DATA 1010
In-class exercises
Samuel S. Watson
26 October 2018

Problem 1

The central limit theorem says that if S_n is a sum of i.i.d. finite-variance random variables is approximately normally distributed with mean $\mathbb{E}[S_n]$ and variance $\mathrm{Var}(S_n)$. Also, about 95% of the probability mass of a normal distribution is within two standard deviations of the mean.

If a million independent Unif$([a, b])$'s are added, what is the shortest interval containing 95% of the probability mass of the distribution of the resulting sum?

Solution

The mean of the million random variables is 5 million, and the variance is a million times the variance of each random variable, which is $(b-a)^2/12 = 10^2/12$ for Unif$([a, b])$. Therefore, the smallest interval capturing 95% of the probability mass is approximately

$$[5{,}000{,}000 - 2\sqrt{10^2/12 \cdot 10^6}, 5{,}000{,}000 + 2\sqrt{10^2/12 \cdot 10^6},] \approx [4994226, 5005774].$$

So typical fluctuations are on the order of a few thousand, which is small compared to the total of 5 million.

Problem 2

The multivariate central limit theorem says that if $\mathbf{X}_1, \mathbf{X}_2, \ldots$ is an independent sequence of random vectors with a common distribution on $\mathbb{R}^n$, then the standardized mean

$$\mathbf{S}_n^* = \frac{X_n - n\mu}{\sqrt{n}}$$

converges in distribution to $\mathcal{N}(\mathbf{0}, \Sigma)$, where Σ is the covariance matrix of $\mathbf{X}_1$.

Investigate the multivariate central limit theorem using by making 2D histograms for i.i.d. sums of (i) uniform samples from the square, and (ii) samples from $((U+V)/2, V)$, where (U, V) is uniformly sampled from the square.

Solution

We obtain a random vector with correlated components by sampling two independent uniform random variables U and V and returning $[X, Y] = [(U+V)/2, V]$. Then

$$\mathrm{Cov}((U+V)/2, V) = \frac{1}{2}(\mathrm{Cov}(U, V) + \mathrm{Cov}(V, V)) = \frac{1}{2}\mathrm{Var}(V) = \frac{1}{24},$$

since $\mathrm{Var}(V) = (b-a)^2/12 = (1-0)^2/12 = 1/12$. So we can see X and Y are correlated. We calculate a running average and plot a 2D histogram as follows:

```
using Plots
function sample()
    U = rand()
    V = rand()
    X = (U + V)/2
    Y = V
    [X,Y]
end

function runningaverage(n)
    X_sum, Y_sum = sum(sample() for i=1:n)
```

```
    μ, ν = 1/2,1/2
    X_stand, Y_stand = [μ,ν] + sqrt(n)*[X_sum/n-μ,Y_sum/n-ν]
    (X_stand,Y_stand)
end

n = 100
numsamples = 10000
histogram2d([runningaverage(n) for i=1:numsamples],
            bins=(60,60),
            aspect_ratio=:equal,
            xlims=(-1,2),
            ylims=(-1,2))
```

Problem 3

Find the mean and covariance of the random vector $[X, Y]$ defined by $X = \frac{1}{2}(U + V), Y = V$, where U and V are independent uniform random variables on $[0, 1]$.

Use the result to find the density of the limiting distribution you plotted in the previous problem.

Solution

We know that $\mathrm{Cov}(X, Y) = \frac{1}{24}$ because we calculated it in the previous problem. Also, the variance of Y is $\frac{1}{12}$, and the variance of X is

$$\mathrm{Var}(X) = \frac{1}{4}(\mathrm{Var}(U) + \mathrm{Var}(V)) = \frac{1}{24}.$$

Therefore, the covariance matrix is

$$\Sigma = \begin{bmatrix} \frac{1}{24} & \frac{1}{24} \\ \frac{1}{24} & \frac{1}{12} \end{bmatrix}.$$

The vector of means is $[\frac{1}{2}, \frac{1}{2}]$.

Therefore, the density of the normalized running sums of independent samples from the distribution of $[X, Y]$ is

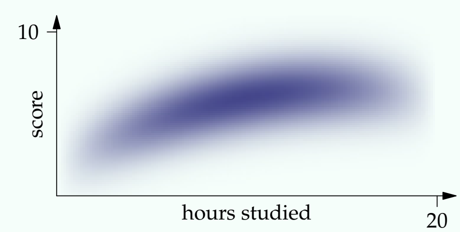
$$\frac{1}{2\pi\sqrt{\det \Sigma}} \exp\left(-\tfrac{1}{2}[x - \tfrac{1}{2}, y - \tfrac{1}{2}]'\Sigma^{-1}[x - \tfrac{1}{2}, y - \tfrac{1}{2}]\right) = \frac{12}{\pi}\mathrm{e}^{-\frac{1}{2}(48x^2+24y^2-48xy-24x+6)}.$$

Problem 4

Find the conditional expectation of Y given X if the joint distribution has density

$$f(x, y) = \frac{3}{4000(3/2)\sqrt{2\pi}} x(20 - x)\mathrm{e}^{-\frac{1}{2(3/2)^2}\left(y-2-\frac{1}{50}x(30-x)\right)^2}.$$

on the strip $[0, 20] \times \mathbb{R}$.



Solution

The restriction of f to a vertical line at position x is proportional to the function

$$y \mapsto \mathrm{e}^{-\frac{1}{2(3/2)^2}\left(y-2-\frac{1}{50}x(30-x)\right)^2}.$$

We recognize this function as proportional to the normal density with mean $\frac{1}{50}x(30 - x)$ and standard deviation $\frac{3}{2}$. Therefore, once this density is suitably normalized, it will be equal to $\mathcal{N}\left(\frac{1}{50}x(30 - x), \frac{3}{2}\right)$. So the desired conditional expectation is

$$\mathbb{E}[Y \mid X] = \frac{1}{50}X(30 - X).$$