Lab 07

Convergence, Central Limits, and Fat Tails

2023-11-03

Table of contents

# 1. Intro

An important idea in statistics is the [Central Limit Theorem (CLT)](https://en.wikipedia.org/wiki/Central_limit_theorem), which states (much more elegantly and precisely) that the sum of many **independent** random variables is approximately normally distributed, even if the underlying process is not randomly distributed (as long as it’s moderately well-behaved). This is often used to justify approximating the average of many data points, and the uncertainty in this average, as a normal distribution.

The CLT is a powerful tool, but intuition we develop about working with distributions that behave “nicely” can be misleading when we are working with extreme value distributions. The purpose of this lab is to explore this idea and to build some intuition about sampling uncertainty and sample statistics when working with extreme value distributions.

## 1.1 Setup

Remember to git clone to your machine, then activate and instantiate the environment, as we have done in previous labs. If you are having trouble, you may want to build IJulia in the Pkg manager.

**Do not use additional packages for this assignment,** though you are welcome to look at the documentation or implementation of other packages for nearest-neighbors methods.

using Distributions  
using Plots  
using StatsBase  
using StatsPlots  
  
Plots.default(; margin=4Plots.mm, size=(700, 400), linewidth=2)

# 2. Normal distribution

To build some intuition, let’s create a Normal distribution with a very large standard deviation relative to its mean:

dist\_normal = Normal(2.5, 10)

We can calculate its mean and standard deviation

mean(dist\_normal), std(dist\_normal)

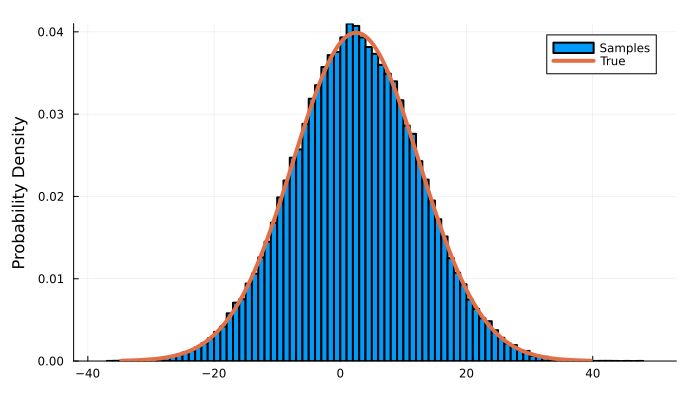
(2.5, 10.0)

Because we’re going to be plotting a lot of distributions, let’s define a function to do this for us:

function plot\_dist(dist; name="", xlims=missing)  
 ub = quantile(dist, 0.998)  
 lb = quantile(dist, 0.002)  
 p = plot(x -> pdf(dist, x); ylabel="Probability Density", label=name, xlims=(lb, ub))  
 !ismissing(xlims) && xlims!(p, xlims)  
 return p  
end  
  
plot\_dist(dist\_normal; name="Normal Distribution")

We can see that even though this distribution has a large standard deviation, it still has that classical “bell” shape. We can also sample from its probability density function and plot a histogram of those samples. That output should look like the probability density function shown above!

samples\_normal = rand(dist\_normal, 100\_000)  
histogram(samples\_normal; normalize=:pdf, label="Samples", ylabel="Probability Density")  
plot!(dist\_normal; label="True", linewidth=4)



We can see that our large sample size leads to us being able to sample well from the mean.

## 2.1 Running statistics

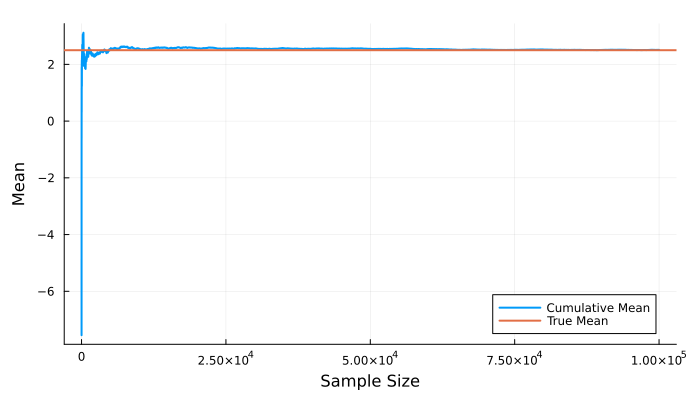
An important idea in statistics is that we can calculate descriptive statistics from our data. For example, things like the mean, standard deviation, skewness, or 99th percentile. An interesting question to ask is how many observations we need to get a reliable estimate of these “sample statistics”.

Let’s start by looking at the mean with our 100,000 samples. We could do this with a loop, but it turns out to be inefficient. Instead, we’ll define some more efficient functions

cumul\_mean(x) = cumsum(x) ./ (1:length(x))  
cumul\_std(x) = sqrt.(cumsum(x .- cum\_mean(x)) .^ 2 ./ (1:length(x)))

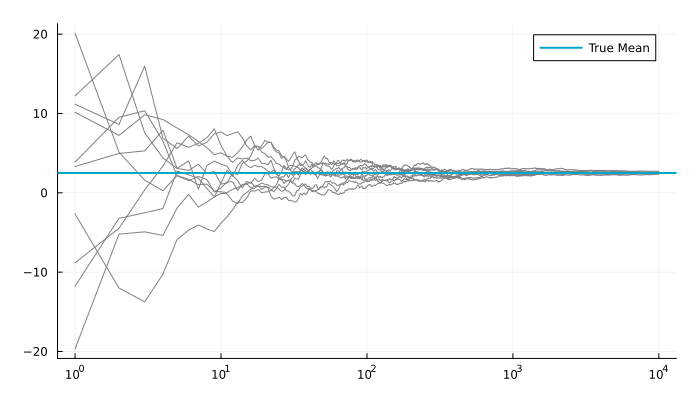
Now let’s plot the cumulative mean of our samples.

plot(  
 cumul\_mean(samples\_normal); label="Cumulative Mean", ylabel="Mean", xlabel="Sample Size"  
)  
hline!([mean(dist\_normal)]; label="True Mean")



We can see that we very rapidly converge to the “true” value. Of course, maybe we just got lucky with our unique sample. We can repeat this experiment many times and see how often we get a good estimate of the mean (note the log scale!)

plot(; xscale=:log10)  
for i in 1:10  
 samples\_normal = rand(dist\_normal, 10\_000)  
 plot!(cumul\_mean(samples\_normal); label=nothing, linewidth=1, color=:gray)  
end  
hline!([mean(dist\_normal)]; label="True Mean")

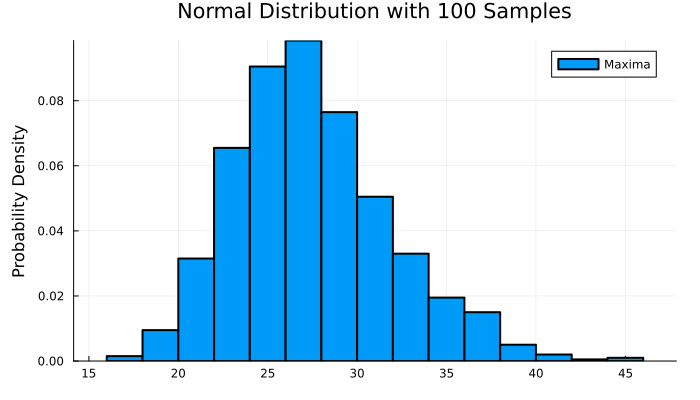


## 2.2 Sample extrema

Another question we can ask is how much extreme values, like the minimum or maximum of our samples, vary from one random realization to the next.

We can answer this question in a very simple experiment. We’ll draw N samples from our distribution, and take the maximum. We’ll repeat this M times, and plot the distribution of the maximums.

N = 100  
M = 1\_000  
maxima = [maximum(rand(dist\_normal, N)) for i in 1:M]  
histogram(  
 maxima;  
 normalize=:pdf,  
 label="Maxima",  
 ylabel="Probability Density",  
 title="Normal Distribution with $N Samples",  
)



We can see from this that even though the Normal distribution is “well-behaved”, the maximum of a sample of 100 observations can vary enormously from one set of 100 draws to the next. (And of course, we are assuming that we know the true distribution!)

# 3. Your turn

Repeat the analysis above for two distributions from the following list

* A distribution (TDist). Make sure you know that the “degrees of freedom” parameter does.
* A GEV (GeneralizedExtremeValue) distribution with a positive shape parameter
* A GEV (GeneralizedExtremeValue) distribution with a negative shape parameter
* A Cauchy (Cauchy) distribution

You may use something else creative if you’re interested. Add plenty of explanatory text to illustrate what you’ve found. Be prepared to put together a very short presentation on your findings.