Topic 11

Central Limit Theorem

The Classical Central Limit Theorem

Outline

Motivation

For the law of large numbers, the sample means from a sequence of independent random variables converge to their common distributional mean as the number n of random variables increases.

$$\frac{1}{n}S_n = \bar{X}_n \to \mu \text{ as } n \to \infty.$$

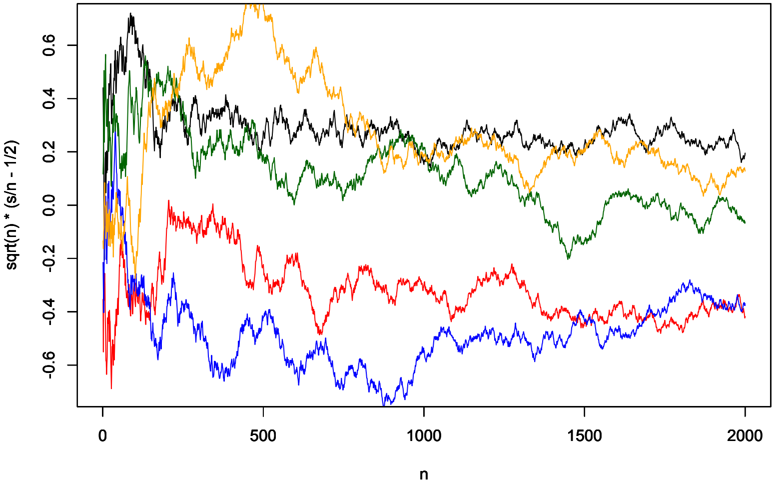
Moreover, the standard deviation of $\bar{X}_n$ is inversely proportional to $\sqrt{n}$. For example, for independent random variables, uniformly distributed on $[0,1]$, $\bar{X}_n$ converges to

$$\mu = \int_0^1 x f_X(x)\,dx = \int_0^1 x\,dx = \frac{x^2}{2}\Big|_0^1 = \frac{1}{2}$$

Because the standard deviation $\sigma_{\bar{X}_n} \propto 1/\sqrt{n}$, we magnify the difference between the running average and the mean by a factor of $\sqrt{n}$ and investigate the graph of

$$\sqrt{n}\left(\frac{1}{n}S_n - \mu\right) \text{ versus } n$$

Motivation



Motivation

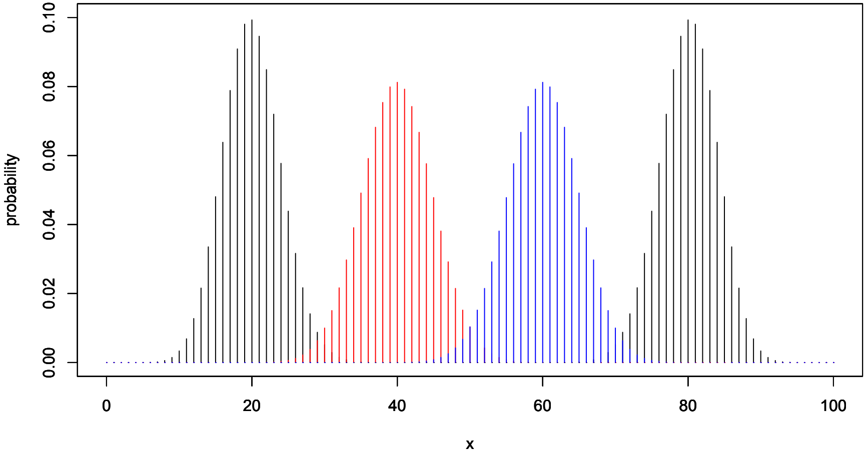
Does the distribution of the size of these fluctuations have any regular and predictable structure? Let's begin by examining the distribution for the sum of $X_1, X_2 \ldots X_n$, independent and identically distributed random variables

$$S_n = X_1 + X_2 + \cdots + X_n.$$

What distribution do we see? We begin with the simplest case, X_i Bernoulli random variables. The sum S_n is a binomial random variable. We examine two cases.

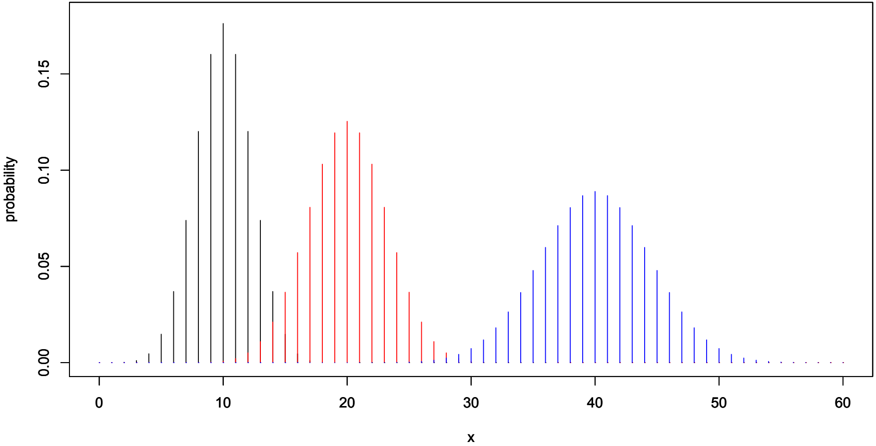
- keep the number of trials the same at $n = 100$ and vary the success probability p.
- keep the success probability the same at $p = 1/2$, but vary the number of trials.

Bernoulli Random Variables



Successes in 100 Bernoulli trials with $p = 0.2, 0.4, 0.6$ and 0.8.

Bernoulli Random Variables



Successes in 20, 40, and 80 Bernoulli trials with $p = 0.5$.

Bernoulli Random Variables

The binomial random variable S_n has

mean np and standard deviation $\sqrt{np(1-p)}$.

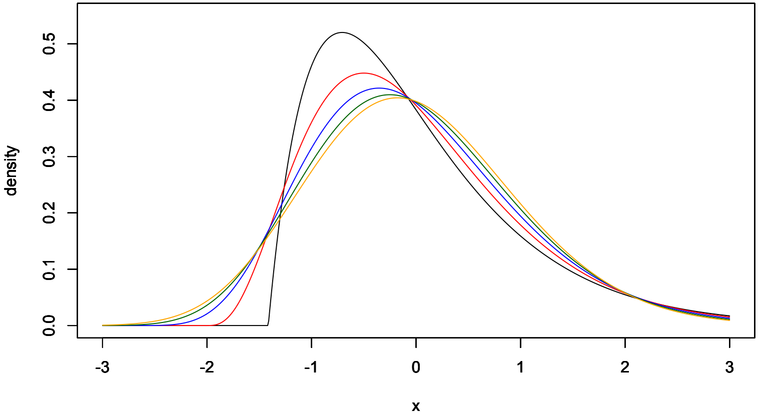
Thus, if we take the standadized version of these sums of Bernoulli random variables

$$Z_n = \frac{S_n - np}{\sqrt{np(1-p)}},$$

then these bell curve graphs would lie on top of each other.

Now, let's consider *exponential* random variables . . . .

Exponential Random Variables



The density of the standardized random variables that result from the sum of 2,4,8,16, and 32 exponential random variables

The Classical Central Limit Theorem

To obtain the standardized random variables,

- we can either standardize using the sum S_n having mean $n\mu$ and standard deviation $\sigma\sqrt{n}$, , *or*
- we can standardize using the sample mean $\bar{X}_n$ having mean μ and standard deviation $\sigma/\sqrt{n}$.

This yields two equivalent versions of the standardized score or z-score.

$$Z_n = \frac{S_n - n\mu}{\sigma\sqrt{n}} = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} = \frac{\sqrt{n}}{\sigma}(\bar{X}_n - \mu).$$

The theoretical result behind these numerical explorations is called the classical central limit theorem.

The Classical Central Limit Theorem

Theorem. Let $\{X_i; i \geq 1\}$ be independent random variables having a common distribution. Let μ be their mean and σ^2 be their variance. Then Z_n, the standardized scores, converges in distribution to Z a standard normal random variable, i.e., the distribution function F_{Z_n} converges to Φ, the distribution function of the standard normal for every value z.

$$\lim_{n\to\infty} F_{Z_n}(z) = \lim_{n\to\infty} P\{Z_n \leq z\} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-x^2/2}\, dx = \Phi(z).$$

In practical terms the central limit theorem states that

$$P\{a < Z_n \leq b\} \approx P\{a < Z \leq b\} = \Phi(b) - \Phi(a).$$

The number value is obtained in R using the command `pnorm(b)-pnorm(a)`.

Uniform Random Variables

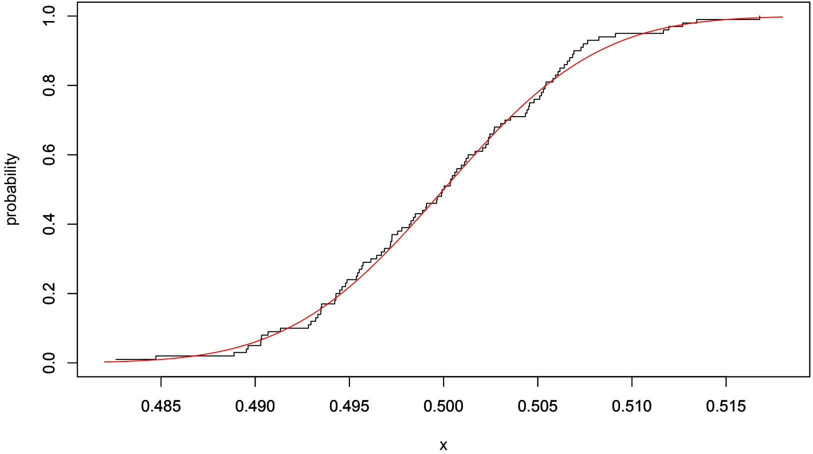
Example.

- For a single $U(0,1)$ random variables,
  - mean $\mu = 1/2$ and standard deviation $\sigma = 1/\sqrt{12}$.
- For $\bar{X}$ the sample mean of 2000 independent of $U(0,1)$ random variables, then $\bar{X}$
  - has mean $\mu = 1/2$ and standard deviation $\sigma = 1/\sqrt{24000}$.

We show the empirical cumulative distribution function for 100 simulations and compare it to the distribution function of a normal with mean $\mu = 1/2$ and standard deviation $\sigma = 1/\sqrt{24000}$.

Exercise. Show that the standard deviation of a $U(0,1)$ random variable is $1/\sqrt{12}$.

Uniform Random Variables



Bernoulli Trials

For a 100 question multiple choice exam with 4 options per question, a student randomly guesses. Each guess is a Bernoulli trial with success probability $p = 1/4$. Thus, the number of correct answers S_{100} has a binomial distribution with

$$\text{mean } np = 100 \cdot \frac{1}{4} = 25 \text{ and standard deviation } \sqrt{np(1-p)} = \sqrt{100 \cdot \frac{1}{4} \cdot \frac{3}{4}} = \frac{5}{2}\sqrt{3} \approx \frac{13}{3}$$

A student has 7 correct answers. This has a z-score

$$z \approx \frac{7-25}{13/3} = \frac{54}{13} < -4$$

Did this student *try* to give incorrect answers?

Exercise. Find the exact z-score and use `pnorm` to estimate the probability of 7 or fewer correct answers. Compare this value to the value obtained using `pbinom`.

Exponential Random Variables

Times between of customer arrivals at a bank are modeled as independent $Exp(1)$ random variables. These random variables have mean and standard deviation 1. We approximate the probability that the 50-th customer arrives within the first hour of business. S_n, the time of arrival of the n-th customer, is the sum of the times between arrivals and thus is the sum of n $Exp(1)$ random variables. S_{50} has mean 50 and standard deviation $\sqrt{50}$. We are asking

$$P\{S_{50} \leq 60\} = P\{S_{50} - 50 \leq 10\} = P\left\{Z_n = \frac{S_{50} - 50}{\sqrt{50}} \leq \frac{10}{\sqrt{50}}\right\}.$$

By the central limit theorem, we have the approximation

```
> pnorm((60-50)/sqrt(50))
[1] 0.9213504
```

We can obtain the same answer using `pnorm(60,50,sqrt(50))`.

Example

You want to store 400 pictures on your smart phone. Pictures have a mean size of 450 kilobytes (KB) and a standard deviation of 50 KB. Assume that the size of the pictures are independent. S_{400}, the total storage space needed for the 400 pictures, has

$$\text{mean } 400 \times 450 = 180,000 \text{ KB and standard deviation } 50\sqrt{400} = 1000 \text{ KB.}$$

To estimate the space required to be 99% certain that the pictures will have storage space on the phone, note that

```
> qnorm(0.99,400*450,50*sqrt(400))
[1] 182326.3
```

So we need about 182.3 megabytes (MB).

Exercise. Give the storage space to be 95% certain to have the space for 300 pictures.