MIT OpenCourseWare
http://ocw.mit.edu

1.010 Uncertainty in Engineering
Fall 2008

For information about citing these materials or our Terms of Use, visit: http://ocw.mit.edu/terms.

Application Example 17

Sums of iid Random Variables

Central Limit Theorem

The Central Limit Theorem establishes that, when properly normalized, the sum of n independent and identically distributed (iid) random variables X_i ($i = 1, 2, \ldots, n$) with finite variance approaches the standard normal distribution as $n \to \infty$. The standard normal distribution has zero mean value and unit variance and probability density function (PDF) given by:

$$f_Z(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}, \quad -\infty < \mathrm{z} < \infty \tag{1}$$

The Case with Finite *n*

In practice, n is finite and the question arises whether n is large enough to assume that the sum has normal distribution. The rate at which the probability density function of the normalized sum

$$Z_n = \frac{\left(\sum_{i=1}^{n} X_i\right) - nm_X}{\sqrt{n}\sigma_X} \tag{2}$$

approaches the function in equation (1) depends on the distribution of X and the region of the distribution one considers. Convergence is usually fast near the mean value but may be slow in the tail regions, especially when the distribution of X is highly skewed.

To illustrate convergence, we consider the cases when X has uniform or exponential distribution.

1. Uniform Distribution

Suppose that X has uniform distribution in the interval [-0.5, 0.5]. In this case the mean value and variance of X are $m_X = 0$ and $(\sigma_X)^2 = 1/12$, and the normalized sum Z_n in equation (2) is:

$$Z_n = \sqrt{12}\frac{\left(\sum_{i=1}^{n} X_i\right)}{\sqrt{n}} \tag{3}$$

Since $\sum_{i=1}^{n} X_i$ ranges from –0.5n to 0.5n, the PDF of Z_n is 0 outside the interval $[-\sqrt{3n}, \sqrt{3n}]$.

For X uniformly distributed, the distribution of Z_n in equation (3) does not have known analytical form but can be obtained numerically, as follows. Consider first the case n = 2. The PDF of $Y_2 = X_1+X_2$ is given by

$$f_{Y_2}(y)=\int_{-\infty}^{\infty} f_X(x)\, f_X(y-x)dx \tag{4}$$

Integrals of the type in equation (4) are called convolution integrals. A convenient way to evaluate such integrals is to work with the characteristic function of X, $\varphi_X(t)$, which is defined as the Fourier transform of f_X:

$$\varphi_X(t) = \int_{-\infty}^{\infty} f_X(x)e^{-ixt}\, dx \tag{5}$$

with inverse transform

$$f_X(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \varphi_X(t)e^{ixt}\, dt \tag{6}$$

One can show that the characteristic function of Y_2 in equation (4) is simply the square of the characteristic function of X. More in general, the characteristic function of $Y_n = \sum_{i=1}^{n} X_i$ is the n^{th} power of the characteristic function of X:

$$\varphi_{Y_n}(t) = [\varphi_X(t)]^n \tag{7}$$

Hence, a simple way to obtain the distribution of Z_n is to: (a) find $\varphi_X(t)$, either analytically or numerically, (b) raise $\varphi_X(t)$ to the n^{th} power, (c) take the inverse Fourier transform to obtain f_{Y_n}, and (d) find $f_{Z_n}(z)$ through:

$$f_{Z_n}(z) = \sqrt{n}\sigma_X f_{Y_n}\left(\sqrt{n}\sigma_X\, z + nm_X\right) \tag{8}$$

We have used this method to obtain the PDF of Z_n in equation (3) for different n. Figure 1 compares f_{Z_n} with the standard normal PDF in equation (1), for n = 2, 5, 10, and 20. For each value of n, the results for f_{Z_n} are shown in both arithmetic scale (left column) and log scale (right column). From the arithmetic plots, it appears that convergence to the normal density is achieved already for n = 5 and definitely for n = 10. However, this is true only up to about 3 or 4 standard deviations away from the mean. As the log plots show, beyond this central region the distributions of Z and Z_n are still different and one needs larger values of n for convergence. Notice that these extreme tail regions are associated with very small values of the PDF, and hence very small exceedance probabilities.

2. Exponential Distribution

Suppose now that X has exponential distribution with PDF

$$f_X(x) = e^{-x}\ ,\ \ x \geq 0 \tag{9}$$

In this case $m_X = (\sigma_X)^2 = 1$ and, using equation (2), Z_n is given by

$$Z_n = \frac{\left(\sum_{i=1}^{n} X_i\right) - n}{\sqrt{n}} \tag{10}$$

The PDF of Z_n can be found analytically, as follows. We know that the sum $Y_n = \sum_{i=1}^{n} X_i$ has gamma distribution with density

$$f_{Y_n}(y) = \frac{y^{n\,-1}\,e^{-y}}{(n\text{-}1)!}\ ,\ \ y \geq 0 \tag{11}$$

Then, using results for distributions of linear functions, the PDF of Z_n in equation (10) is

$$f_{Z_n}(z) = \frac{\sqrt{n}}{(n\text{-}1)!}\,(n+\sqrt{n}z)^{n\,-1}\ \mathrm{e}^{-(n+\sqrt{n}z)}\ ,\ \ z \geq -\sqrt{n} \tag{12}$$

Figure 2 compares the density in equation (12) with the density in equation (1) for n = 2, 10, 20, 130. Convergence to the normal distribution is slower than in the case of the uniform distribution. The reason is that the exponential distribution is highly skewed and one needs a large *n* to "undo" that skewness. This is especially clear from the log plots.

Problem 17.1

(a) Notice that, if X has uniform distribution, Z_2 *has symmetrical triangular distribution. Using the results in Figure 1, discuss how large n should be for convergence to the normal distribution in the range [-3, 3] if X has symmetrical triangular distribution.*

(b) Computers usually simulate normally distributed variables as the sum of 12 uniformly distributed variables. See what equation (3) becomes for n = 12. Suppose now that you need to accurately reproduce the tail of the normal distribution up to about 5 standard deviations from the mean. Using Figure 1, estimate a reasonable value of n that would give you the desired accuracy. How could you use the normal variable simulator of the computer (which uses n = 12) to simulate normal variables that meet your higher-accuracy criterion?

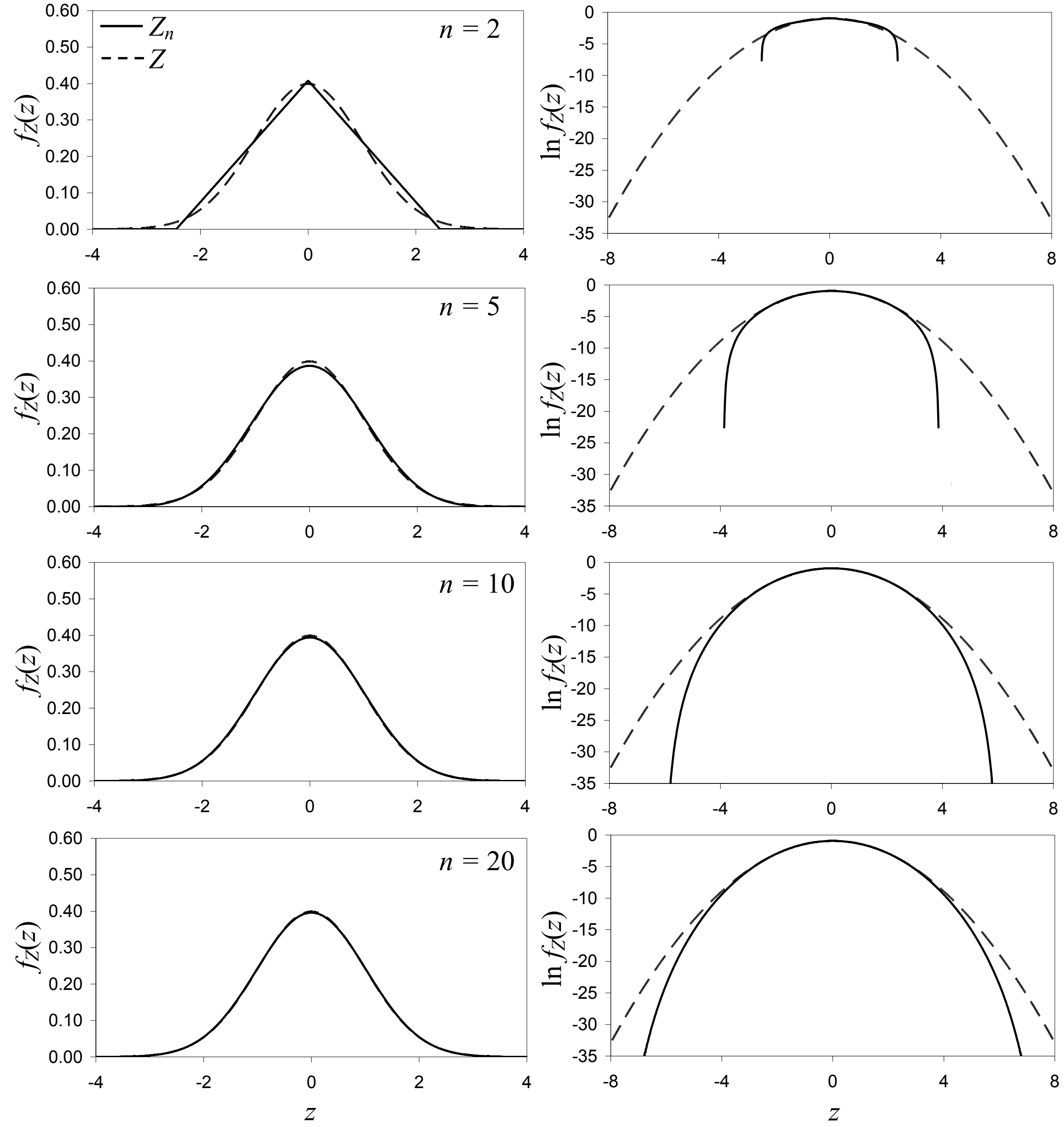

Figure 1: Comparison of the distributions of Z_n (solid lines) and Z (dashed lines) for different n, for the case when X has uniform distribution. The plots on the left are in arithmetic scale, whereas the plots on the right are in log scale and emphasize the tail regions of the distributions.

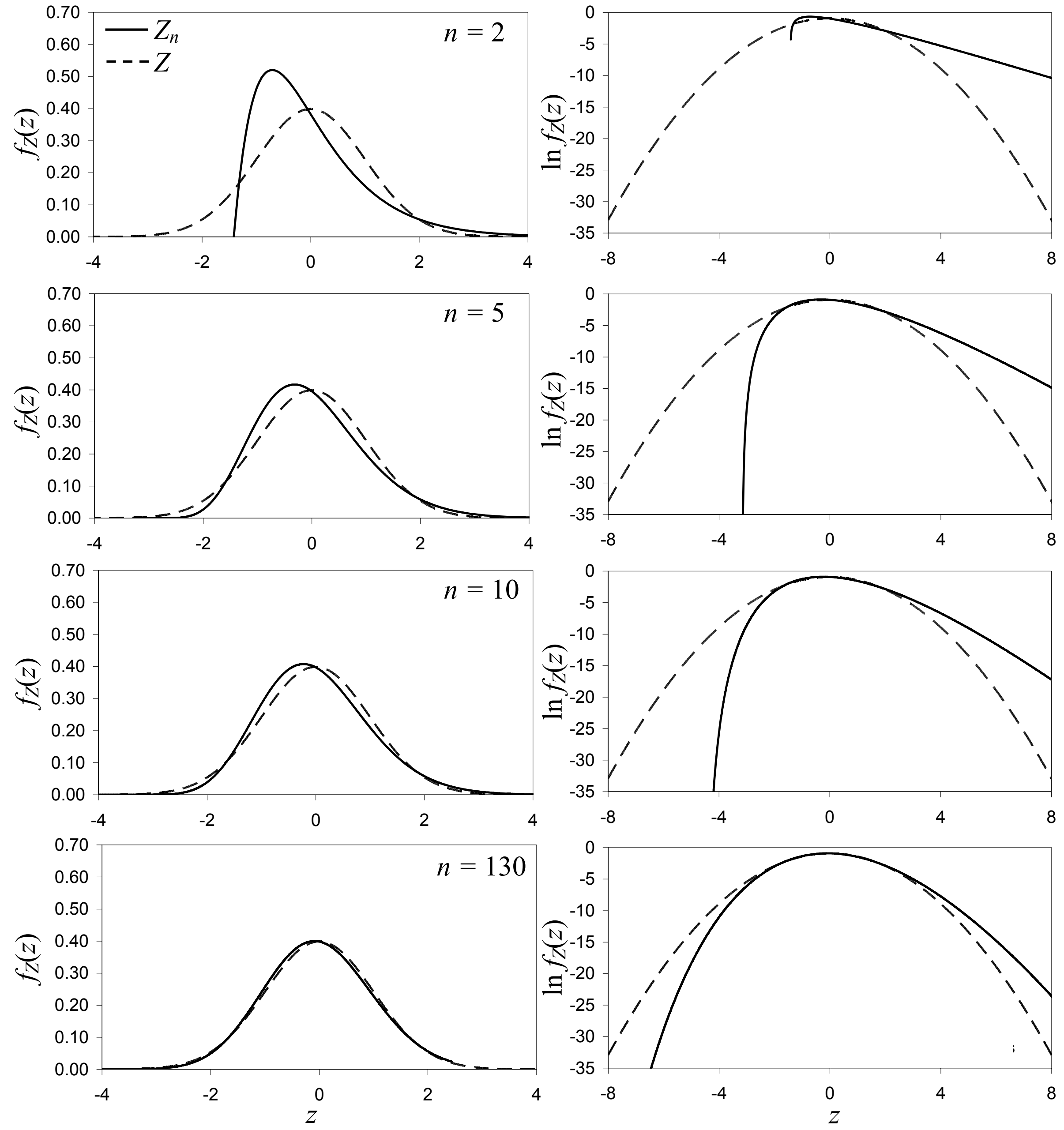

Figure 2: Comparison of the distributions of Z_n (solid lines) and Z (dashed lines) for different n, for the case when X has exponential distribution. The plots on the left are in arithmetic scale, whereas the plots on the right are in log scale and emphasize the tail regions of the distributions.