

NORMAL APPROXIMATION TO THE BINOMIAL

Talk by David Pollard at the Orsay meeting in honor of

JEAN BRETAGNOLLE
DIDIER DACUNHA-CASTELLE
and
ILDAR IBRAGIMOV

8 June 2001

The quantile coupling between the $\text{Bin}(n, \frac{1}{2})$ and the $N(n/2, n/4)$ distributions has played a surprisingly important role in probability and statistics. For example, it is the basic ingredient in the KMT coupling of the empirical process with a Brownian Bridge. More recently, it has emerged as a tool for bounding the Le Cam distance between density and white noise models.

The talk will sketch some applications and explain a simple method for deriving bounds for the quantile coupling.

<http://www.stat.yale.edu/~pollard/Paris2001>

$$\text{Bin}(n, \frac{1}{2}) \approx N\left(\frac{n}{2}, \frac{n}{4}\right)$$

De Moivre (1733)

⋮

⋮

Peizer & Pratt (1968); Molenaar (1970)

⋮

Komlós, Major & Tusnády (1975)

⋮

Tusnády (1977)

⋮

Csörgő & Révész (1981, Section 4.4)

⋮

Bretagnolle & Massart (1989)

⋮

Carter & Pollard (2000)

TUSNÁDY BOUND

There exist $X \sim \text{Bin}(n, \frac{1}{2})$ and $Y \sim N(n/2, n/4)$ with

$$|X - n/2| \leq |Y - n/2| + 1$$

and

$$|X - Y| \leq 1 + \frac{(Y - n/2)^2}{2n}$$

Note:

$$X - \frac{n}{2} \text{ is of order } \sqrt{n}$$

$$Y - \frac{n}{2} \text{ is of order } \sqrt{n}$$

$$\text{Tusnady: } X - Y \text{ is of order } O_p(1)$$

COUPLING EMPIRICAL PROCESS WITH BROWNIAN BRIDGE

P = Lebesgue measure on $[0, 1]$.

Haar basis for $L^2(P)$:

$$\Psi = \{1\} \cup \{\psi_{i,k} : 0 \leq i < 2^k, k \in \mathbb{N}_0\}$$

$$f - Pf = \sum_{k=0}^{\infty} \sum_i \psi_{i,k} \langle f, \psi_{i,k} \rangle,$$

Brownian Bridge:

$$\nu(f) = \sum_{k=0}^{\infty} \sum_i \eta_{i,k} \langle f, \psi_{i,k} \rangle$$

with $\eta_{i,k} = \nu(\psi_{i,k})$ independent $N(0, 1)$.

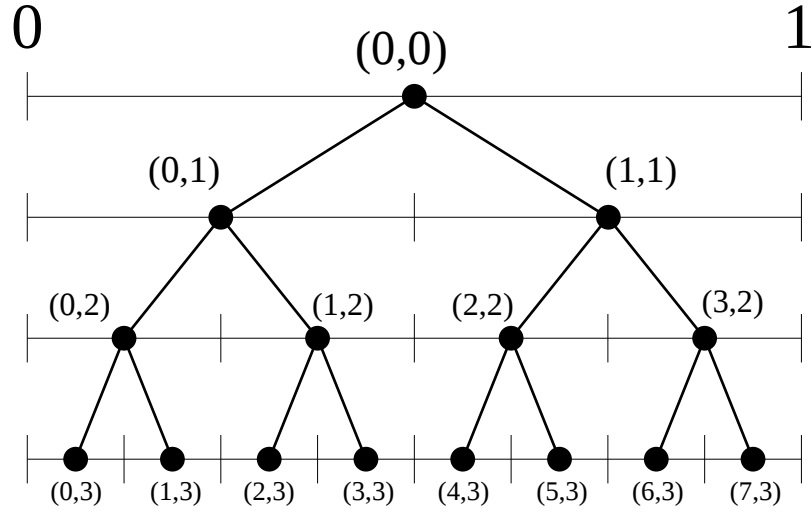
Empirical process $\nu_n := \sqrt{n}(P_n - P)$:

$$\nu_n(f) \stackrel{?}{=} \sum_{k=0}^{\infty} \sum_i \nu_n(\psi_{i,k}) \langle f, \psi_{i,k} \rangle.$$

Use Tusnády to couple $\nu_n(\psi_{i,k})$ with $\eta_{i,k}$.

KMT: $f \in \{[0, t] : 0 \leq t \leq 1\}$

cf. UGMTP §10.6



Contribution from coupling down a path

$(i_0, 0), (i_1, 1), (i_2, 2) \dots, (i_m, m) :$

$$\sum_{k=0}^m 2^{-k/2} |\nu_n(\psi_{i_k, k}) - \eta_{i_k, k}| \leq \frac{C}{\sqrt{n}} \sum_{k=0}^m (1 + \eta_{i_k, k}^2)$$

QUANTILE TRANSFORMATION

Find

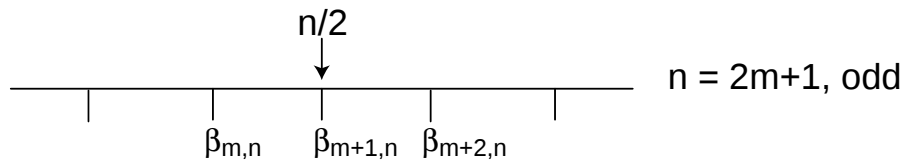
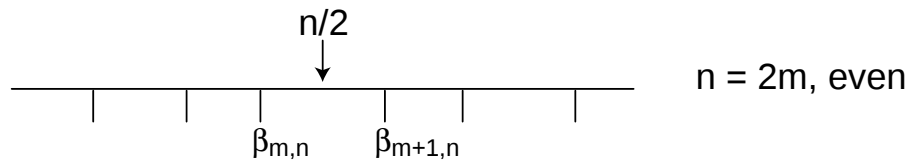
$$-\infty = \beta_{0,n} < \beta_{1,n} < \dots < \beta_{n,n} < \beta_{n+1,n} = \infty$$

with

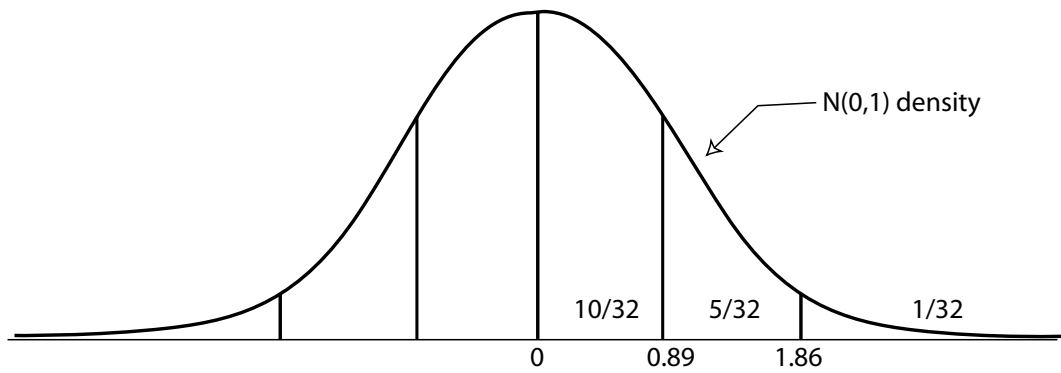
$$\mathbb{P} \left\{ \text{Bin}(n, \frac{1}{2}) \geq k \right\} = \mathbb{P} \left\{ N \left(\frac{n}{2}, \frac{n}{4} \right) > \beta_{k,n} \right\}.$$

Put

$$X = k \quad \text{when } \beta_{k,n} \leq Y < \beta_{k+1,n}$$



Bin(5, 1/2) EXAMPLE



N(5/2, 5/4) cutpoints:

0.42

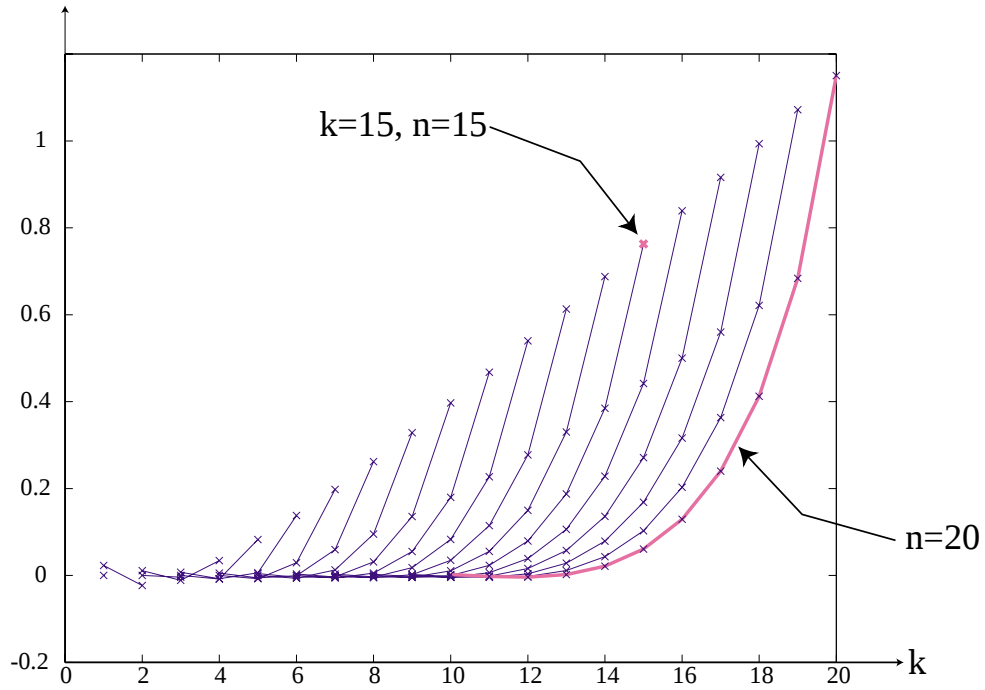
1.51

2.50

3.49

4.58

$\beta_{k,n} - (k - 1/2)$ versus k
for
 $n/2 \leq k \leq n$ and $1 \leq n \leq 20$



$$\mathbb{P}\{X \geq k\} \approx \mathbb{P}\{Y \geq k - \frac{1}{2}\} \quad \text{and} \quad \beta_{k,n} \approx k - \frac{1}{2} \quad ?$$

APPROXIMATIONS FOR $k > n/2$

Tusnády, via Bretagnolle & Massart (1989, Appendix):

$$\begin{aligned} k - 1 &\leq \beta_{k,n} \leq 3n/2 - n\sqrt{1 - \epsilon_k} \quad \text{with } \epsilon_k := (2k - n)/n \\ &= k + \frac{(k - n/2)^2}{2n} + \dots \quad \text{for } k \ll n \end{aligned}$$

Carter & Pollard (2000):

$$k - \frac{1}{2} + f_{k,n} - \frac{C}{\sqrt{n}} \leq \beta_{k,n} \leq k - \frac{1}{2} + f_{k,n} + \frac{C \log n}{\sqrt{n}}$$

with

$$f_{k,n} = \frac{(k - n/2)^3}{3n^2} + \dots \quad \text{for } k \ll n$$

Compare when $|k - n/2| = o(n^{2/3})$ or $k \approx n$.

EXTREME TAIL, FOR FIXED B

$$\beta_{n-B,n} = \frac{1+c}{2}n - \frac{1+2B}{4c} \log n + O(1)$$

where

$$c = \sqrt{2 \log 2} \approx 1.177 \quad \text{and} \quad \frac{1+c}{2} \approx 1.088$$

Tusnády:

$$\beta_{n-B,n} \leq \frac{3n}{2} - \sqrt{2nB}$$

THEOREM

Standardized cutpoints:

$$\begin{aligned} z_{k,n} &:= 2(\beta_{k,n} - n/2)/\sqrt{n} \\ u_{k,n} &:= 2(k - 1/2 - n/2)/\sqrt{N} \quad \text{where } N := n - 1. \end{aligned}$$

Uniformly in $(n + 1)/2 \leq k < n$,

$$z_{k,n} = u_{k,n} S(u_{k,n}/\sqrt{N}) + \frac{\log(1 - u_{k,n}^2/N)}{2c|u_{k,n}|} + R_{k,n}$$

with

$$-O\left(\frac{|u_{k,n}| + 1}{n}\right) \leq R_{k,n} \leq O\left(\frac{|u_{k,n}| + \log n}{n}\right)$$

where $S(\epsilon) := \sqrt{1 + 2\epsilon^2\gamma(\epsilon)}$, for the increasing function $\gamma(\cdot)$ defined for $0 < \epsilon < 1$ by

$$\begin{aligned} \gamma(\epsilon) &:= \frac{(1 + \epsilon) \log(1 + \epsilon) + (1 - \epsilon) \log(1 - \epsilon) - \epsilon^2}{2\epsilon^4} \\ &= \sum_{r=0}^{\infty} \frac{\epsilon^{2r}}{(2r + 3)(2r + 4)}. \end{aligned}$$

Note: $\gamma(0) = 1/12$ and $1 + 2\gamma(1) = c^2 := 2 \log 2$.

METHOD OF PROOF

$$\mathbb{P}\{X \geq k\} = \frac{n!}{(k-1)!(n-k)!} \int_0^{1/2} t^{k-1} (1-t)^{n-k} dt.$$

Put $N := n - 1$ and $K := k - 1$ and

$$\alpha := \frac{1 + \epsilon}{2} := K/N := 1 - \bar{\alpha}.$$

$$H(t) := \alpha \log t + \bar{\alpha} \log(1-t)$$

Stirling:

$$\frac{n!}{(k-1)!(n-k)!} \approx \sqrt{\frac{N}{2\pi\alpha\bar{\alpha}}} \exp(-NH(\alpha))$$

$$\begin{aligned} \mathbb{P}\{X \geq k\} &\approx \sqrt{\frac{N}{2\pi\alpha\bar{\alpha}}} \int_0^{1/2} \exp(NH(t) - NH(\alpha)) dt \\ &\approx \sqrt{\frac{N}{2\pi\alpha\bar{\alpha}}} \int_0^{1/2} \exp\left(-N\epsilon^4\gamma(\epsilon) - 2N\left(s + \frac{\epsilon}{2}\right)^2\right) ds. \end{aligned}$$

INTERPRETATION OF NUSSBAUM (1996)

$$Q_\delta := N\left(\frac{n}{2} + n\delta, \frac{n}{4}\right) \quad \text{and} \quad P_\delta := \text{Bin}(n, \frac{1}{2} + \delta)$$

Under quantile coupling, $X \sim P_0$ and $Y \sim Q_0$, and

$$Y \mid X = x \sim K_x := Q_0(\cdot \mid [\beta_{x,n}, \beta_{x+1,n}))$$

Define $\tilde{Q}_\delta$ as the distribution of $\tilde{Y}$, where

$$X \sim P_\delta \quad \text{and} \quad \tilde{Y} \mid X = x \sim K_x$$

Hellinger distance:

$$H^2(\tilde{Q}_\delta, Q_\delta) = O(\delta^2) \quad \text{for } \delta = O(n^{-1/2}).$$

Carter (2000, 2001)

REFERENCES

- Bretagnolle, J. & Massart, P. (1989), ‘Hungarian constructions from the nonasymptotic viewpoint’, *Annals of Probability* **17**, 239–256.
- Carter, A. (2000), Asymptotic equivalence of nonparametric experiments, PhD thesis, Yale.
- Carter, A. (2001), Le Cam distance between multinomial and multivariate normal experiments under smoothness constraints on the parameter set, Technical report, University of California, Santa Barbara. <http://www.pstat.ucsb.edu/~carter>.
- Carter, A. & Pollard, D. (2000), Tusnády’s inequality revisited, Technical report, Yale University. <http://www.stat.yale.edu/~pollard>.
- Csörgő, M. & Révész, P. (1981), *Strong Approximations in Probability and Statistics*, Academic Press, New York.
- Komlós, J., Major, P. & Tusnády, G. (1975), ‘An approximation of partial sums of independent rv-s, and the sample df. I’, *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* **32**, 111–131.
- Molenaar, W. (1970), *Approximations to the Poisson, Binomial, and Hypergeometric distribution functions*, Center for Mathematics and Computer Science. CWI Tract 31.
- Nussbaum, M. (1996), ‘Asymptotic equivalence of density estimation and gaussian white noise’, *Annals of Statistics* **24**, 2399–2430.
- Peizer, D. B. & Pratt, J. W. (1968), ‘A normal approximation for Binomial, F, and other common, related tail probabilities I’, *Journal of the American Statistical Association* **63**, 1416–1456.
- Pollard, D. (2001), *A User’s Guide to Measure Theoretic Probability*, Cambridge University Press.
- Tusnády, G. (1977), A study of Statistical Hypotheses, PhD thesis, Hungarian Academy of Sciences, Budapest. In Hungarian.