A Conjecture

Jenő Reiczigel[1] Lídia Rejtő[2,3] Gábor Tusnády[3]

October 7, 2011

Abstract

Let $\varepsilon_1, \ldots, \varepsilon_m$ be i.i.d. random variables with

$$P(\varepsilon_i = 1) = P(\varepsilon_i = -1) = 1/2,$$

and $X_m = \sum_{i=1}^{m} \varepsilon_i$. Let Y_m be a normal random variable with the same first two moments as that of X_m. There is a uniquely determined function Ψ_m such that the distribution of $\Psi_m(Y_m)$ equals to the distribution of X_m. Tusnády's inequality states that

$$| \Psi_m(Y_m) - Y_m | \leq \frac{Y_m^2}{m} + 1.$$

Here we propose a sharpened version of this inequality.
AMS 2000 subject classification. Primary 62E17; secondary 62B15
Key words and phrases. Quantile transformation; normal approximation; binomial distribution; Tusnády's inequality

1 Conjecture

Let $\varepsilon_1, \ldots, \varepsilon_m$ be i.i.d. random variables with

$$P(\varepsilon_i = 1) = P(\varepsilon_i = -1) = 1/2,$$

and $X_m = \sum_{i=1}^{m} \varepsilon_i$. Let Y_m be a normal random variable with the same first two moments as that of X_m. Using quantile transformation we can

[1]Szent István University, Department of Biomathematics and Informatics, Faculty of Veterinary Science, Budapest, Hungary
[2]University of Delaware, Statistics Program, FREC, CANR, Newark, Delaware, USA
[3]Alfréd Rényi Mathematical Institute of the Hungarian Academy of Sciences, Budapest, Hungary

see that there is a uniquely determined function Ψ_m such that the distribution of $\Psi_m(Y_m)$ equals to the distribution of X_m. The central limit theorem implies that the function Ψ_m is close to the identity for large m. A sharp inequality of Tusnády [12] raised certain interest in the literature ([1],[2],[3],[4],[5],[6],[7],[8],[9],[10],[11],[13]).

Let us define the function f on the interval $(0,1)$ as

$$f(x) = \sqrt{(1+x)\log(1+x) + (1-x)\log(1-x)},$$

set $f(0) = 0, f(1) = \sqrt{\log(4)}$. Let us put

$$x_{k,m} = \frac{k - \frac{m}{2}}{\frac{m}{2}}$$

for positive even integers m with k such that $m/2 < k \leq m$, and set

$$p_{k,m} = P(X_m \geq k) = 2^{-m} \sum_{i=k}^{m} \binom{m}{i}.$$

Let us define the function Q on the reals as

$$Q(x) = \frac{1}{\sqrt{2\pi}} \int_x^\infty e^{-u^2/2} du.$$

With those ingredients our conjecture states that

$$Q(\sqrt{m} f(x_{k,m})) < p_{k,m} < Q(\sqrt{m} f(x_{k-1,m}))$$

holds true for $\frac{m}{2} < k \leq m$. Or more sharply

$$2(k-1) - \frac{m}{2} + 0.8964 < m f^{-1}(Q^{-1}(p_{k,m})/\sqrt{m}) < 2(k-1) - \frac{m}{2} + 1.0000 \quad (1)$$

holds true with pessimal parameters $m = k = 10$. It implies that Tusnády's inequality is sharpened to

$$\left| \Psi_m(Y_m) - m f^{-1}\left(\frac{Y_m}{m}\right) \right| < 1.1036.$$

2 Generalization

For an arbitrary random variable X let us consider the function on reals

$$R(t) = E e^{tX}$$

restricting ourselves for distributions having finite momentum generators. Next we define

$$\psi(t) = \frac{R'(t)}{R(t)},$$

$$\alpha(x) = t \quad \text{iff} \quad \psi(t) = x,$$

$$\rho(x) = R(\alpha(x)) \exp(-x\alpha(x)).$$

The probability $P(\sum_{i=1}^{m} X_i \geq mx)$ is approximately $\rho(x)^{-m}$ if $x > EX$. The function ρ depends on the distribution of X, it is the Chernoff function of X. Let us denote the Chernoff function of the distribution F of X by ρ_F, and the corresponding function for standard normal by ρ_G. The quantile transformation between the partial sums of distribution F with Gaussian ones resemble us to the equation

$$\rho_F(x) = \rho_G(y)$$

having the property that it gives sharp values for any m. Perhaps the error term is bounded with a bound depending on the distribution of X. For the case symmetrical binomial distribution the error term might be as small as that the quantile curve jumps over its limiting function: it is the informal explanation of our conjecture.

3 Numerical Illustration

The function Ψ_m is shown in Figure 1. called "step" for $m = 50$ with a rescaling for random variables

$$\xi_m = \frac{X_m}{m}, \quad \eta_m = \frac{Y_m}{m}.$$

The function f is called "limit", for the sequence of step functions goes to f after rescaling. The conjecture comes from the observation that the limit function crosses all steps near to their middle. Let us introduce the blownup error term

$$\Delta_{k,m} = 10 \left(2k - 1 - m \, f^{-1} \left(\frac{1}{\sqrt{m}} \, Q^{-1} \left(\sum_{i=k}^{m} \binom{m}{i} 2^{-m} \right) \right) \right),$$

for $0 < k \leq m/2$. In Figure 1. it is labelled as "Delta". With these notations (1) is equivalent with $0 < \Delta_{k,m} < 1.036$. These error terms are shown in Figure 2. for $2 \leq m \leq 1000$. Figure 2. prompts the conjecture that even these curves are convergent. We are a bit perplexed: even the inequality

$0 < \Delta_{1,2} < 1.036$ means that $Q(0.723359) < 0.25 < Q(0.6435214)$. How can we prove such an inequality theoretically?

References

[1] Bretagnolle, J. and Massart, P. (1989). Hungarian constructions from the nonasymptotic viewpoint, *The Annals of Probability* **17**, 239–256.

[2] Castelle, N. (2009). Improvement of two Hungarian bivariate theorems, Manuscript

[3] Carter, A. and Pollard, D. (2004). Tusnády's inequality revisited, *The Annals of Statistics* **32**, 2731–2741.

[4] Csörgő, M. (2007). A glimpse of the KMT (1975): approximation of empirical processes by Brownian bridges via quantiles, Acta Sci. Math. (Sseged) **73**, 349–366

[5] Csörgő, M. and Révész, P. (1981). *Strong Approximations in Probability and Statistics*, Academic Press, New York.

[6] Dudley, R. M. (1999). *Notes on empirical processes,* Lecture notes for a course given at Aarhus Univ., August 24, 1999.

[7] Dudley, R. M. (2008). On the quantile transformation for asymmetric binomial and hypergeometric distributions, Manuscript

[8] Major, P. (1999). *The approximation of the empirical distribution function,* Technical report, Alfréd Rényi Mathematical Institute of the Hungarian Academy of Sciences. Notes available from http://www.renyi.hu/ ~major/probability/empir.html, August 3, 1999.

[9] Mason, D. M. (2001). Notes on the KMT Brownian bridge approximation to the uniform empirical process, In *Asymptotic Methods in Probability and Statistics with Applications*, (N. Balakrishnan, I. A. Ibragimov and V. B. Nevzorov, eds.) 351–369. Birkhäuser, Boston.

[10] Massart, P. (2002). Tusnády's lemma, 24 years later, *Ann. Inst. H. Poincaré Probab. Statist.* **38**, 991–1007.

[11] Pollard, D. (2002). *A user's guide to measure theoretic probability,* Cambridge University Press

[12] Tusnády, G. (1977). *Study of statistical hypotheses*, Dissertation for Habilitation, Hungarian Academy of Sciences, Budapest.

[13] Zhou, H. H. (2006). A note on quantile coupling inequalities and their applications, Manuscript

4 Appendix

R- program of Figures 1 and 2.

```
Q=function(p) -qnorm(p)
G=function(x) ((1+x)*log(1+x)+(1-x)*log(1-x))**0.5
Ginv=function(u) {
GG=function(x) G(x)-u
uniroot(GG,c(0,1),f.lower=-u,f.upper=log(4)^.5-u,tol=10^-100)
}

m=50; k=m/2
sum=0; divisor=2**m; bin=
xx=c(1:k+1); yy=c(1:k+1); zz=c(1:k+1);

for (i in 1:k-1){
sum=sum+bin
x=(m-2*i)/m
y=Q(sum/divisor)/(m**.5)
b=Ginv(y)$root
yy[i+1]=y; xx[i+1]=x
bin=(m-i)*bin/(i+1)
zz[i+1]=10*(m-2*i-1-m*b)}
xx[k+1]=0; yy[k+1]=0; zz[k+1]=0
kerx=c(0,1.25); kery=c(0,1.15)
plot(kerx, kery, type="n",xlab="eta", ylab="xi",
main="Figure1. Quantile transform, its limit and blownup error, m=50")

for (i in 1:k){
bb=seq(from=yy[i+1], to=yy[i], by=0.01)
cc=bb*0+1; cc=cc*xx[i+1]
points(bb,cc,type="l", col="blue", lwd=2)}
cc=seq(from=0, to=0.999, by=0.001)
bb=((1+cc)*log(1+cc)+(1-cc)*log(1-cc))**0.5
points(bb,cc, type="l", col="red", lwd=2)
points(yy,zz, type="l", col="green", lwd=2)
legend(locator(1),c("Limit","Step","Delta"),
lty=c(1,1,1),
col=c("red","blue","green"))
```

```
kerx=c(0,1.25); kery=c(0,1.15)
plot(kerx, kery, type="n",xlab="eta", ylab="Delta",
      main="Figure 2. The blownup error")

for (k in 1:500){m=2*k;
sum=0; divisor=2**m; bin=1
 yy=c(1:k+1); zz=c(1:k+1);
for (i in 1:k-1){
sum=sum+bin
y=Q(sum/divisor)/(m**.5)
b=Ginv(y)$root
yy[i+1]=y;
bin=(m-i)*bin/(i+1)
zz[i+1]=10*(m-2*i-1-m*b)}
 yy[k+1]=0; zz[k+1]=0
 if (k<100) clr="red" else
 if (k<200) clr="blue" else
 if (k<300) clr="purple" else
 if (k<400) clr="gray" else clr="green"
points(yy,zz, type="l", col=clr)}
legend(locator(1),c("0<m <= 200","200<m<=400","400<m<=600",
"600<m<=800","800<m<=1000"),
lty=c(1,1,1,1,1),
col=c("red","blue","purple","gray","green"))
```