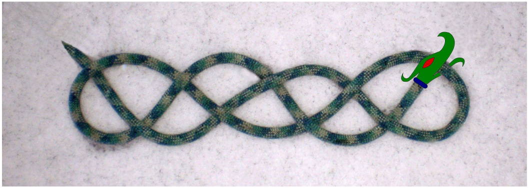
Inequalities for the Binomial Distributions

Peter Harremoës

Copenhagen Business College
Denmark



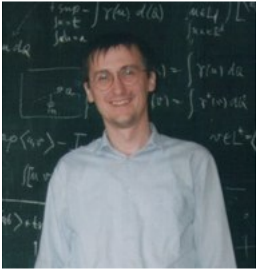
Workshop on Probability and Information Theory, Hong Kong
2019

Thanks to my coauthors

Lásló Györfi

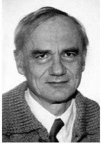
Oliver Johnson

Ioannis Kontoyiannis



František Matúš

Pavel Ruzankin



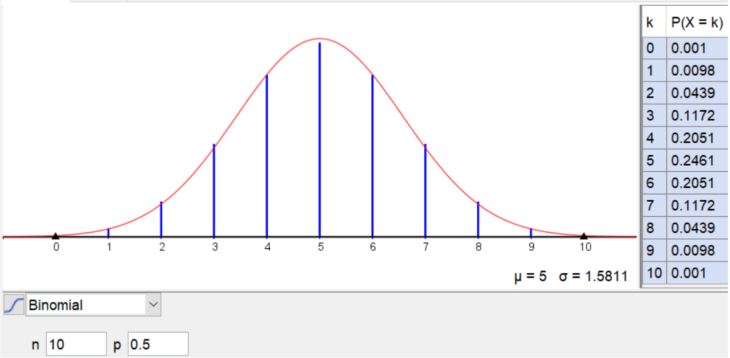
Gábor Tusnády

Christophe Vignat

What is the problem?

The random variable X is binomial if

$$\Pr(X = j) = \binom{n}{j} p^j (1-p)^{n-j}.$$

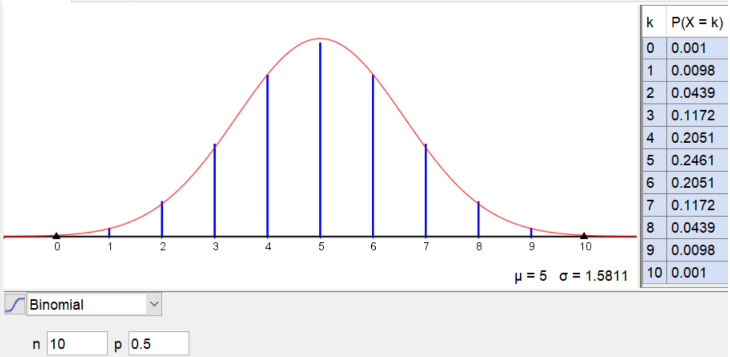


k	P(X = k)
0	0.001
1	0.0098
2	0.0439
3	0.1172
4	0.2051
5	0.2461
6	0.2051
7	0.1172
8	0.0439
9	0.0098
10	0.001

What is the problem?

The random variable X is binomial if

$$\Pr(X = j) = \binom{n}{j} p^j (1-p)^{n-j}.$$



- Often n or p are not known.

The binomial distribution and its cousins

- Hypergeometric distribution.
- Bernoulli sum.
- Poisson distribution.
- Negative binomial distribtuion.
- Gaussian distribution.
- Multinomial distribution.

Maximum entropy

Let $B_n(\lambda)$ denote the set of distributions of sums $S_n = X_1 + X_2 + \cdots + X_n$ with mean λ where X_i is a Bernoulli random variable with $\Pr(X_i = 1) = p_i$.

Lemma (Shepp and Olkin 1978, E. Hillion and O. Johnson 2015)

The map $(p_1, p_2, \ldots, p_n) \to H(S_n)$ is concave.

Theorem (PH 2001)

The $H(P)$ entropy restricted to $P \in B_n(\lambda)$ has maximum when $p_i = \lambda/n$, i.e. when P is $bin(n, \lambda/n)$.
Let $B_\infty(\lambda) = cl(\bigcup B_n(\lambda))$.

Corollary (PH 2001)

The entropy restricted to $B_\infty(\lambda)$ has maximum at $Po(\lambda)$. Further $H(bin(n, \lambda/n)) \to H(Po(\lambda))$ for $n \to \infty$.

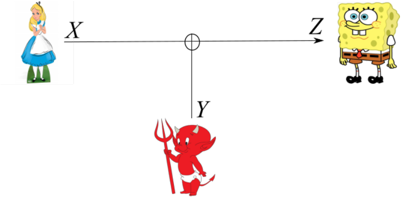
Universal coding interpretation

Assume that we are going to code a data point in $\mathbb{N}$ that are generated by some $P \in B_n(\lambda)$, but the exact distribution P is unknown. The code $\kappa : \mathbb{N} \to A^*$ is characterized by a the code length function $j \to |\kappa(j)|$ satisfying Kraft's inequality $\sum_j a^{|\kappa(j)|} \leq 1$ where $a = |A|$. The goal is to minimize the maximum mean code length.

$$\min_{\kappa} \max_{P} E_P(|\kappa(j)|).$$

The solution is $|\kappa(n)| = -\log(bin(n, p, j))$, i.e. use the code that is optimal if we knew $P = bin(n, \lambda/n)$.
Similarly, assume that we are going to code a data point in $\mathbb{N}$ that are generated by some $P \in B_n(\lambda)$, but both P and n are unknown. The it is optimal to code as if $P = Po(\lambda)$.

Relation to the Poisson channel



- The goal for Alice is to maximize $I(X,Z)$ over $X \in B_\infty(\lambda)$.
- The goal for the devil is to minimize $I(X,Z)$ over $Y \in B_\infty(\mu)$.

$$\begin{aligned} I(X,Z) &= H(X+Y) - H(X+Y \mid X) \\ &= H(X+Y) - H(Y \mid X) \\ &= H(X+Y) - H(Y). \end{aligned}$$

For any Y it is optimal for Alice to choose $X \sim Po(\lambda)$. If $X \sim Po(\lambda)$ then it is optimal for the devil to choose $Z \sim Po(\mu)$ [PH and C. Vignat, 2003].

Entropy power inequality

Theorem ([PH and C. Vignat 2004])

Assume that $X \sim bin(m, 1/2)$ *and* $Y \sim bin(n, 1/2)$. *Then*

$$e^{2H(X)} + e^{2H(Y)} \leq e^{2H(X+Y)}.$$

For $X \sim bin(m, p)$ and $Y \sim bin(n, q)$ the inequality does not hold for small values of m, n ☹
but it holds for sufficiently large values of m, n [N. Sharma, S. Das, S. Muthukrishnan, 2010].

Entropy power inequality

Theorem ([PH and C. Vignat 2004])

Assume that $X \sim bin(m, 1/2)$ *and* $Y \sim bin(n, 1/2)$. *Then*

$$e^{2H(X)} + e^{2H(Y)} \leq e^{2H(X+Y)}.$$

For $X \sim bin(m, p)$ and $Y \sim bin(n, q)$ the inequality does not hold for small values of m, n ☹
but it holds for sufficiently large values of m, n [N. Sharma, S. Das, S. Muthukrishnan, 2010].

From entropy to divergence

Bernoulli sum and hypergeometric distributions

For $P \in B_n(\lambda)$ we have

$$H(P) + D(P \| bin(n, \lambda/n)) \leq H(bin(n, \lambda/n))$$

so if $H(P_k) \to H_{\max}(B_n(\lambda))$ for $k \to \infty$ then $D(P_n \| bin(n, \lambda/n)) \to 0$ for $k \to \infty$.

Law of small numbers

Since $bin(n, \lambda/n) \in B_\infty(\lambda)$ we have

$$H(bin(n, \lambda/n)) + D(bin(n, \lambda/n) \| Po(\lambda)) \leq H(Po(\lambda))$$

so

$$H(bin(n, \lambda/n)) = H_{\max}(B_n(\lambda)) \to H_{\max}(B_\infty(\lambda))$$

for $k \to \infty$ then $D(bin(n, \lambda/n) \| Po(\lambda)) \to 0$ for $k \to \infty$.

Upper bounds on total variation

[Babour and Hall, 1984] has

$$\frac{1}{16}\min\left\{p, np^2\right\} \leq V\left(bin\left(n, p\right), Po\left(\lambda\right)\right)$$
$$\leq 2\min\left\{p, np^2\right\}$$

Upper bounds on total variation

[Babour and Hall, 1984] has

$$\frac{1}{16}\min\left\{p, np^2\right\} \leq V\left(bin\left(n, p\right), Po\left(\lambda\right)\right)$$
$$\leq 2\min\left\{p, np^2\right\}$$

A factor of 32 in difference between upper and lower bound ☹

Bounds on divergence

We have $D(P\|Q) = \sum f\left(\frac{p_i}{q_i}\right) \cdot q_i$ where $f(x) = x \ln(x)$. For

$$x - 1 \leq f(x) \leq x - 1 + (x-1)^2 .$$

Some better bound

$$\begin{aligned} x - 1 + \frac{1}{2}(x-1)^2 - \frac{1}{6}(x-1)^3 &\leq f(x) \\ &\leq x - 1 + \frac{1}{2}(x-1)^2 - \frac{1}{6}(x-1)^3 + \frac{1}{3}(x-1)^4 . \end{aligned}$$

$$\begin{aligned} D(P\|Q) &\leq \chi^2(P,Q) , \\ D(P\|Q) &\approx \frac{1}{2}\chi^2(P,Q) . \end{aligned}$$

Orthogonal polynomials

Assume that $f_0, f_1, f_2, \ldots$ are orthogonal normalized polynomials with respect to Q. Then

$$\frac{\mathrm{d}P}{\mathrm{d}Q}(x) = \sum_{i=0}^{\infty} f_i(x) \cdot \left\langle f_i \middle| \frac{\mathrm{d}P}{\mathrm{d}Q} \right\rangle ,$$

$$\begin{aligned} \left\langle f_i \middle| \frac{\mathrm{d}P}{\mathrm{d}Q} \right\rangle &= \int f_i(x) \frac{\mathrm{d}P}{\mathrm{d}Q}(x) \, \mathrm{d}Qx \\ &= E_P[f_i(X)] . \end{aligned}$$

Therefore

$$\chi^2(P, Q) = \sum_{i=1}^{\infty} (E_P[f_i(X)])^2 .$$

We have

$$
\begin{aligned}
D\left(bin\left(n,p\right)\|Po\left(\lambda\right)\right) &= \sum_{j=0}^{n} \ln\left(\frac{bin\left(n,p,j\right)}{Po\left(\lambda,j\right)}\right)\cdot bin\left(n,p,j\right)\\
&= \sum_{j=0}^{n} \ln\left(\frac{\binom{n}{j}p^{j}\left(1-p\right)^{n-j}}{\frac{\lambda^{j}}{j!}e^{-\lambda}}\right)\cdot bin\left(n,p,j\right)\\
&= \sum_{j=0}^{n}\left(\lambda+\left(n-j\right)\ln\left(1-p\right)+\ln\left(\frac{n^{\underline{j}}}{n^{j}}\right)\right)\cdot bin\left(n,p,j\right)\\
&= \lambda+\left(n-\lambda\right)\ln\left(1-p\right)+\sum_{j=0}^{n}\left(\ln\left(\prod_{i=0}^{j-1}\left(1-\frac{i}{n}\right)\right)\right)\cdot bin\left(n,p,j\right)\ .
\end{aligned}
$$

Stirling numbers

Expand

$$\ln\left(\prod_{i=0}^{j-1}\left(1-\frac{i}{n}\right)\right)=\sum_{i=0}^{j}\ln\left(1-\frac{i}{n}\right)$$
$$=-\sum_{i=0}^{j}\sum_{k=1}^{\infty}\frac{1}{k}\cdot\left(\frac{j}{n}\right)^{k}.$$

Introduce Stirling numbers

$$j^{\ell}=\sum_{m=1}^{\ell}j^{\underline{\ell}}\left\{\begin{array}{c}\ell\\ m\end{array}\right\},$$
$$j_{[m]}=\sum_{m=0}^{\ell}j^{\ell}\left[\begin{array}{c}\ell\\ m\end{array}\right].$$

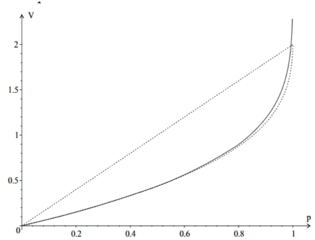
Truncations of these identities leads to inequalities.

Upper bounds

Theorem (PH and P. Ruzankin 2005)

For all

$$D\left(bin\left(n,p\right)\|Po\left(\lambda\right)\right)\leq-\frac{\ln\left(1-p\right)+p}{2}+\frac{p^{2}}{12n\left(1-p\right)}+\frac{p^{2}\left(2+11p+11p^{2}\right)}{12n^{2}\left(1-p\right)^{5}}.$$



Observe that $\limsup n^{2}\cdot D\left(bin\left(n,p\right)\|Po\left(\lambda\right)\right)\leq \lambda^{2}/4$.

Lower bound

Theorem

If $\lambda = np$ *then*

$$D\left(bin\left(n,p\right)\|Po\left(\lambda\right)\right) \geq \frac{p^2}{4}.$$

Key observation: Assume that $S_n \sim bin\left(n,p\right)$ and $Y \sim Po\left(\lambda\right)$ where $\lambda = np$. Then

$$E\left[S_n\right] = E\left[Y\right]$$

and

$$\begin{aligned} Var\left(S_n\right) &= np\left(1-p\right) \\ &< np \\ &= Var\left(Y\right). \end{aligned}$$

Improved rate of convergence

Theorem

Let $Po_\beta(\lambda)$ denote the information projection of $Po(\lambda)$ on the set of distributions with the same 1st and 2nd moment as $bin(n, \lambda/n)$. Then

$$n^2 \cdot D\left(bin\left(n, \lambda/n\right) \| Po_\beta(\lambda)\right) \to 0$$

for $n \to \infty$.

Proof.

We have

$$\begin{aligned} D(bin(n,p) \| Po(\lambda)) &= D(bin(n,p) \| Po_\beta(\lambda)) + D(Po_\beta(\lambda) \| Po(\lambda)) \\ &\geq D(bin(n,p) \| Po_\beta(\lambda)) + \frac{p^2}{4} \end{aligned}$$

Multiply both sides by n^2. $\square$

Poisson Charlier polynomials

The orthogonal polynomials with respect to $Po(\lambda)$ are

$$C_k^\lambda(x) = (\lambda k!)^{-1/2} \sum_{\ell=0}^{k} \binom{k}{\ell} (-\lambda)^{k-\ell} x^{\underline{\ell}}$$

If $E[X] = \lambda$ then

$$E\left[C_2^\lambda(X)\right] = \frac{Var(X) - \lambda}{2^{1/2}\lambda}$$

Conjecture For any random variable with $E\left[C_k^\lambda(X)\right] \leq 0$ we have

$$D(X \| Po(\lambda)) \geq \frac{1}{2}\left(E\left[C_k^\lambda(X)\right]\right)^2$$

The conjecture has been proved for $k = 1, 2$ and for any value of k when $E\left[C_k^\lambda(X)\right]$ is small [PH, Johnson and Kontoyiannis 2015].

Hypergeometric distributions and Bernoulli sums

A hypergeometric distribution is given by

$$\Pr\left(X=j\right)=\frac{\binom{K}{j}\binom{N-K}{n-j}}{\binom{N}{n}}$$

Then there exist $p_1, p_2, \ldots, p_n$ such that

$$\Pr\left(X=j\right)=\Pr\left(S_n=j\right)$$

where $S_n=\sum_{i=1}^{n} X_i$ is a Bernoulli sum and $\Pr\left(X_i=1\right)=p_i$. The mean is $E\left[S_n\right]=\sum p_i$. Then $bin\left(n,\bar{p}\right)$ has the same means as S_n if $\bar{p}=\frac{\sum p_i}{n}$. The variance is

$$\begin{aligned} Var\left(S_n\right) &= \sum p_i\left(1-p_i\right) \\ &\leq n\bar{p}\left(1-\bar{p}\right) \\ &= Var\left(bin\left(n,\bar{p}\right)\right) \end{aligned}$$

Kravchuk polynomials

The Kravchuk polynomials $\tilde{K}\,(n,x)$ are orthogonal with respect to $bin\,(n,p)$. are

$$C_k^\lambda\,(x) = (\lambda k!)^{-1/2} \sum_{\ell=0}^{k} \binom{k}{\ell} (-\lambda)^{k-\ell}\, x^{\underline{\ell}}$$

If $E\,[X] = \lambda$ then

$$E\left[C_2^\lambda\,(X)\right] = \frac{Var\,(X) - \lambda}{2^{1/2}\lambda}$$

Conjecture For any random variable with $E\left[\tilde{K}_k\,(X)\right] \leq 0$ we have

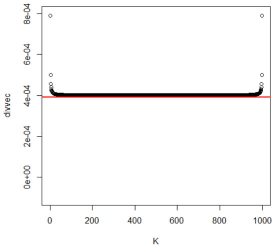
$$D\,(X\|bin\,(n,p)) \geq \frac{1}{2}\left(E\left[\tilde{K}_k\,(X)\right]\right)^2$$

The conjecture has been proved for $k = 1, 2$ and for any value of k when $E\left[C_k^\lambda\,(X)\right]$ is small [PH and F. Matúš, 2019].

Lower bound for hypergeometric distributions

The hypergeometric distribution satisfies

$$D\left(hyp\left(N,K,n\right)\|bin\left(n,\frac{K}{N}\right)\right)\geq\frac{n\left(n-1\right)}{4\left(N-1\right)^{2}}.$$



This result confirms the rule of thump:
Assume independence when sample size is less than 5 % of population size.

Upper bound for hypergeometric distributions

Stam 1978 proved

$$D\left(hyp\left(N,K,n\right)\|bin\left(n,\frac{K}{N}\right)\right)\leq\frac{n\left(n-1\right)}{2\left(N-1\right)\left(N-n+1\right)}.$$

Upper bound for hypergeometric distributions

Stam 1978 proved

$$D\left(hyp\left(N,K,n\right)\|bin\left(n,\frac{K}{N}\right)\right)\leq\frac{n\left(n-1\right)}{2\left(N-1\right)\left(N-n+1\right)}.$$

By taking higher order terms into account we get

$$D\left(hyp\left(N,K,n\right)\|bin\left(n,\frac{K}{N}\right)\right)\leq\frac{N\ln\frac{N-1/2}{N-n-3/2}-n+\frac{N}{N-n-1}}{N-1}.$$

Weak approximations

Let $N(\mu, \sigma^2)$ denote a Gaussian with mean μ and standard deviation σ. Then

$$D\left(N\left(\lambda, \sigma^2\right) \| N\left(\mu, \sigma^2\right)\right) = \frac{(\lambda - \mu)^2}{2\sigma^2}.$$

For the binomial distributions we have

$$D(bin(n, p) \| bin(n, q)) = n\left(p \ln \frac{p}{q} + (1-p) \ln \frac{1-p}{1-q}\right).$$

Signed log-likelihood

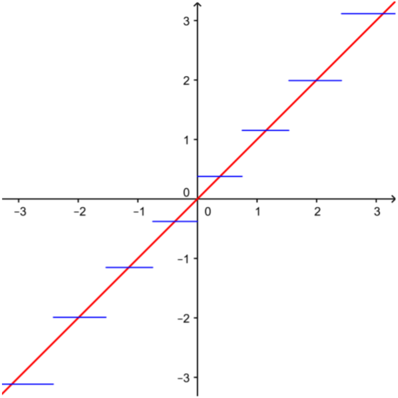
Let $\left(P^{\lambda}\right)$ denote elements of an exponential family in its mean value parametrization. Define

$$G(x) = \begin{cases} +\left(2D\left(P^{x} \| P^{\mu}\right)\right)^{1/2}, & \text{for } \lambda \geq \mu\,; \\ -\left(2D\left(P^{x} \| P^{\mu}\right)\right)^{1/2}, & \text{for } \lambda < \mu\,. \end{cases}$$

If $P^{\lambda} = N\left(\lambda, \sigma^2\right)$ then $G(x) = \frac{x-\mu}{\sigma}$.
For any exponential family $G(x) = \frac{x-\mu}{\sigma}$ is the first part of the Taylor expansion of G around $x = \mu$.

QQ-plot for binomial

Assume that $X \sim bin(n, p)$. For each $q \in (0,1)$ plot the q-quantile of a standard Gaussian against the q-quantile of $G(X)$.



$$\Pr(X < j) \leq \Pr(Z \leq G(j)) \leq \Pr(X \leq j).$$

[Serov and Zubkov, 2013]

Where do they intersect?

The intersection point is approximately given by the following result. If $X \sim bin\,(n, p)$ then if nq is an integer we have

$$\Pr\left(X \leq nq\right) = \Phi\left(G\left(j + c_q\right)\right) \cdot \left(1 + O\left(\frac{1}{n}\right)\right)$$

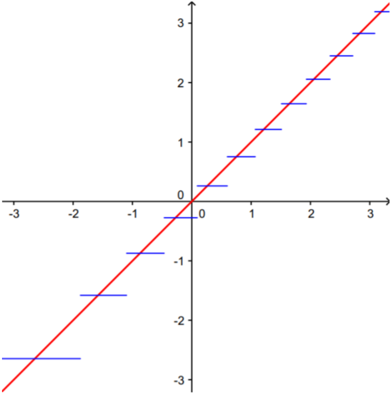
where

$$c_q = \frac{1}{2} + \frac{\ln\left(\frac{2D(q\|p)}{(q-p)^2}\,p\left(1-p\right)\right)}{2\ln\left(\frac{q(1-p)}{p(1-q)}\right)}.$$

[PH, L. Györfi and G. Tusnády, 2012]

QQ-plot for Poisson

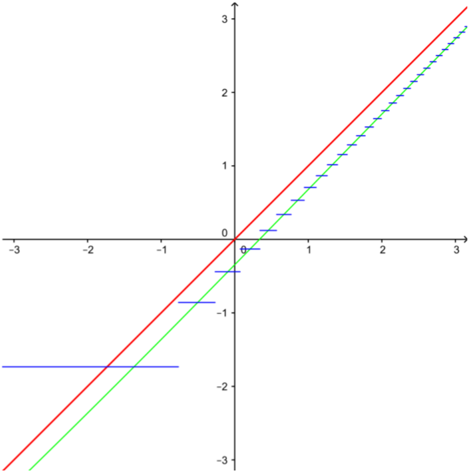
Assume that $X \sim Po(\lambda)$. For each $q \in (0,1)$ plot the q-quantile of a standard Gaussian against the q-quantile of $G(X)$.



[PH and Tusnády, 2011]

QQ-plot for negative binomial

Assume that $X \sim negbin(k, p)$. For each $q \in (0,1)$ plot the q-quantile of a standard Gaussian against the q-quantile of $G(X)$. [PH 2016]

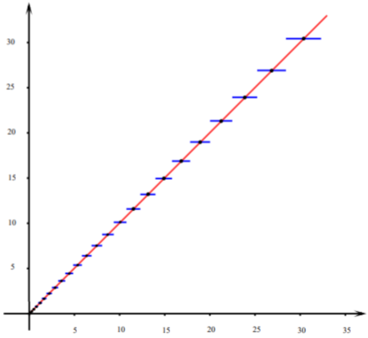


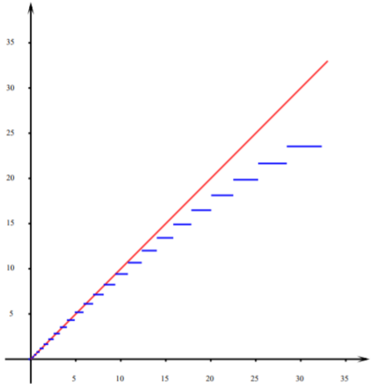
Proof structure

1. Prove majorization for Gamma distributions.
2. Prove intersection for negative binomial and Gamma distributions.
3. Combine to get upper bound for binomial.
4. Use upper bound on the binomial variable $n - X$ to get a lower bound for X.

Application

Information divergence is more χ^2-distributed than the χ^2-statistic.





Conclusion

- If you expand too little you will get punished by a factor of 2.
- Lower bounds can be found using othogonal polynomials.
- Saddlepolint approximations can often be replaced by powerful inequalities.
- Use information divergence rather than total variation or χ^2-divergence.

Work in progress:

- Simplify upper bounds.
- Bounds on moment generating functions.
- Generalizations to multivariate disributions.