EXTENDING CENTRAL LIMIT THEOREM

Independent sum

Consider RIS of size n from a distribution with the mean of μ, the standard deviation of σ and MGF $M(t)$. The distribution of

$$Z_n \equiv \frac{\sum_{k=1}^{n} X_k - n\mu}{\sigma\sqrt{n}}$$

tends, as $n \to \infty$, to $\mathcal{N}(0,1)$. This can be seen by first expanding $\ln M(t)$ in terms of so called cumulants:

$$\ln M(t) = \sum_{j=1}^{\infty} \frac{\kappa_j t^j}{j!}$$

where $\kappa_1 = \mu$, $\kappa_2 = \sigma^2$, $\kappa_3 = \mathbb{E}\left[(X-\mu)^3\right]$, $\kappa_4 = \mathbb{E}\left[(X-\mu)^4\right] - 3\sigma^4$, etc. Note that the conversion between cumulants and central moments is done by expanding

$$\ln M_{X-\mu}(t) = -\mu t + \ln M_X(t) = \ln\left(1 + \sum_{j=2}^{\infty} \frac{\mu_j t^j}{j!}\right)$$

where μ_j denote the corresponding central moment.

The MGF of Z_n is constructed by taking the MGF of $X - \mu$, namely

$$\exp\left(\sum_{j=2}^{\infty} \frac{\kappa_j t^j}{j!}\right)$$

raising it to the power of n, and finally replacing t by $\frac{t}{\sigma\sqrt{n}}$:

$$\exp\left(\sum_{j=2}^{\infty} \frac{\kappa_j t^j}{n^{j/2-1}\sigma^j j!}\right) = \exp(\tfrac{t^2}{2}) \cdot \exp\left(\sum_{j=3}^{\infty} \frac{\kappa_j t^j}{n^{j/2-1}\sigma^j j!}\right)$$

The last expression clearly tends, as $n \to \infty$, to $\exp(\frac{t^2}{2})$ of the Normal distribution, thus proving the Central Limit Theorem.

To get a better approximation, we expand the last factor in powers of $\frac{1}{\sqrt{n}}$, thus:

$$1 + \frac{\kappa_3 t^3}{6\sigma^3\sqrt{n}} + \frac{3\sigma^2\kappa_4 t^4 + \kappa_3^2\ t^6}{72\sigma^6 n} + \frac{54\sigma^4\kappa_5 t^5 + 45\kappa_3\kappa_4\sigma^2 t^7 + 5\kappa_3^3\ t^9}{6480\sigma^9 n^{3/2}} + ...$$

To incorporate terms of this expansion in our approximation, we need the following Fourier-transform inverse

$$\exp(\tfrac{t^2}{2}) \cdot t^j \to \frac{\exp(-\frac{z^2}{2})}{\sqrt{2\pi}} H_j(z)$$

where $H_j(z)$ are monic polynomials (closely related to the usual Hermite polynomials), namely

0	1
1	z
2	z^2-1
3	$z(z^2-3)$
4	z^4-6z^2+3
5	$z(z^4-10z^2+15)$
6	$z^6-15z^4+45z^2-15$

Note that

$$H_{j+1}(z) = z \cdot H_j(z) - H_j'(z)$$

which is a consequence of: multiplying a MGF by t results in *differentiating* the corresponding Fourier-transform inverse (which can be used repeatedly).

`EXAMPLE`: Suppose X is exponential, with $\mu = 1$. Z_n is thus equal to $Y - \sqrt{n}$ where $Y \in \gamma(n, \frac{1}{\sqrt{n}})$, which means that is exact pdf is

$$\frac{(z+\sqrt{n})^{n-1} \exp[-(z+\sqrt{n})\sqrt{n}]}{(n-1)!} n^{n/2} \qquad \text{for } z > -\sqrt{n}$$

We can easily find $\sigma = 1$, $\kappa_3 = 2$ and $\kappa_4 = 6$. The previous distribution can be better approximated (discarding $\frac{1}{n^{3/2}}$ terms and beyond) by

$$\frac{\exp(-\frac{z^2}{2})}{\sqrt{2\pi}} \left(1 + \frac{z(z^2-3)}{3\sqrt{n}} + \frac{2z^6-21z^4+36z^2-3}{36n} + ...\right)$$

General case

This technique can be extended to the case of $g(\bar{X}, \bar{Y}, ...)$. First we have to define

$$Z_n \equiv \frac{g - \mathbb{E}[g]}{\sqrt{\text{Var}[g]}}$$

and then compute the first few terms of

$$1 + \frac{\kappa_3 t^3}{6\sigma^3} + \frac{3\sigma^2\kappa_4 t^4 + \kappa_3^2\ t^6}{72\sigma^6} + ...$$

where now the cumulants and σ are those of $g(\bar{X}, \bar{Y}, ...)$, correspondingly expanded in powers of $\frac{1}{n}$.

A common technique for incorporating the skewness correction is to consider a function (transformation) of g, say $H(g)$, such that the skewness of H is, to the $\frac{1}{\sqrt{n}}$ approximation, equal to zero.

`EXAMPLE:` Returning to the exponential distribution (with mean β), we investigate the skewness of $g(\bar{X})$. According to our previous formula, it is equal to

$$\frac{g'(\mu)^3\mu_3 + 3g'(\mu)^2 g''(\mu)\sigma^4}{n^2} =$$

$$[g'(\beta)]^3 \frac{2\beta^3}{n^2} + 3[g'(\beta)]^2 g''(\beta)\frac{\beta^4}{n^2} + O(\frac{1}{n^3})$$

Making this equal to zero one has to solve

$$g'(\beta) + \frac{3}{2} g''(\beta) \cdot \beta = 0$$

or

$$\begin{array}{rcl} \frac{g''(\beta)}{g'(\beta)} & = & \frac{-2}{3\beta} \\ \ln g'(\beta) & = & -\frac{2}{3} \ln \beta + \tilde{c} \\ g'(\beta) & = & c.\beta^{-2/3} \end{array}$$

whose simplest solution is $g = \sqrt[3]{\beta}$.

For the expected value, we then get

$$\begin{array}{c} g(\mu) + \frac{g''(\mu)\sigma^2}{2n} + ... = \\ \sqrt[3]{\beta} - \frac{\beta^{-5/3}.\beta^2}{9n} ... = \sqrt[3]{\beta} - \frac{\sqrt[3]{\beta}}{9n} ... \end{array}$$

Similarly, the variance becomes

$$\begin{array}{c} \frac{g'(\mu)^2\sigma^2}{n} + \frac{g'(\mu)g''(\mu)\kappa_3 + \frac{1}{2}g''(\mu)^2\sigma^4 + g'(\mu)g'''(\mu)\sigma^4}{n^2} + ... = \\ \frac{\beta^{2/3}}{9n} + \frac{-\frac{4}{27}\beta^{2/3} + \frac{2}{81}\beta^{2/3} + \frac{10}{81}\beta^{2/3}}{n^2} + ... = \frac{\beta^{2/3}}{9n} + ... \end{array}$$

the third central moment is (to this level of approximation) equal to zero be design, and the forth cumulant is

$$\begin{array}{c} \frac{g'(\mu)^2\left[g'(\mu)^2\kappa_4 + 12g'(\mu)g''(\mu)\kappa_3\sigma^2 + 12g''(\mu)^2\sigma^6 + 4g'(\mu)g'''(\mu)\sigma^6\right]}{n^3} + ... = \\ \frac{\frac{1}{9}\left[\frac{2}{3} - \frac{16}{9} + \frac{16}{27} + \frac{40}{81}\right]\beta^{4/3}}{n^3} + ... = \frac{2\beta^{4/3}}{9^3 n^3} \end{array}$$

The distribution of

$$\frac{\sqrt[3]{\bar{X}} - \sqrt[3]{\beta} + \frac{\sqrt[3]{\beta}}{9n} + ...}{\frac{\sqrt[3]{\beta}}{3\sqrt{n}} + ...} = 3\sqrt{n}\left(\sqrt[3]{\frac{\bar{X}}{\beta}} - 1 + \frac{1}{9n} + ...\right)$$

is approaching $\mathcal{N}(0,1)$ a lot faster than

$$\frac{\bar{X} - \beta}{\frac{\beta}{\sqrt{n}}}$$

does. We can make it three times more accurate by using

$$\frac{e^{-z^2/2}}{\sqrt{2\pi}}\left(1 + \frac{2H_4(z)}{24 \times 9n}\right)$$

`Second EXAMPLE:` We would like to have a good approximation to the distribution of

$$r \equiv \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2 \cdot \sum_{i=1}^n (Y_i - \bar{Y})^2}}$$

Its exact distribution is known, but it is rather complicated.

One can show that the first four cumulants of r, divided by the corresponding power of σ_r and expanded to the appropriate level of accuracy, are

$$\begin{aligned}
\mu_r &= \rho\left(1-\tfrac{1-\rho^2}{n}+...\right)\\
\sigma_r^2 &= \tfrac{(1-\rho^2)^2}{n}\left(1+\tfrac{2+11\rho^2}{2n}+...\right)\\
\tfrac{\kappa_3}{\sigma_r^3} &= \tfrac{-6\rho}{\sqrt{n}}+...\\
\tfrac{\kappa_4}{\sigma_r^4} &= \tfrac{72\rho^2-6}{n}+...
\end{aligned}$$

It is easy to construct the approximate pdf of $\frac{r-\mu_r}{\sigma_r}$.

If a similar procedure is followed to find the skewness of $g(r)$, one gets:

$$-3\tfrac{2\rho g'(\rho)-(1-\rho^2)g''(\rho)}{|g'(\rho)|\sqrt{n}}+...$$

To make it identically equal to zero, one needs to solve

$$\tfrac{2\rho}{1-\rho^2}=\tfrac{g''(\rho)}{g'(\rho)}$$

or

$$\begin{aligned}
-\ln(1-\rho^2) &= \ln g'(\rho)\\
g'(\rho) &= \tfrac{1}{1-\rho^2}\\
g(\rho) &= \int\tfrac{d\rho}{1-\rho^2}=\tfrac{1}{2}\ln\tfrac{1+\rho}{1-\rho}\equiv \operatorname{arc\,tanh}\rho
\end{aligned}$$

This implies that $\operatorname{arctanh} r$ has zero skewness (to this level of accuracy), and can thus be approximated by the (modified) Normal distribution a lot more easily. We just have to re-compute

$$\begin{aligned}
\mu_g &= \operatorname{arc\,tanh}\rho+\tfrac{\rho}{2n}+...\\
\sigma_g^2 &= \tfrac{1}{n}+\tfrac{6-\rho^2}{2n^2}+...
\end{aligned}$$

and

$$\tfrac{\kappa_4}{\sigma_g^4}=\tfrac{2}{n}+...$$

Now, it is even easier to construct the approximate pdf of $\frac{\operatorname{arc\,tanh} r-\mu_g}{\sigma_g}$.

Bivariate and multivariate extension

Again, the first (and easiest) possibility is that we are dealing with two or more sample means, say $\bar{X}$, $\bar{Y}$, ... To find a good approximation for their joint distribution, we must first construct the joint MGF of the distribution from which we sample (should this become too difficult, we substitute the corresponding expansion, up to and including the fourth moments). We then change it to the MGF of $X_i-\mu_x$, $Y_i-\mu_y$, ... by multiplying it by

$$e^{-\mu_x t_1-\mu_y t_2-...}$$

and finally to the MGF of $a(X_i-\mu_x)$, $a(Y_i-\mu_y)$, ... by replacing each t_i by at_i (where a represents $\frac{1}{\sqrt{n}}$).

Multiplying the log of this MGF by n (i.e. dividing it by a^2) yields the CGF of $\sqrt{n}(\bar{X}-\mu_x)$, $\sqrt{n}(\bar{Y}-\mu_y)$, ... This, we expand in powers of a (up to and including a^2 terms), getting $C_0(t_1,t_2,..)+aC_1(t_1,t_2,..)+a^2C_2(t_1,t_2,..)+...$

Taking the inverse transform of

$$e^{C_0}\cdot\left(1+aC_1+a^2C_2+\tfrac{1}{2}a^2C_1^2+...\right)$$

yields the joint pdf of $\sqrt{n}(\bar{X}-\mu_x)$, $\sqrt{n}(\bar{Y}-\mu_y)$, ... To make it into the joint pdf of $\bar{X}$, $\bar{Y}$, ..., we have to make the following replacement

$$\begin{array}{rcl} x & \rightarrow & \frac{x-\mu_x}{a} \\ y & \rightarrow & \frac{y-\mu_y}{a} \\ & \vdots & \end{array}$$

multiplying the result by a^2 (a^3 for three variables, etc.). Finally. we change a to $\frac{1}{\sqrt{n}}$.

To find a good approximation to the joint pdf of $\frac{g(\bar{X},\bar{Y},..)-\mu_g}{\sigma_g}$, $\frac{h(\bar{X},\bar{Y},..)-\mu_h}{\sigma_h}$, ...(now it is more convenient to work in the 'Z scale'), we would have to work out the mean of each $g(\bar{X},\bar{Y},..)$, $h(\bar{X},\bar{Y},..)$, ... to the $\frac{1}{n}$ accuracy, the variance and all covariances to the $\frac{1}{n^2}$ accuracy, all third-order cumulants to the $\frac{1}{n^2}$ accuracy, and all forth-order cumulants to the $\frac{1}{n^3}$ accuracy. The CGF of $\frac{g(\bar{X},\bar{Y},..)-\mu_g}{\sigma_g}$ and $\frac{h(\bar{X},\bar{Y},..)-\mu_h}{\sigma_h}$ (I will show the bivariate case only) can then be written as

$$\frac{t_1^2+t_2^2+2\rho_{gh}t_1t_2}{2}+\frac{\gamma_{30}t_1^3+3\gamma_{21}t_1^2t_2+3\gamma_{12}t_1t_2^2+\gamma_{03}t_2^3}{6}+$$
$$\frac{\gamma_{40}t_1^4+4\gamma_{31}t_1^3t_2+6\gamma_{22}t_1^2t_2^2+4\gamma_{13}t_1t_2^3+\gamma_{04}t_2^4}{6}$$

where γ_{ij} are the 'normalized' cumulants, i.e. $\frac{\kappa_{ij}}{\sigma_g^i\sigma_h^j}$. Note that the third-order (fourth-order) γ cumulants are proportional to $\frac{1}{\sqrt{n}}$ ($\frac{1}{n}$). Also note that ρ_{gh} would (in general) consist of an absolute term, and an $\frac{1}{n}$ proportional term, say $\rho_0+\frac{\rho_1}{n}+...$ To get the answer, we ask Maple to give us the inverse transform of

$$\exp\left(\frac{t_1^2+t_2^2+2\rho_{gh}t_1t_2}{2}\right)\cdot\left[1+\frac{\gamma_{30}t_1^3+3\gamma_{21}t_1^2t_2+3\gamma_{12}t_1t_2^2+\gamma_{03}t_2^3}{6}+\right.$$
$$\frac{\gamma_{40}t_1^4+4\gamma_{31}t_1^3t_2+6\gamma_{22}t_1^2t_2^2+4\gamma_{13}t_1t_2^3+\gamma_{04}t_2^4}{6}+$$
$$\left.\left(\frac{\gamma_{30}t_1^3+3\gamma_{21}t_1^2t_2+3\gamma_{12}t_1t_2^2+\gamma_{03}t_2^3}{6}\right)^2+...\right]$$

`Appendix A:` Third-order cumulants are just the corresponding 3^{rd} central moments. Fourth-order cumulants are a bit more tricky, in general

$$\kappa_{1111}=\mu_{1111}-\mu_{1100}\mu_{0011}-\mu_{1010}\mu_{0101}-\mu_{1001}\mu_{0110}$$

This reduces to the old

$$\kappa_4 = \mu_4 - 3\sigma^4$$

in case of one random variable, to

$$\begin{aligned} \kappa_{31} &= \mu_{31} - 3\mu_{20}\mu_{11} \\ \kappa_{22} &= \mu_{22} - \mu_{20}\mu_{02} - 2\mu_{11}^2 \end{aligned}$$

in case of two, and to

$$\kappa_{211} = \mu_{211} - \mu_{200}\mu_{011} - 2\mu_{110}\mu_{101}$$

in case of three.

`Appendix B:` The inverse transform of

$$\exp\left(\frac{t_1^2+t_2^2+2\rho_0 t_1 t_2}{2}\right)$$

is

$$\frac{\exp\left[\frac{x^2+y^2-2\rho_0 xy}{2(1-\rho^2)}\right]}{2\pi\sqrt{1-\rho^2}}$$

(this is done in MATH 2P81/2). If the first function is multiplied by $t_1^j t_2^k$, the second function is differentiated j times with respect to x and k times with respect to y. Thus, we get

j	k	$H(x,y)$
1	1	$XY + \frac{\rho}{1-\rho^2}$
2	1	$X^2Y - \frac{Y-2\rho X}{1-\rho^2}$
3	1	$X^3Y - \frac{3X(Y-\rho X)}{1-\rho^2} - \frac{3\rho}{(1-\rho^2)^2}$
2	2	$X^2Y^2 - \frac{X^2+Y^2-4\rho XY}{1-\rho^2} + \frac{1+2\rho^2}{(1-\rho^2)^2}$

where $X \equiv \frac{x-\rho y}{1-\rho^2}$ and $Y \equiv \frac{y-\rho x}{1-\rho^2}$

For more than two variables, one would express the MGF in the vector form of

$$\exp\left(\frac{\mathbf{t}^{\mathrm{T}}\mathbb{C}\mathbf{t}}{2}\right)$$

and express its inverse as

$$\frac{\exp\left(\frac{\mathbf{x}^{\mathrm{T}}\mathbb{C}^{-1}\mathbf{x}}{2}\right)}{(2\pi)^{\ell/2}\sqrt{\det\mathbb{C}}}$$

where ℓ is the number of variables. Multiplying the first function by a product of t powers, one would have to correspondingly differentiate the second function.

`Example:` The method-of-moments estimators of α and β parameters of the gamma distribution are

$$\frac{\bar{X}^2}{\frac{\overline{X^2}-\bar{X}^2}{\bar{X}}}$$

$$\frac{\overline{X^2}-\bar{X}^2}{\bar{X}}$$

respectively.

Expanding these, one gets:

$$\alpha + 2\frac{1+\alpha}{\beta}(\bar{X}-\mu_x) - \frac{\overline{X^2}-\mu_y}{\beta^2} + \frac{1+5\alpha+4\alpha^2}{\alpha\beta^2}(\bar{X}-\mu_x)^2$$
$$+\frac{(\overline{X^2}-\mu_y)^2}{\alpha\beta^4} - 2\frac{1+2\alpha}{\alpha\beta^3}(\bar{X}-\mu_x)(\overline{X^2}-\mu_y) + ...$$

and

$$\beta - \frac{1+2\alpha}{\alpha}(\bar{X}-\mu_x) + \frac{\overline{X^2}-\mu_y}{\alpha\beta} + \frac{1+\alpha}{\alpha^2\beta}(\bar{X}-\mu_x)^2$$
$$-\frac{(\bar{X}-\mu_x)(\overline{X^2}-\mu_y)}{\alpha^2\beta^2} + ...$$

where $\mu_x = \alpha\beta$ and $\mu_y = \mathbb{E}(X^2) = \alpha(1+\alpha)\beta^2$.

Using

$$\begin{aligned}
\mathbb{E}\left[(X-\mu_x)^2\right] &= \alpha\beta^2 \\
\mathbb{E}\left[(X^2-\mu_y)^2\right] &= 2\alpha(1+\alpha)(3+2\alpha)\beta^4 \\
\mathbb{E}\left[(X-\mu_x)(X^2-\mu_y)\right] &= 2\alpha(1+\alpha)\beta^3
\end{aligned}$$

and similar formulas for third powers, one finds that the expected values of the two estimators are $\alpha+\frac{3(1+\alpha)}{n}+...$ and $\beta-\frac{(1+\alpha)\beta}{n\alpha}+...$, their variances are $\frac{2\alpha(1+\alpha)}{n}+...$ and $\frac{(3+2\alpha)\beta^2}{n\alpha}+...$, and their correlation coefficient equals $-\sqrt{\frac{2(1+\alpha)}{3+2\alpha}}+....$ The four values of skewness are

$$2(2\alpha-1)\sqrt{\frac{2}{n\alpha(1+\alpha)}} + ...$$

$$4(2-\alpha)\sqrt{\frac{1}{n\alpha(3+2\alpha)}} + ...$$

$$-14\sqrt{\frac{2(1+\alpha)}{n\alpha(3+2\alpha)}} + ...$$

$$\frac{46+52\alpha+8\alpha^2}{3+2\alpha}\sqrt{\frac{1}{n\alpha(3+2\alpha)}} + ...$$

which would enable us to build a $\frac{1}{\sqrt{n}}$ accurate approximation.

The maximum likelihood technique gives the following (different, but more complicated) estimators:

$$\begin{aligned}
\hat{\alpha} &= g\left(\ln\bar{X} - \overline{\ln X}\right) \\
\hat{\beta} &= \frac{\bar{X}}{\hat{\alpha}}
\end{aligned}$$

where $g(x)$ is the inverse function to $\ln x - \Psi(x)$.

Expanding these, one gets:

$$\alpha + \frac{\bar{X}-\mu_x}{\beta[1-\alpha\Psi'(\alpha)]} - \frac{\alpha}{1-\alpha\Psi'(\alpha)}(\overline{\ln X}-\mu_y)$$
$$+\frac{2\Psi'(\alpha)-\alpha\Psi'(\alpha)^2+\alpha\Psi''(\alpha)}{2\beta^2[1-\alpha\Psi(\alpha)]^3}(\bar{X}-\mu_x)^2 + \frac{\alpha[1+\alpha^2\Psi''(\alpha)]}{2[1-\alpha\Psi(\alpha)]^3}(\overline{\ln X}-\mu_y)^2$$
$$-\frac{1+\alpha^2\Psi''(\alpha)}{\beta[1-\alpha\Psi(\alpha)]^3}(\bar{X}-\mu_x)(\overline{\ln X}-\mu_y) + ...$$

and

$$\begin{array}{l}\beta - \frac{\bar{X}-\mu_x}{1-\alpha\Psi'(\alpha)} + \frac{\beta}{1-\alpha\Psi'(\alpha)}(\overline{\ln X} - \mu_y) \\ -\frac{\Psi'(\alpha)^2+\Psi''(\alpha)}{2\beta[1-\alpha\Psi(\alpha)]^3}(\bar{X}-\mu_x)^2 + \frac{\beta[1-2\alpha\Psi'(\alpha)-\alpha^2\Psi''(\alpha)]}{2[1-\alpha\Psi(\alpha)]^3}(\overline{\ln X} - \mu_y)^2 \\ +\frac{\alpha[\Psi'(\alpha)^2+\Psi''(\alpha)]}{[1-\alpha\Psi(\alpha)]^3}(\bar{X}-\mu_x)(\overline{\ln X} - \mu_y) + ...\end{array}$$

where now $\mu_y = \Psi(\alpha) + \ln\beta$. With the help of

$$\begin{aligned}\mathbb{E}\left[(\ln X - \mu_y)^2\right] &= \Psi'(\alpha) \\ \mathbb{E}\left[(X-\mu_x)(\ln X - \mu_y)\right] &= \beta\end{aligned}$$

the expected values are now equal to

$$\begin{array}{l}\alpha + \frac{1+\alpha+\alpha^2\Psi(2,\alpha)}{2n\alpha[\alpha\Psi(1,\alpha)-1]} + ... \\ \beta\left(1 - \frac{1-\alpha+\alpha^2\Psi(2,\alpha)}{2n\alpha^2[\alpha\Psi(1,\alpha)-1]}\right) + ...\end{array}$$

the variances are

$$\begin{array}{l}\frac{\alpha}{n[\alpha\Psi(1,\alpha)-1]} + ... \\ \frac{\beta^2\Psi(1,\alpha)}{n[\alpha\Psi(1,\alpha)-1]} + ...\end{array}$$

and the correlation coefficient is

$$-\sqrt{\frac{1}{\alpha\Psi(1,\alpha)}} + ...$$