null hypotheses?

Multiple testing: example

Example with dependence: portfolio of assets

- k Brownian motions measured at n_t times
- Test if the correlation between each pair is as expected considering n_s independent repetitions.



Outline

1. Impact of dependency on the distribution of p-value
2. Distribution of multivariate p-values
3. Distribution of sums of indicators

Multiple testing: control procedures

Family-wise error rate (FWER)

Probability to reject at least one null hypothesis.
Examples: Bonferroni, Sidák, Hochberg

False discovery rate (FDR)

Rate of falsely rejected null hypotheses:

$$E\left[\frac{\#\text{falsely rejected}}{\#\text{rejects}}\right]$$

Examples: Benjamini-Hochberg (1995), Benjamini-Yekutieli (2001), Cai-Liu (2016)

Multiple testing: control procedures

Family-wise error rate (FWER)

Probability to reject at least one null hypothesis.
Examples: Bonferroni, Sidák, Hochberg

False discovery rate (FDR)

Rate of falsely rejected null hypotheses:

$$E\left[\frac{\#\text{falsely rejected}}{\#\text{rejects}}\right]$$

Examples: Benjamini-Hochberg (1995), Benjamini-Yekutieli (2001), Cai-Liu (2016)

$\rightsquigarrow$ Instead of a correction, could we consider a better distribution for multivariate p-values?

Distributions of p-values

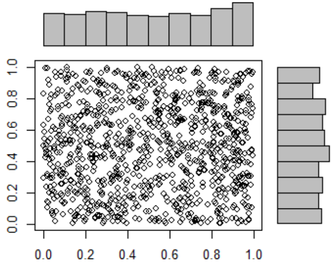
Under $H_{0,i}$, p-value $p_i \sim U[0,1]$ for $i = 1, \ldots, m$

- With independence, $\mathbf{p} \sim U[0,1]^m$
- Without independence, only the margins are uniformly distributed.

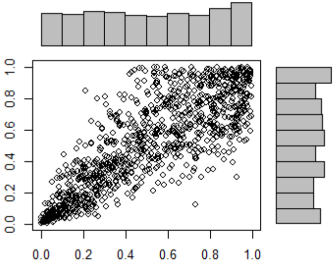
Distributions of p-values

Under $H_{0,i}$, p-value $p_i \sim U[0,1]$ for $i = 1, \ldots, m$

- With independence, $\mathbf{p} \sim U[0,1]^m$
- Without independence, only the margins are uniformly distributed.



With independence



Without independence

Distributions under H_0

Distributions of p-values

Difficult problem (see Wang 2014)

$\rightsquigarrow$ "Easier" distributions can be considered

Distributions under H_0

Distributions of p-values

Difficult problem (see Wang 2014)
$\rightsquigarrow$ "Easier" distributions can be considered

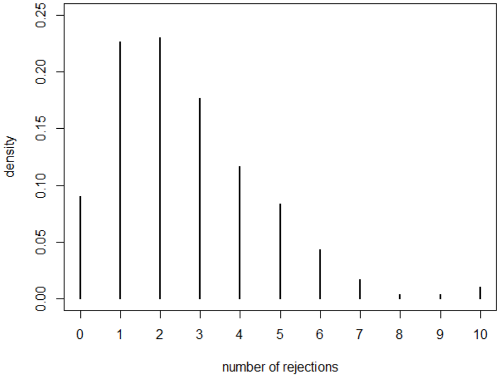
Distributions of the number W of rejections under H_0

For a level α for each test,

- With independence, $W \sim Bin(m, \alpha)$.
- Without independence, Binomial distribution does not fit anymore.

Example: portfolio of assets

Testing correlation of 11 Brownian motions measured at 300 times

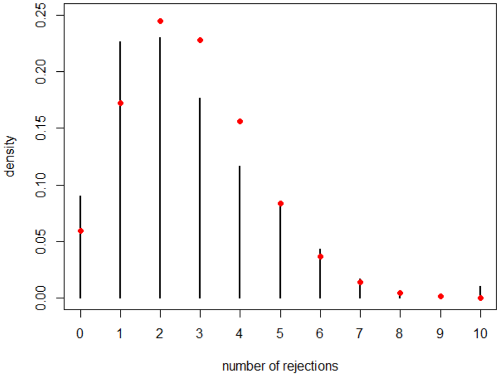


Which distribution fits the data ?

Distribution under H_0

Example: portfolio of assets

Testing correlation of 11 Brownian motions measured at 300 times



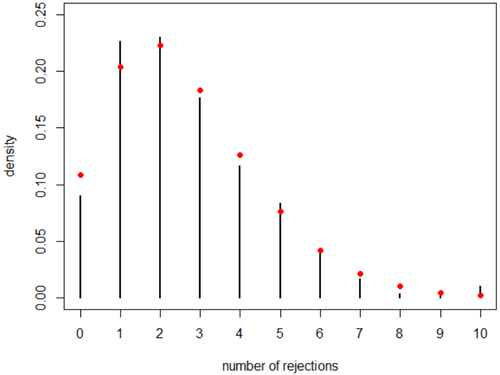
Which distribution fits the data ?

- Binomial ?

Distribution under H_0

Example: portfolio of assets

Testing correlation of 11 Brownian motions measured at 300 times



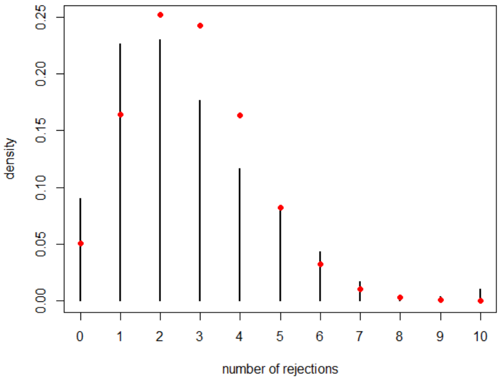
Which distribution fits the data ?

- Binomial ?
- Beta binomial?

Distribution under H_0

Example: portfolio of assets

Testing correlation of 11 Brownian motions measured at 300 times



Which distribution fits the data ?

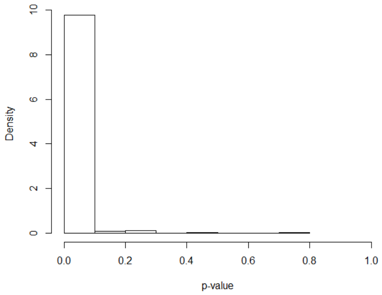
- Binomial ?
- Beta binomial?
- Hypergeometric?

Distribution under H_0

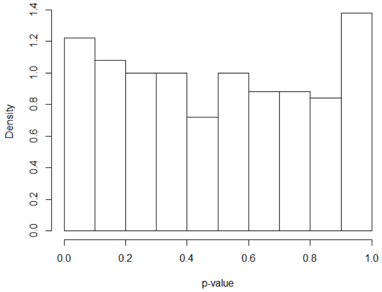
Example

Pearson's Chi-squared test to determine the distribution of the number of rejections.

$$W \sim \text{Bin}(55, 0.05)?$$



$$W \sim \text{Beta Bin}(55, \alpha, \beta)?$$



Sum of correlated indicators

General question

How "close" are the distributions of sums of indicators?
Can we identify when they are alike or different?

Some examples of law

We consider random variables which admit representations of the form

$$\sum_{i=1}^{n} \mathbb{I}_i$$

for $(\mathbb{I}_1, \ldots, \mathbb{I}_n) \in \{0,1\}^n$ with $\mathbb{I}_i \sim \mathrm{Bern}(p_i)$

Some laws $X = \sum_{i=1}^{n} \mathbb{I}_i$

- Binomial $\text{Bin}(n, p)$: $\mathbb{I}_i \overset{i.i.d.}{\sim} \text{Bern}(p)$
- Poisson binomial S: $\mathbb{I}_i \sim \text{Bern}(p_i)$ and $\mathbb{I}_i \perp\!\!\!\perp$
- Beta Binomial $\mathcal{BB}(n, \alpha, \beta)$: $\mathbb{I}_i \sim \text{Bern}(\frac{\alpha}{\alpha+\beta})$ and $cor(\mathbb{I}_i, \mathbb{I}_j) = \frac{1}{\alpha+\beta+1} > 0$
- Hypergeometric $\mathcal{H}(M, N, n)$: $\mathbb{I}_i \sim \text{Bern}(\frac{M}{M+N})$ and $cor(\mathbb{I}_i, \mathbb{I}_j) = \frac{-1}{M+N-1} < 0$
- A general case
  $\text{Bin}(n_1, p) \oplus \sum_{i=1}^{n_2} X_i$ with some $X_i \sim \text{Bern}(p_i)$
- k-runs
  $\mathbb{I}_i = \prod_{j=1}^{k} X_{i+j}$ with $X_i \overset{iid}{\sim} Bern(p)$
- m-dependent $N(n; k_1, k_2)$
  $\mathbb{I}_i = \prod_{j=1}^{k_1} X_{i+j} \prod_{j=1}^{k_2}(1 - X_{i+k_1+j})$ with $X_i \overset{iid}{\sim} Bern(p)$

Some laws $X = \sum_{i=1}^{n} \mathbb{I}_i$

- Binomial $\mathrm{Bin}(n, p)$: $\mathbb{I}_i \overset{i.i.d.}{\sim} \mathrm{Bern}(p)$ explicit law
- Poisson binomial S: $\mathbb{I}_i \sim \mathrm{Bern}(p_i)$ and $\mathbb{I}_i \perp\!\!\!\perp$
- Beta Binomial $\mathcal{BB}(n, \alpha, \beta)$: $\mathbb{I}_i \sim \mathrm{Bern}(\frac{\alpha}{\alpha+\beta})$ and $cor(\mathbb{I}_i, \mathbb{I}_j) = \frac{1}{\alpha+\beta+1} > 0$ explicit law
- Hypergeometric $\mathcal{H}(M, N, n)$: $\mathbb{I}_i \sim \mathrm{Bern}(\frac{M}{M+N})$ and $cor(\mathbb{I}_i, \mathbb{I}_j) = \frac{-1}{M+N-1} < 0$ explicit law
- A general case
  $\mathrm{Bin}(n_1, p) \oplus \sum_{i=1}^{n_2} X_i$ with some $X_i \sim \mathrm{Bern}(p_i)$
- k-runs
  $\mathbb{I}_i = \prod_{j=1}^{k} X_{i+j}$ with $X_i \overset{iid}{\sim} Bern(p)$
- m-dependent $N(n; k_1, k_2)$
  $\mathbb{I}_i = \prod_{j=1}^{k_1} X_{i+j} \prod_{j=1}^{k_2} (1 - X_{i+k_1+j})$ with $X_i \overset{iid}{\sim} Bern(p)$

Distance between distributions

Let W and Z be two random variables

Total variation distance

$$\mathrm{TV}(W,Z) = \sup_A |P(W \in A) - P(Z \in A)|$$

for measurable sets A.

$\rightsquigarrow$ Stein's method allows us to get bounds

Some bounds for TV Distances

$\sum_{\mathbf{i=1}}^{\mathbf{n}} \mathbb{I}_{\mathbf{i}}$	Bin(n,p)	$Pois(\lambda)$	$NB(r,q)$ (Neg. bin.)	$TP(\mu,\sigma^2)$ (Translated Pois.)	$N^d(\mu,\sigma^2)$ (Discretized)
Bin(n,p)	0	$\leq (1-e^{np})p$ (Barbour et al., 1992)			
$\mathcal{BB}(n,\alpha,\beta)$	$\leq \frac{n(n-1)}{(n+1)(1+\alpha+\beta)}$ (Teerapabolarn, 2008)				
$\mathcal{H}(M,N,n)$	$TV \leq \frac{n-1}{N-1}$ (Holmes, 2004)	$\frac{\varepsilon}{11+3\max(0,\frac{\gamma}{\lambda\varepsilon})} \leq TV$ $\leq \frac{1-e^{-\lambda}}{N-1}(n+R-nR-1)$ (Barbour et al., 1992)			
$S(n,\mathbf{p})$	$\sim \sum_{i=1}^{n}(p_i-p)^2$ (Ehm, 1991)	$\sim \lambda^{-1}\sum p_i^2$ (Barbour-Hall, 1984)			
Bin$(n-k,p)$ $\oplus k$Bern(p)					
2-runs			$\leq \frac{32.2p}{\sqrt{(n-1)(1-p)^3}}$ (Brown-Xia, 2001)		$\leq \frac{C_p}{\sqrt{n}}$ (Fang, 2014)
k-runs	$TV \leq \mathcal{O}\left(\frac{k^2}{p(1-p)}\left(\frac{k}{k-1}\right)^{k-1}\right)$ (Kumar-Upadhye, 2016)		$\leq \mathcal{O}\left(\frac{(4k-3)(2k-1)p^2}{\sqrt{(n-4k+2)p^k(1-p)^3}}\right)$ (Wang-Xia, 2008)	$\leq \frac{K(k,p)}{\sqrt{n}}$ (Röllin, 2005)	
m-dependent $N(n,k_1,k_2)$	$\leq \frac{(1-p)^{1.5k_1}p^{1.5k_2}m^2}{\sqrt{n-m+1}}$ (Cekanavicius-Vell., 2015) (Zhang, 2016)	$\leq \mathcal{O}\left(\frac{(2+qp)(n-k)}{q(1+(n-k-1)p)}\right)$ (Kumar-Upadhye, 2017)			

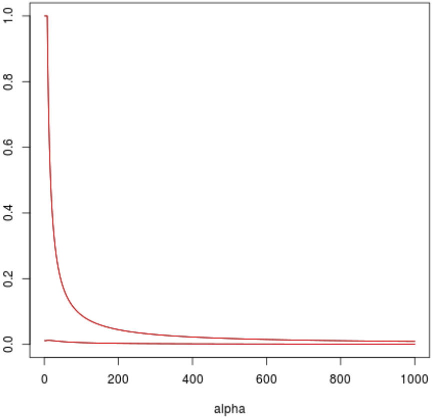
Some bounds for TV Distances

$\sum_{\mathbf{i=1}}^{\mathbf{n}} \mathbb{I}_{\mathbf{i}}$	$\mathrm{Bin}(n,p)$	$Pois(\lambda)$	$NB(r,q)$ (Neg. bin.)	$TP(\mu,\sigma^2)$ (Translated Pois.)	$N^d(\mu,\sigma^2)$ (Discretized)
$\mathrm{Bin}(n,p)$	0	$\leq (1-e^{np})p$ (Barbour et al., 1992)			
$\mathcal{BB}(n,\alpha,\beta)$	$\mathcal{O}\left(\frac{n^2(\alpha+\beta)^2}{(\alpha+\beta)^3+n^2\alpha\beta}\right) \leq TV$ $\leq \dfrac{n(n-1)}{(n+1)(1+\alpha+\beta)}$ (Teerapabolarn, 2008)				
$\mathcal{H}(M,N,n)$	$\mathcal{O}\left(\left(\frac{n}{n-R}\right)^{n+\frac{1}{2}}\left(\frac{n-R}{N-R}\right)^{R}\left(\frac{N-R}{N}\right)^{n}\right) \leq$ $TV \leq \dfrac{n-1}{N-1}$ (Holmes, 2004)	$\frac{\varepsilon}{11+3\max(0,\frac{\gamma}{\lambda\varepsilon})} \leq TV$ $\leq \frac{1-e^{-\lambda}}{N-1}(n+R-nR-1)$ (Barbour et al., 1992)			
$S(n,\mathbf{p})$	$\sim \sum_{i=1}^{n}(p_i-p)^2$ (Ehm, 1991)	$\sim \lambda^{-1}\sum p_i^2$ (Barbour-Hall, 1984)			
$\mathrm{Bin}(n-k,p)$ $\oplus k\mathrm{Bern}(p)$	$\mathcal{O}\left(\frac{k^2pq}{npq+kpq+k^2}\right) \leq TV$ $\leq \mathcal{O}\left(\frac{k(k-1)}{n+1}\right)$				
2-runs			$\leq \frac{32.2p}{\sqrt{(n-1)(1-p)^3}}$ (Brown-Xia, 2001)		$\leq \frac{C_p}{\sqrt{n}}$ (Fang, 2014)
k-runs	$TV \leq \mathcal{O}\left(\frac{k^2}{p(1-p)}\left(\frac{k}{k-1}\right)^{k-1}\right)$ (Kumar-Upadhye, 2016)		$\leq \mathcal{O}\left(\frac{(4k-3)(2k-1)p^2}{\sqrt{(n-4k+2)p^k(1-p)^3}}\right)$ (Wang-Xia, 2008)	$\leq \frac{K(k,p)}{\sqrt{n}}$ (Röllin, 2005)	
m-dependent $N(n,k_1,k_2)$	$\leq \frac{(1-p)^{1.5k_1}p^{1.5k_2}m^2}{\sqrt{n-m+1}}$ (Cekanavicius-Vell., 2015) (Zhang, 2016)	$\leq \mathcal{O}\left(\frac{(2+qp)(n-k)}{q(1+(n-k-1)p)}\right)$ (Kumar-Upadhye, 2017)			

Illustration

$$TV(\text{Bin}(n,p), \mathcal{BB}(n,\alpha,\beta))$$

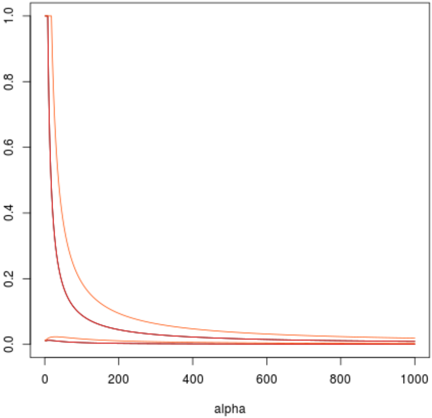
where $p = \frac{\alpha}{\alpha+\beta}$, $\beta = 1$ and $n = 10$



Illustration

$$TV(\text{Bin}(n,p), \mathcal{BB}(n,\alpha,\beta))$$

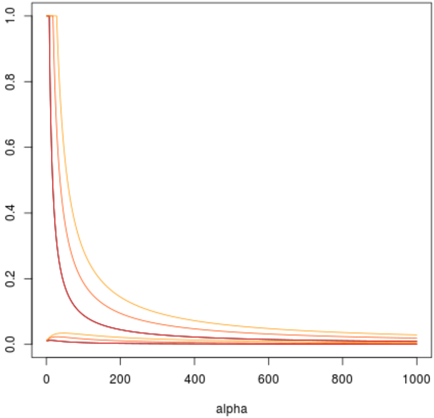
where $p = \frac{\alpha}{\alpha+\beta}$, $\beta = 1$ and $n = 10, 20$



Illustration

$$TV(\text{Bin}(n,p), \mathcal{BB}(n,\alpha,\beta))$$

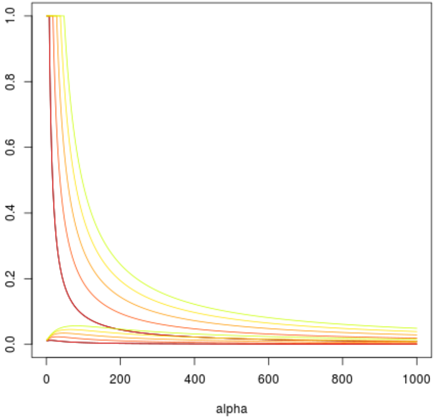
where $p = \frac{\alpha}{\alpha+\beta}$, $\beta = 1$ and $n = 10, 20, 30$



Illustration

$$TV(\text{Bin}(n,p), \mathcal{BB}(n,\alpha,\beta))$$

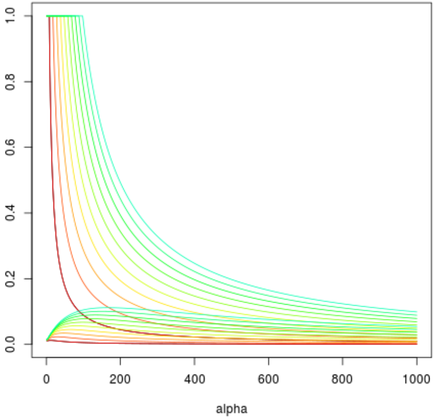
where $p = \frac{\alpha}{\alpha+\beta}$, $\beta = 1$ and $n = 10, 20, \ldots, 50$.



Illustration

$$TV(\text{Bin}(n,p), \mathcal{BB}(n,\alpha,\beta))$$

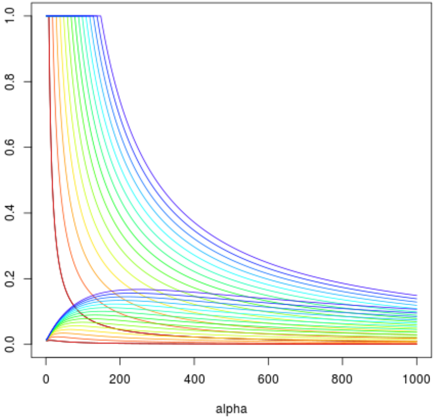
where $p = \frac{\alpha}{\alpha+\beta}$, $\beta = 1$ and $n = 10, 20, \ldots, 100$.



Illustration

$$TV(\mathrm{Bin}(n,p), \mathcal{BB}(n,\alpha,\beta))$$

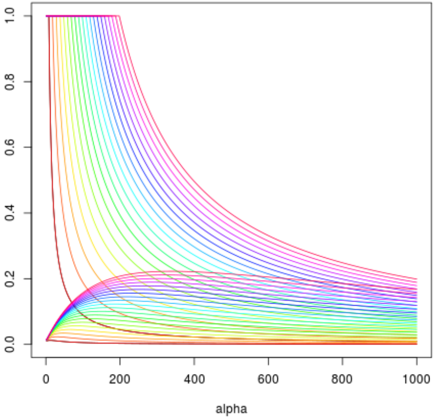
where $p = \frac{\alpha}{\alpha+\beta}$, $\beta = 1$ and $n = 10, 20, \ldots, 150$.



Illustration

$$TV(\text{Bin}(n,p), \mathcal{BB}(n,\alpha,\beta))$$

where $p = \frac{\alpha}{\alpha+\beta}$, $\beta = 1$ and $n = 10, 20, \ldots, 200$.



Example: family-wise error rate

Testing correlation for 20 Brownian motions measured at 300 times

	FWER method	Proportion of rejections
FWER method on p-values	Bonferroni	0.324%
	Sidák	0.332%
	Hochberg	0.327%
FWER method on indicators	Binomial quantile	16%
	Beta Binomial quantile	6%

Example: family-wise error rate

Testing correlation for 20 Brownian motions measured at 300 times

	FWER method	Proportion of rejections	
FWER method on p-values	Bonferroni	0.324%	conservative tests
	Sidák	0.332%	
	Hochberg	0.327%	
FWER method on indicators	Binomial quantile	16%	anticonservative tests
	Beta Binomial quantile	6%	

Related problems

Power in multiple testing

Number of rejections $W \sim \text{Bin}(m_0, p) \oplus \sum_{i=m_0+1}^{m} \text{Bern}(\beta_i)$

Exact distances

The exact distances between distributions is important (Adell 2005, 2008)

Distribution of multivariate p-values

A better understanding of the joint distribution can lead to refined confidence/rejection regions (already studied in Chi 2008). This is a difficult problem (see Wang 2014).

References

Stein's method

- Barbour & Hall (1984). On the rate of Poisson convergence. Cambridge University Press.
- Barbour, Holst & Janson (1992). Poisson approximation. Clarendon Press Oxford.
- Brown & Xia (2001). Stein's method and birth-death processes. *Annals of probability*.
- Čekanavičius & Vellaisamy (2013). Discrete approximations for sums of m-dependent random variables. arXiv.
- Ehm (1991). Binomial approximation to the Poisson binomial distribution. *Statistics & Probability Letters*.
- Fang (2014). Discretized normal approximation by Stein's method. *Bernoulli*.
- Holmes (2004). Stein's method for birth and death chains. *Stein's Method*, Institute of Mathematical Statistics.
- Kumar & Upadhye (2017). On discrete Gibbs measure approximation to runs. arXiv.
- Kumar & Upadhye (2016). Pseudo-binomial Approximation to (k_1, k_2)-runs. arXiv.
- Röllin (2005). Approximation of sums of conditionally independent variables by the translated Poisson distribution. *Bernoulli*.
- Soon (1996). Binomial approximation for dependent indicators. *Statistica Sinica*.
- Teerapabolarn (2008). A bound on the binomial approximation to the beta binomial distribution. *International Mathematical Forum*.
- Wang & Xia (2008). On negative binomial approximation to k-runs. *Journal of Applied Probability*.
- Zhang (2016). Binomial approximation for sum of indicators with dependent neighborhoods. *Statistics & Probability Letters*.

Thanks to the IAP StUDyS for financial support