Central Limit Theorem: Discrete Independent Trials

May 15, 2006

Central Limit Theorem for Discrete Independent Trials

- Let $S_n = X_1 + X_2 + \cdots + X_n$ be the sum of n independent discrete random variables of an independent trials process with common distribution function $m(x)$ defined on the integers, with mean μ and variance σ^2.

- **Standardized Sums**

$$S_n^* = \frac{S_n - n\mu}{\sqrt{n\sigma^2}} \ .$$

- This standardizes S_n to have expected value 0 and variance 1.

- If $S_n = j$, then S_n^* has the value x_j with

$$x_j = \frac{j - n\mu}{\sqrt{n\sigma^2}} \ .$$

Approximation Theorem

Theorem. *Let X_1, X_2, ..., X_n be an independent trials process and let $S_n = X_1 + X_2 + \cdots + X_n$. Assume that the greatest common divisor of the differences of all the values that the X_j can take on is 1. Let $E(X_j) = \mu$ and $V(X_j) = \sigma^2$. Then for n large,*

$$P(S_n = j) \sim \frac{\phi(x_j)}{\sqrt{n\sigma^2}} \ ,$$

where $x_j = (j - n\mu)/\sqrt{n\sigma^2}$, and $\phi(x)$ is the standard normal density.

Central Limit Theorem for a Discrete Independent Trials Process

Theorem. *Let* $S_n = X_1 + X_2 + \cdots + X_n$ *be the sum of* n *discrete independent random variables with common distribution having expected value* μ *and variance* σ^2*. Then, for* $a < b$*,*

$$\lim_{n\to\infty} P\left(a < \frac{S_n - n\mu}{\sqrt{n\sigma^2}} < b\right) = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-x^2/2}\, dx\ .$$

Example

A die is rolled 420 times. What is the probability that the sum of the rolls lies between 1400 and 1550?

Example

A die is rolled 420 times. What is the probability that the sum of the rolls lies between 1400 and 1550?

The sum is a random variable

$$S_{420} = X_1 + X_2 + \cdots + X_{420} \ .$$

We have seen that $\mu = E(X) = 7/2$ and $\sigma^2 = V(X) = 35/12$.

Thus, $E(S_{420}) = 420 \cdot 7/2 = 1470$, $\sigma^2(S_{420}) = 420 \cdot 35/12 = 1225$, and $\sigma(S_{420}) = 35$.

$$\begin{aligned} P(1400 \le S_{420} \le 1550) &\approx P\left(\frac{1399.5 - 1470}{35} \le S_{420}^* \le \frac{1550.5 - 1470}{35}\right) \\ &= P(-2.01 \le S_{420}^* \le 2.30) \\ &\approx NA(-2.01, 2.30) = .9670\ . \end{aligned}$$

A More General Central Limit Theorem

Theorem. *Let* $X_1, \; X_2, \; \ldots, \; X_n \;, \; \ldots$ *be a sequence of independent discrete random variables, and let* $S_n = X_1 + X_2 + \cdots + X_n$*. For each* n*, denote the mean and variance of* X_n *by* μ_n *and* σ_n^2*, respectively. Define the mean and variance of* S_n *to be* m_n *and* s_n^2*, respectively, and assume that* $s_n \to \infty$*. If there exists a constant* A*, such that* $|X_n| \le A$ *for all* n*, then for* $a < b$*,*

$$\lim_{n\to\infty} P\left(a < \frac{S_n - m_n}{s_n} < b\right) = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-x^2/2}\, dx \; .$$

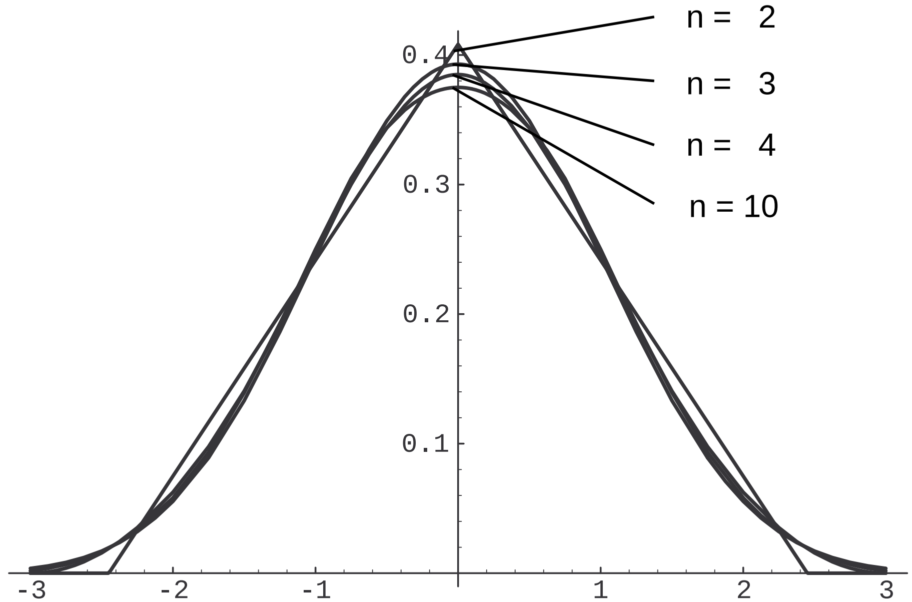
Central Limit Theorem for Continuous Independent Trials

Standardized Sums

- Suppose we choose n random numbers from the interval $[0, 1]$ with uniform density. Let X_1, X_2, . . . , X_n denote these choices, and $S_n = X_1 + X_2 + \cdots + X_n$ their sum.

- The standardized sum is

$$S_n^* = \frac{S_n - n\mu}{\sqrt{n}\sigma} .$$


n = 2
n = 3
n = 4
n = 10
0.4
0.3
0.2
0.1
-3
-2
-1
1
2
3

Exponential Density

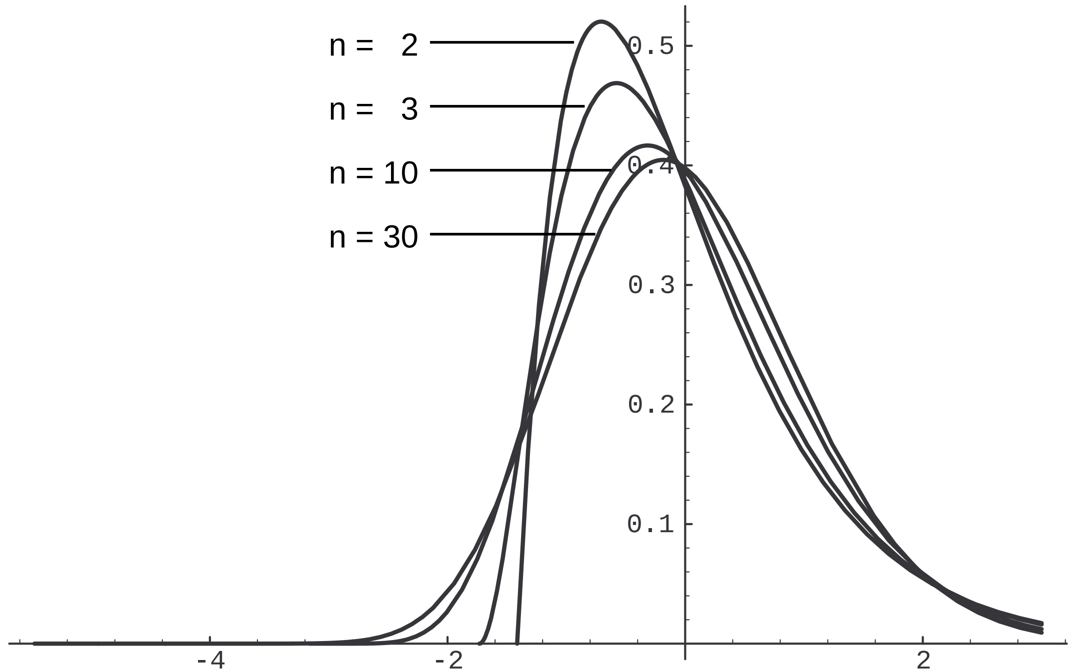
- Choose numbers from the interval $[0, +\infty)$ with an exponential density with parameter λ.

- Then

$$\begin{aligned} \mu &= E(X_i) = \frac{1}{\lambda}\ , \\ \sigma^2 &= V(X_j) = \frac{1}{\lambda^2}\ . \end{aligned}$$

- There are formulas for the density function for S_n and the density function for S_n^*:

$$
\begin{aligned}
f_{S_n}(x) &= \frac{\lambda e^{-\lambda x}(\lambda x)^{n-1}}{(n-1)!}\,, \\
f_{S_n^*}(x) &= \frac{\sqrt{n}}{\lambda} f_{S_n}\left(\frac{\sqrt{n}x+n}{\lambda}\right)\,.
\end{aligned}
$$


n = 2
n = 3
n = 10
n = 30
0.5
0.4
0.3
0.2
0.1
-4
-2
2

Central Limit Theorem

Theorem. *Let* $S_n = X_1 + X_2 + \cdots + X_n$ *be the sum of* n *independent continuous random variables with common density function* p *having expected value* μ *and variance* σ^2. *Let* $S_n^* = (S_n - n\mu)/\sqrt{n}\sigma$. *Then we have, for all* $a < b$,

$$\lim_{n\to\infty} P(a < S_n^* < b) = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-x^2/2}\, dx \ .$$

Example

- Suppose a surveyor wants to measure a known distance, say of 1 mile, using a transit and some method of triangulation.

- He knows that because of possible motion of the transit, atmospheric distortions, and human error, any one measurement is apt to be slightly in error.

- He plans to make several measurements and take an average.

- He assumes that his measurements are independent random variables with a common distribution of mean $\mu = 1$ and standard deviation $\sigma = .0002$.

- What can he say about the average?

Estimating the Mean

- Now suppose our surveyor is measuring an unknown distance with the same instruments under the same conditions.
- He takes 36 measurements and averages them.
- How sure can he be that his measurement lies within .0002 of the true value?

Sample Mean

- The *sample mean* of n measurements:

$$\bar{\mu} = \frac{x_1 + x_2 + \cdots + x_n}{n} \ ,$$

- Moreover

$$P(|\bar{\mu} - \mu| < .0002) \approx .997 \ .$$

- The interval $(\bar{\mu} - .0002, \bar{\mu} + .0002)$ is called the 99.7% confidence interval.

Sample Variance

- If he does not know the variance σ^2 of the error distribution, then he can estimate σ^2 by the sample variance:

$$\bar{\sigma}^2 = \frac{(x_1 - \bar{\mu})^2 + (x_2 - \bar{\mu})^2 + \cdots + (x_n - \bar{\mu})^2}{n} .$$