Transformation of Random Variables

Sujit Sandipan Chaugule[1*], Dr. Amiya Ranjan Bhowmick[2]

[1*]Department of Pharmaceutical Sciences and Technology, Institute of Chemical Technology, Mumbai

[2]Department of Mathematics, Institute of Chemical Technology, Mumbai

Introduction

Suppose that we have a random variable X with the probability density function $f_X(x)$ and we are interested in obtaining the PDF of $Y = g(X)$, which is a transformation of the random variable X. To start this concept, we start with an illustrative example. Suppose that $X \sim \mathcal{N}(0,1)$ and we aim to find the probability distribution of $Y = X^2$. We can think of Y as the squared distance of the random variable X from 0. We define the support of a random variable X as the set of all points on the real line for which the PDF $f_X(x)$ is positive and denoted by the symbol $\mathcal{X}$. Therefore,

$$\mathcal{X} = \{x \in \mathbb{R} : f_X(x) > 0\}.$$

Y is a transformation of X and we define the sample space of Y as

$$\mathcal{Y} = \{y \in \mathbb{R} : f_Y(y) > 0\}.$$

In this context, we have $g(x) = x^2$, $\mathcal{X} = (-\infty, \infty)$ and $\mathcal{Y} = (0, \infty)$ .Before going into the mathematical computations, let us do some simulation and try to see whether the distributions show some commonly known patterns. From an algorithmic point of view, we do the following steps:

- Fix m
- Simulate $X_1, X_2, \ldots, X_m \sim \mathcal{N}(0,1)$
- Compute $Y_1 = X_1^2, \ldots, Y_m = X_m^2$
- Draw the histogram using the values $Y_1, Y_2, \ldots, Y_m$.

In [1]:
```julia
using Plots, Distributions, Statistics, StatsBase
```

In [2]:
```julia
n = 1000
x = rand(Normal(0,1), n)
p1 = histogram(x, normalize = true, xlabel = "x", ylabel = "density",
     label = "",title  = "Histogram of x")
y = x.^2
p2 = histogram(y, normalize = true, xlabel = "y", ylabel = "density",
     label = "",title  = "Histogram of y")
plot(p1, p2, layout = (1,2), size = (800,400))
```

Out[2]:

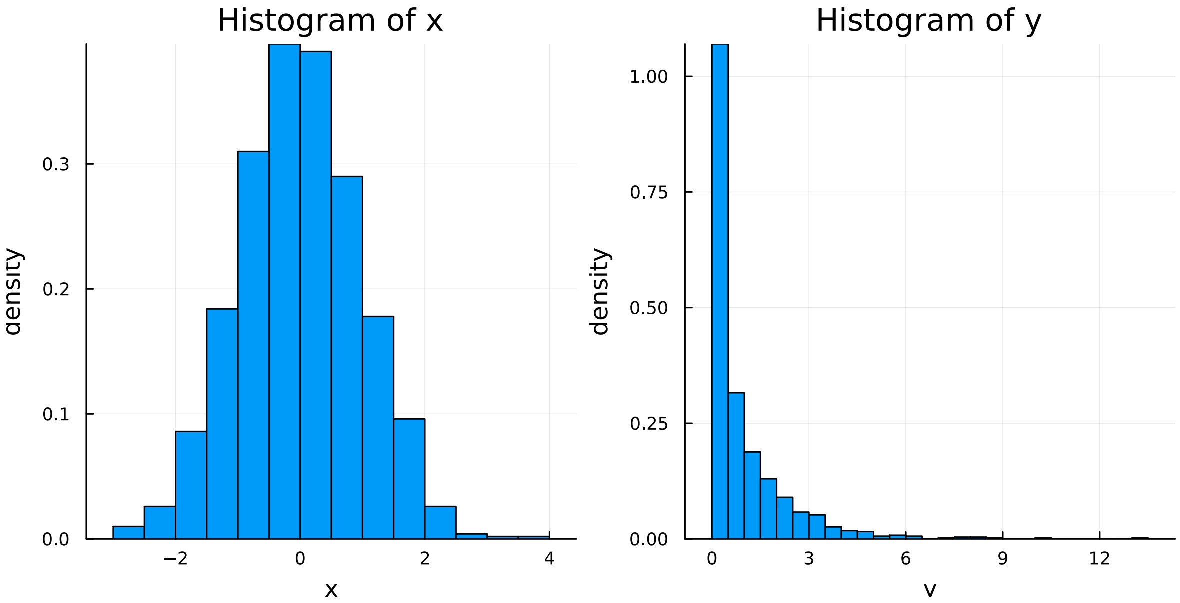

Figure 1:The histogram of the simulated realizations from the standard normal distribution is shown in the left panel. The simulated values of X are squared to obtain the realizations of Y.The distribution of realizations from the distribution of $Y$is shwon in the right panel.

It can be observed that the simulated realizations from the distribution of Y is highly positively skewed. Let us try to obtain the PDF of Y by explicit computation. We take the following strategy:

- Compute the CDF of Y, $F_Y(y)$, $y \in \mathcal{Y}$.
- Take the derivative of $F_Y(y)$ to obtain the PDF $f_Y(y)$.

$$\begin{aligned} F_Y(y) &= P(Y \leq y) = P(X^2 \leq y) \\ &= P(-\sqrt{y} \leq X \leq \sqrt{y}) \\ &= 2P(0 \leq X \leq \sqrt{y}) \quad \text{(even function)} \\ &= 2 \times \int_0^{\sqrt{y}} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx \end{aligned}$$

It is worthwhile to remember the Leibnitz rule for differentiation under the integral sign:

$$\frac{d}{dy} \int_{a(y)}^{b(y)} \psi(x,y)dx = \psi(b(y),y)b'(y) - \psi(a(y),y)a'(y) + \int_{a(y)}^{b(y)} \frac{\partial}{\partial y} \psi(x,y)dx,$$

provided the required mathematical requirements on the function $\psi(x,y)$ is satisfied.

The PDF of Y is given by

$$f_Y(y) = \frac{e^{-\frac{y}{2}} y^{\frac{1}{2}-1}}{\Gamma\left(\frac{1}{2}\right) 2^{\frac{1}{2}}}, \quad 0 < y < \infty,$$

and zero otherwise. The important observation is that the square of the standard normal distribution belongs to the $\mathcal{G}(\alpha, \beta)$ family of distributions. Let us expand this problem in a two-dimensional scenario. During the lecture, we posed the following question: Suppose that we consider the two-dimensional plane and call it the (x_1, x_2) plane. We have $X_1 \sim \mathcal{N}(0,1)$ and $X_2 \sim \mathcal{N}(0,1)$, and we are interested in computing the probability that

$$P(X_1^2 + X_2^2 \leq 1).$$

Let us understand this probability statement in simple terms: Suppose we randomly choose a point on the (x_1, x_2) plane, where each coordinate is chosen randomly and independently from the $\mathcal{N}(0,1)$ distribution. What is the probability that the selected point will fall within the circle of unit radius? There are two strategies to compute this probability.
The first idea is more intuitive rather than mathematically rigorous. We consider the following steps to be performed using Julia:

- Fix m.
- Randomly select $X_1 \sim \mathcal{N}(0,1)$.
- Randomly select $X_2 \sim \mathcal{N}(0,1)$.
- Plot the point on the (x_1, x_2) plane and compute the squared distance $Y = X_1^2 + X_2^2$.
- If $Y \leq 1$, then set counter = 1; otherwise, set counter = 0.
- Repeat these steps m times and compute $\frac{\text{counter}}{m}$, which will be approximately equal to the probability.

In [3]:

```
using Plots, Distributions, Statistics, StatsBase, LaTeXStrings
```

In [4]:

```
m = 1000
x1 = rand(Normal(0,1), m)
x2 = rand(Normal(0,1), m)
p1 = scatter(x1, x2, color="grey", markersize= 3, xlabel=L"x_1",
        ylabel=L"x_2", label="")
hline!([0], color = "red", lw = 2, label = "" )
vline!([0], color = "red", lw = 2, label = "" )
y = (x1.^2).+(x2.^2)
plot!(x -> sqrt(1 - x^2), -1, 1, color="blue", lw = 3 ,
    linestyle=:dash, label="")
plot!(x -> -sqrt(1 - x^2), -1, 1, color="blue", lw = 3 ,
    linestyle=:dash, label="")
p2 = histogram(y, normalize = true, bins = 30, xlabel = "y",
    ylabel = "density", title = "histogram of y", label = "")
plot(p1, p2, layout = (1,2), size = (800,400))
```

Out[4]:

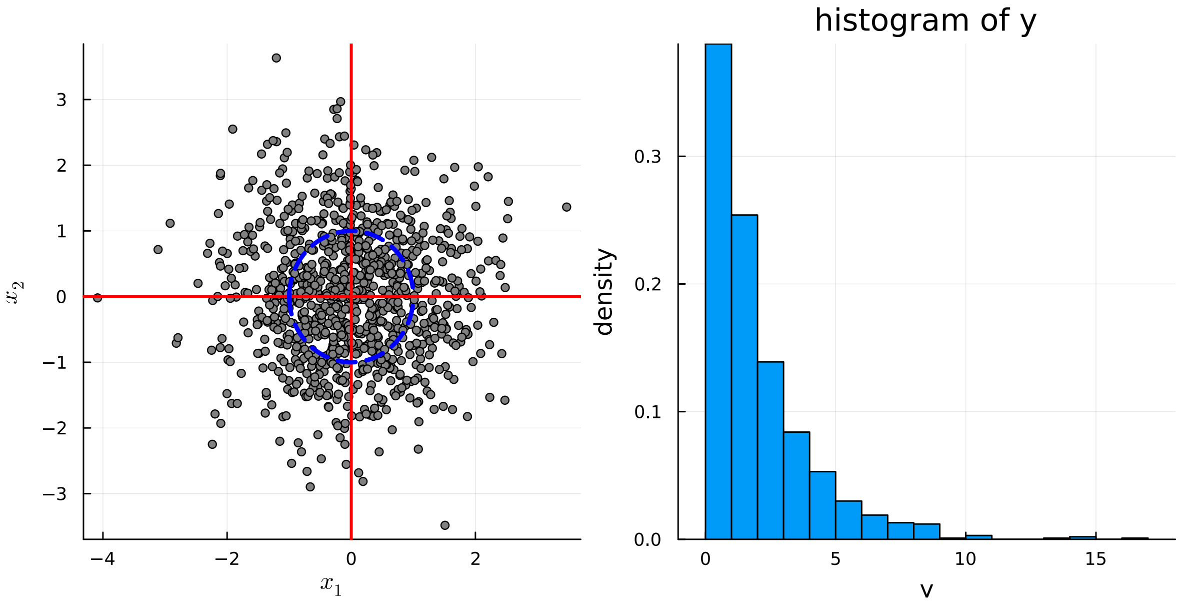

Figure 2: The histogram of the simulated realizations of Y based on 1000 simuations of pairsof independent standard normal random variables. The proportion of the number of points falling in inside the unit circle gives an approximation of the probability $P(Y \leq 1)$.

In [5]:

```
println("Approximate probability based on m = 1000 simulated data point is:",
        mean(y.<=1))
```

```
Approximate probability based on m = 1000 simulated data point is:0.388
```

The above simulation scheme suggested a highly positively skewed distribution for the random variable Y. Let us try to compute the exact PDF of Y. We have already learnt how to compute $f_Y(y)$ by computing the CDF from the definition. In this case, the following integration needs to be carried out:

$$F_Y(y) = P\left(X_1^2 + X_2^2 \leq y\right) = \int\int_{\{(x_1,x_2):x_1^2+x_2^2\leq y\}} \frac{1}{2\pi} e^{-\frac{1}{2}(x_1^2+x_2^2)} dx_1 dx_2.$$

You are encouraged to perform this integration, however, we can try some alternative approach as well. For example, we now understand that both X_1^2 and X_2^2 follow $\mathcal{G}(\alpha, \beta)$ with $\alpha = \frac{1}{2}$, $\beta = 2$, and Y is nothing but the sum of two independent $\mathcal{G}(\cdot, \cdot)$ distributions. We recollect the moment generating function of the $\mathcal{G}(\alpha, \beta)$ distribution, where α and β are the shape and scale parameters, respectively.
The MGF is given by

$$M(t) = (1 - \beta t)^{-\alpha}, \quad t < \frac{1}{\beta}.$$

In addition, the expected value and variance of this distribution is $\alpha\beta$ and $\alpha\beta^2$, respectively.

Suppose that $W_1 \sim \mathcal{G}(\alpha_1, \beta)$ and $W_2 \sim \mathcal{G}(\alpha_2, \beta)$ and W_1, W_2 are independent random variables and let $W = W_1 + W_2$. Then the MGF of W can be computed as

$$\begin{aligned} M_W(t) &= \mathbb{E}\left(e^{tW}\right) = \mathbb{E}\left(e^{tW_1+tW_2}\right) \\ &= \mathbb{E}\left(e^{tW_1}\right)\mathbb{E}\left(e^{tW_2}\right) = M_{W_1}(t)M_{W_2}(t) \quad \text{(independence)} \\ &= (1 - t\beta)^{-(\alpha_1+\alpha_2)}, \quad t < \frac{1}{\beta}. \end{aligned}$$

Therefore, $W \sim \mathcal{G}(\alpha_1 + \alpha_2, \beta)$; in particular, the addition of these two independent random variables also belongs to the $\mathcal{G}$ family of distributions. Using this result, we can see that

$$Y = X_1^2 + X_2^2 \sim \mathcal{G}(\alpha = 1, \beta = 2).$$

Let us check whether the theory is matching with the simulated histograms of the Y values in the previous figure.

In [6]:

```
using Plots, Distributions, Statistics, StatsBase, LaTeXStrings
```

In [7]:

```
m = 1000
x1 = rand(Normal(0,1), m)
x2 = rand(Normal(0,1), m)
p1 = scatter(x1, x2, color="grey", markersize= 3, xlabel=L"x_1",
        ylabel=L"x_2", label="")
hline!([0], color = "red", lw = 2, label = "" )
vline!([0], color = "red", lw = 2, label = "" )
y = (x1.^2).+(x2.^2)
```

```
p2 = histogram(y, normalize = true, bins = 30, xlabel = "y",
    ylabel = "density", title = "histogram of y",ylims = (0, 0.45),
    label = "")
x_vals = range(0,15,length = 1000)
pdf_vals = pdf.(Gamma(1,2), x_vals)
plot!(x_vals, pdf_vals, color = "red", label = "")
plot(p1, p2, layout = (1,2), size = (800,400))
```

Out[7]:

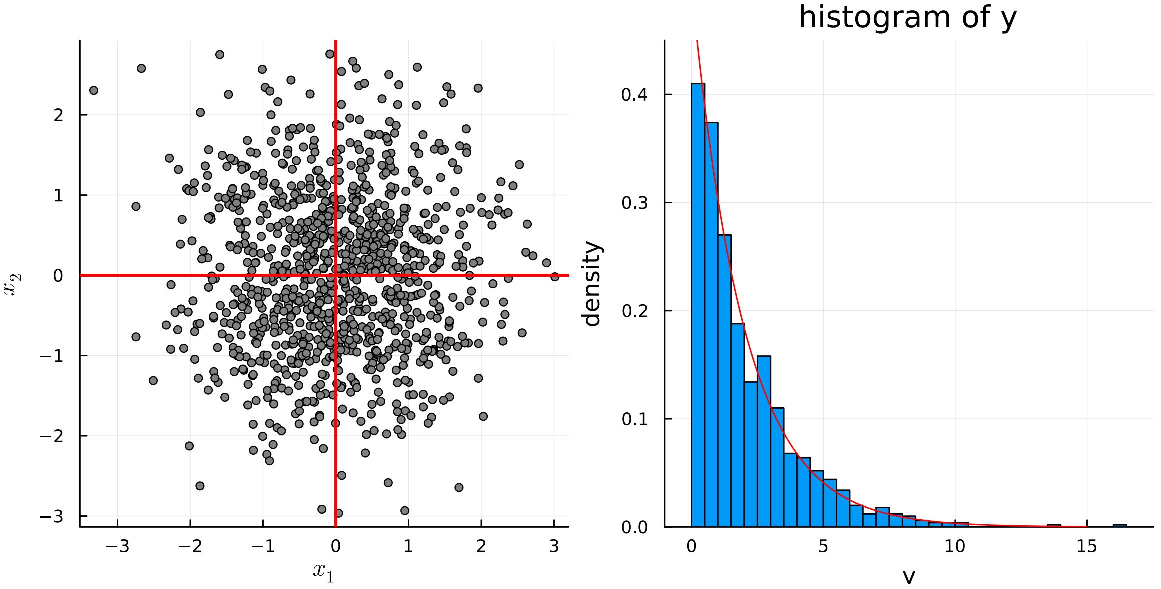

Figure 3: The histogram of the simulated realizations of Y based on 1000 simulations of pairs of independent standard normal random variables. In addition, we observe that the histogram is well approximated by the $\mathcal{G}(\alpha = 1, \beta = 2)$ PDF which basically the exponential PDF with mean 2.

Now are in a position to generalize this idea. We just extend the idea using an important result which states that if $W_i \sim \mathcal{G}(\alpha_i, \beta)$ for $1 \leq i \leq n$ and they are independent, then

$$\sum_{i=1}^{n} W_i \sim \mathcal{G}$$

Let us write them in a sequential manner:

- $X_1 \sim \mathcal{N}(0,1)$, then $Y_1 \sim \mathcal{G}(\alpha = \frac{1}{2}, \beta = 2)$
- $X_1, X_2 \sim \mathcal{N}(0,1)$, then $Y_2 = X_1^2 + X_2^2 \sim \mathcal{G}(\alpha = 1, \beta = 2)$
- $\vdots$
- $X_1, X_2, \ldots, X_n \sim \mathcal{N}(0,1)$, then $Y_n = \sum_{i=1}^{n} X_i^2 \sim \mathcal{G}(\alpha = \frac{n}{2}, \beta = 2)$

Therefore, for $n \in \mathbb{N}$, the PDF of Y_n is given by

$$f_{Y_n}(y) = \frac{e^{-\frac{y}{2}} y^{\frac{n}{2}-1}}{2^{\frac{n}{2}} \Gamma\left(\frac{n}{2}\right)}, \quad 0 < y < \infty.$$

and zero otherwise. This probability density function is celebrated as the chi-squared distribution with n degrees of freedom and usually denoted as $Y_n \sim \chi_n^2$. From the properties of the $\mathcal{G}(\cdot, \cdot)$ distribution, we can easily conclude that $E(Y_n) = n$

$$\text{Var}(Y_n) = \frac{n}{2} \times 2^2 = 2n.$$

Let us visualize the distribution of the Y_n for different choices of n values based on $m = 1000$ replications.

```
In [8]: using Plots, Distributions, Statistics, StatsBase, LaTeXStrings
```

```
In [9]: plt = plot(layout=(2, 3), size=(900, 600))
        n_vals = [2, 5, 10, 30, 40, 100]

        for (idx, n) in enumerate(n_vals)
            m = 1000
            sim_data = Matrix{Float64}(undef, m, n)
            for j in 1:n
                sim_data[:, j] .= randn(m)
            end
            y = zeros(m)
            for i in 1:m
                y[i] = sum(sim_data[i, :].^2)
            end
            histogram!(plt, y, norm=true, title="n = $n", bins=30,
                color="lightgrey",label = "",subplot=idx)
            x_vals = range(minimum(y), maximum(y), length=1000)
            plot!(plt, x_vals, pdf.(Gamma(n/2, 2), x_vals), color="red",
                linewidth=2, subplot=idx, label = "")
        end

        display(plt)
```

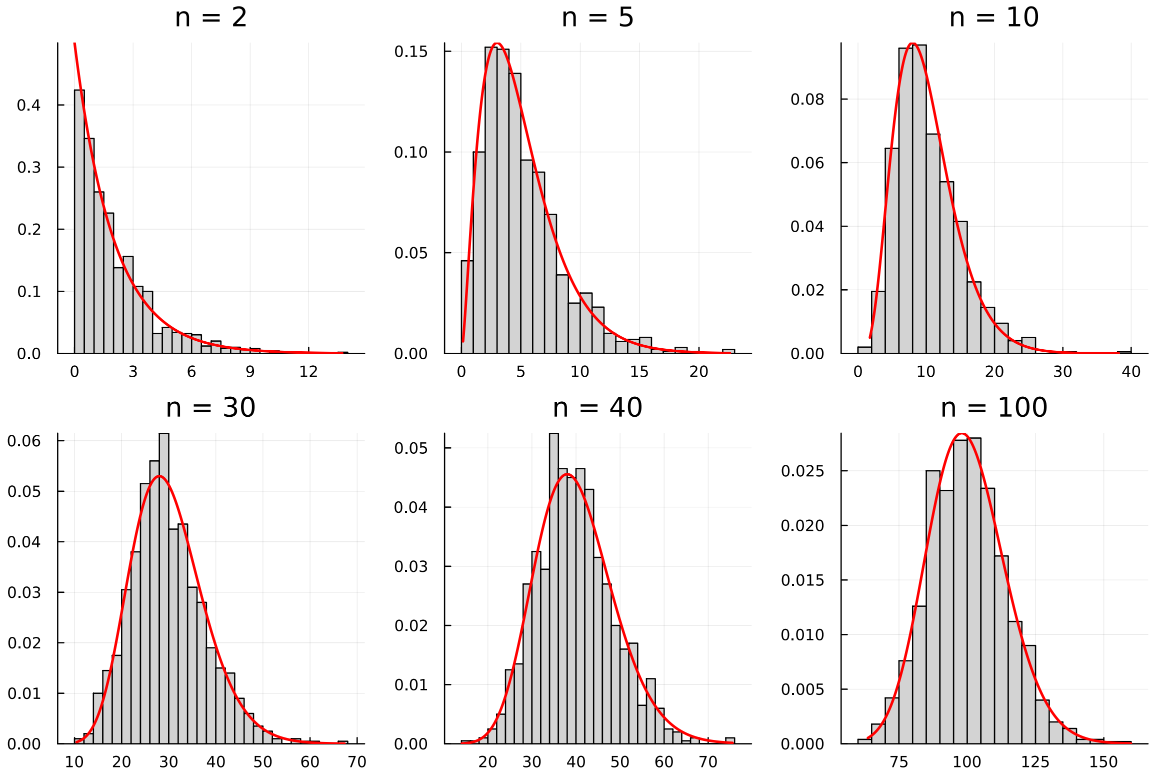

Figure 4: The histogram of the simulated realizations of Y based on 1000 simuations of n copies of independent standard normal random variables and their sum of squared values. The Chi-squared distribution with n degrees of freedom is overlaid on the histograms.

While these histograms are well approximated by the χ^2_n distribution, one student notices that as n increases, the histograms and the PDFs are behaving similar to a bell curve, that is,

the normal distribution. Therefore, a natural question arises, is the χ_n^2 PDF looks like a bell curve for large n?
In addition, what will be the mean and variance of this bell curve if these distributions are well approximated by a bell curve? For experiment purposes, let us draw the χ_n^2 PDF and the normal PDF with mean n and variance $2n$, that is $\mathcal{N}(n, 2n)$, for different choices of n.

In [10]:
```julia
plt = plot(layout=(2, 3), size=(900, 600))
n_vals = [2,8,10,30,50,100]

for (idx, n) in enumerate(n_vals)
    x_vals = range(n - 4*sqrt(2*n), n + 4*sqrt(2*n), length=1000)
    pdf_chisq = pdf.(Chisq(n), x_vals)
    pdf_norm = pdf.(Normal(n, sqrt(2*n)), x_vals)
    plot!(plt, x_vals, pdf_chisq, color="red", lw=2, xlabel=L"x",
          ylabel=L"f(x)", xlims=(n - 4*sqrt(2*n), n + 4*sqrt(2*n)),
          label=L"\chi_n^2", subplot=idx)
    plot!(plt, x_vals, pdf_norm, color="blue", lw=2, linestyle=:dash,
          label=L"N(n,2n)", subplot=idx)
end

display(plt)
```

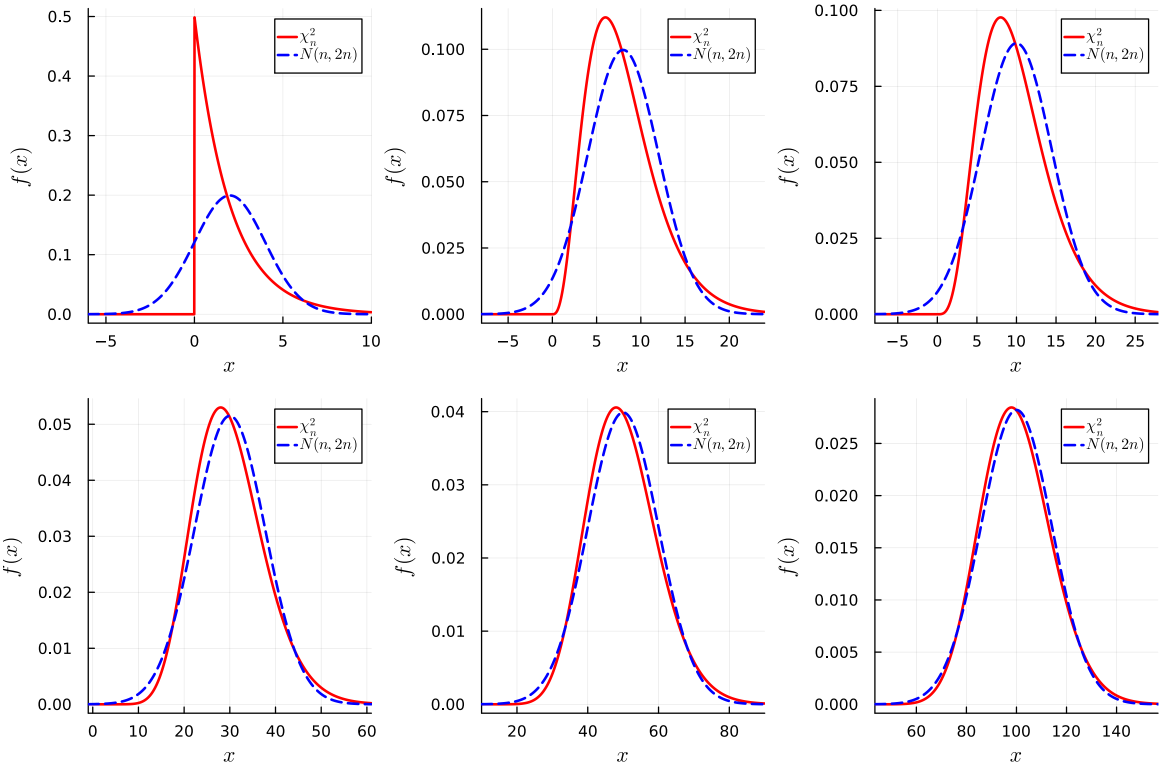

Figure 5: The figures clearly suggest that as the degrees of freedom increases, the χ_n^2 distributions are well approximated by the normal distribution. The normal distributions are overlaid with dotted blue color line

For a theoretical proof of the above observations, we can prove in the class that for large n, the Moment Generating Function of $\frac{X-n}{\sqrt{2n}}$ is approximately equal to $e^{\frac{t2}{2}}$ which is the MGF of the standard normal distribution.
Since

$$\frac{X-n}{\sqrt{2n}} \sim \mathcal{N}$$

therefore,

$X \sim \mathcal{N}$

It is important to plot the curves of the PDFs for different choices of the parameters.

Most surprising transformation (Probability Integral Transform)

The probability integral transform is a fundamental theorem in statistical simulation. Suppose that we are interested in simulating random numbers from the distribution of X, whose CDF is given by $F_X(x)$, say.
If we consider the transformation $U = F_X(X)$, basically, the transformation function $g(\cdot)$ is the CDF itself $F_X(\cdot)$, then

$$U \sim \text{Uniform}(0, 1).$$

This is quite surprising as the result is true for any random variable.
Therefore, if we simulate $U \sim \text{Uniform}(0, 1)$, then we can obtain a realization from the distribution of X, by considering the following inverse transformation:

$$X = F_X^{-1}(U).$$

We must take caution in defining the inverse mapping F_X^{-1} of F_X. For example, if X is a discrete random variable, then $F_X(x)$ is a step function, therefore, the inverse cannot be properly defined, in usual sense. We will not discuss it further here and close this session with an illustrative example using the exponential(1) distribution. The CDF is given by $F_X(x) = 1 - e^{-x}, 0 \leq x < \infty$ and zero for $x < 0$. Therefore, for $F_X^{-1}(U) = -\log(1-U), 0 < U < 1$. The following code is used to demonstrate the simulation of exponential(1) random variables starting with the Uniform(0,1) random numbers.

In [11]:

```
using Plots, Distributions, Statistics, StatsBase, LaTeXStrings
using StatsPlots
```

In [12]:

```
CDF_exp(x) = (1 .-exp.(-x))*(x>0)

n = 1000
x = rand(Exponential(1),n)
p1 = histogram(x, normalize = true, color = "lightgrey",
     xlabel = "x", ylabel = "density", label = "")
p2 = plot(ecdf(x), color = "blue" ,xlabel="x",
    ylabel=L"F_n(x)", label="")
plot!(CDF_exp, -1, 6, color = "red", lw = 2, label = "")

U = CDF_exp.(x)
p3 = histogram(U, normalize = true, color = "lightgrey",
     xlabel = "U", ylabel = "density", label = "",
     title = "histogram of U", xlims = (0,1) )
x_vals = range(minimum(x),maximum(x), length = 1000)
pdf_vals = pdf.(Uniform(0,1), x_vals)
plot!(x_vals, pdf_vals, color = "red", lw = 2, label = "")

inv_CDF_Exp(u) = -log.(1 .-u)
```

```
U = rand(Uniform(0,1),n)
p4 = histogram(U, normalize = true, color = "lightgrey",
    xlabel = "u", ylabel = "density", label = "")
x = inv_CDF_Exp.(U)
p5 = plot(ecdf(x), color = "blue" ,xlabel="x",
    ylabel=L"F_n(x)", label="")
plot!(CDF_exp, -1, 6, color = "red", lw = 2, label = "")
p6 = histogram(x, normalize = true, color = "lightgrey",
     xlabel = "x", ylabel = "density", label = "")
plot(p1,p2,p3, layout = (2,2), size = (800, 600))

plot(p1,p2,p3,p4,p5,p6, layout = (2,3), size = (800, 600))
```

Out[12]:

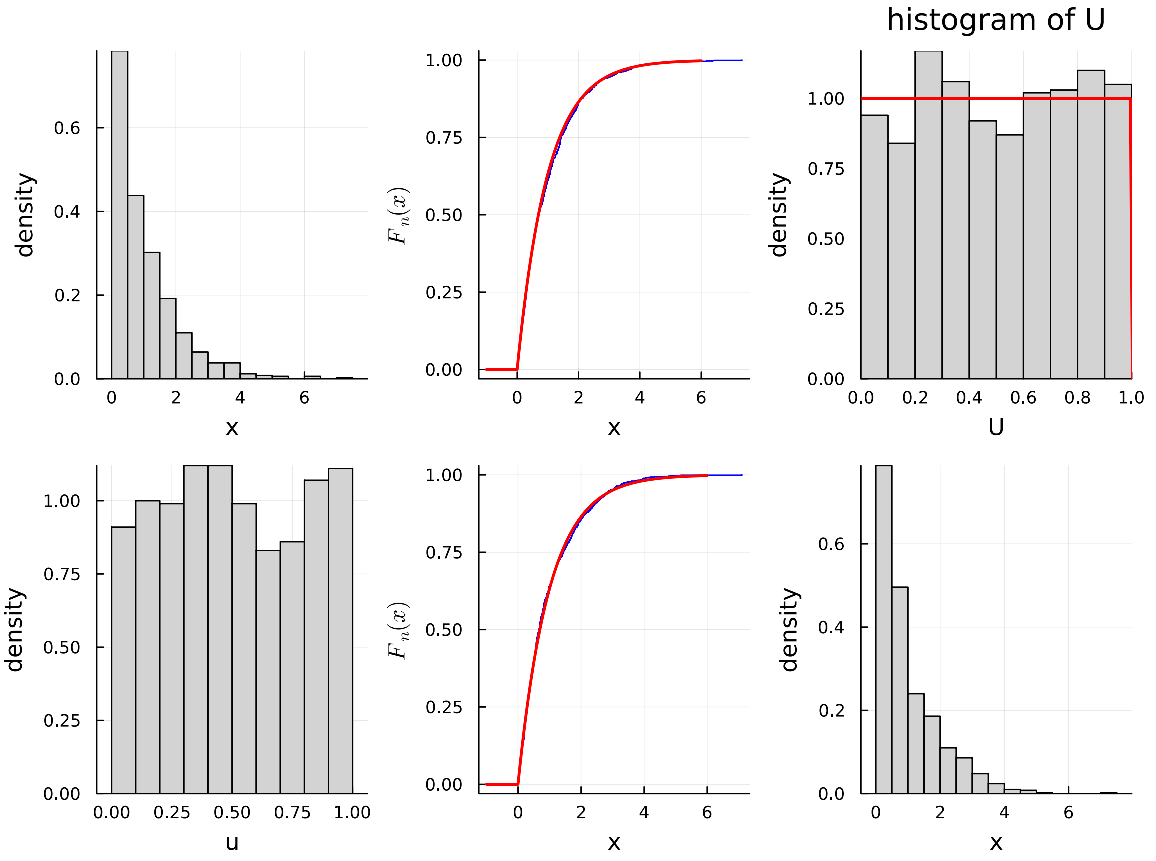

Figure 6: Top panel: We simulated observations from the $\text{Exponential}(1)$ distribution.Then simulated values are transformed using $F_X(\cdot)$ function, and a histogram of the resulting values is drawn. It is evident that the transformed values are indeed uniform(0,1) distributed. In the bottom panel, we simulated from the uniform(0,1) distribution and transformed them using $F_X^{-1}(\cdot)$, and the histogram of the resulting values confirms the Exponential(1) distribution.

Maximum and Minimum of Two random variable

Suppose U_1 and U_2 are two independent uniformly distributed random variables from the interval $(0, t)$ for $t > 0$. We are interested to understand the sampling distribution of the $\max(U_1, U_2)$ and $\min(U_1, U_2)$. In the following code, we simulate the sampling distribution of these two functions. You are encouraged to simulate the sampling distributions for the maximum and minimum of n independent and identically distribution Uniform(0, t) random variables. These are also called the maximum and minimum order statistics and typically denoted as $U_{(n)}$ and $U_{(1)}$, respectively.

In [13]:

```
using Plots, Statistics, StatsModels, LaTeXStrings
```

```julia
using Distributions
```

In [14]:

```julia
t = 3
U = rand(Uniform(0, t), 2)
length(U)
```

Out[14]:

```
2
```

In [15]:

```julia
U1 = minimum(U)
print(U1)
```

```
1.7296226254433211
```

In [16]:

```julia
U2 = maximum(U)
print(U2)
```

```
2.4716298671104986
```

In [17]:

```julia
M = 10000  # Number of simulations
n = 2      # Number of samples per trial

U1 = zeros(M)
U2 = zeros(M)

for i in 1:M
    U = rand(Uniform(0, t), 2)
    U1[i] = minimum(U)
    U2[i] = maximum(U)
end

function f_U2(x)
     n * (x/t)^(n-1) * (1/t) * (0 < x < t)
end

function f_U1(x)
     n * (1 - x/t)^(n-1) * (1/t) * (0 < x < t)
end

x_vals = range(0, t, length=100)

p1 = histogram(U1, normalize=:pdf, xlabel=L"U_1",
    ylabel="Density", bins=30, label="", title = L"f_{U_1}(U_1)")
plot!(x_vals, f_U1.(x_vals), color="red", lw=2, label="")

p2 = histogram(U2, normalize=:pdf, xlabel=L"U_2",
    ylabel="Density", bins=30, label="", title = L"f_{U_2}(U_2))")
plot!(x_vals, f_U2.(x_vals), color="red", lw=2, label="")

p3 = histogram2d(U1, U2, bins=30, color=cgrad(:heat, [5]),
    xlabel=L"U_{(1)}", ylabel=L"U_{(2)}")

plot(p1, p2 , p3, layout = (2,2), size = (800, 600))
```

Out[17]:

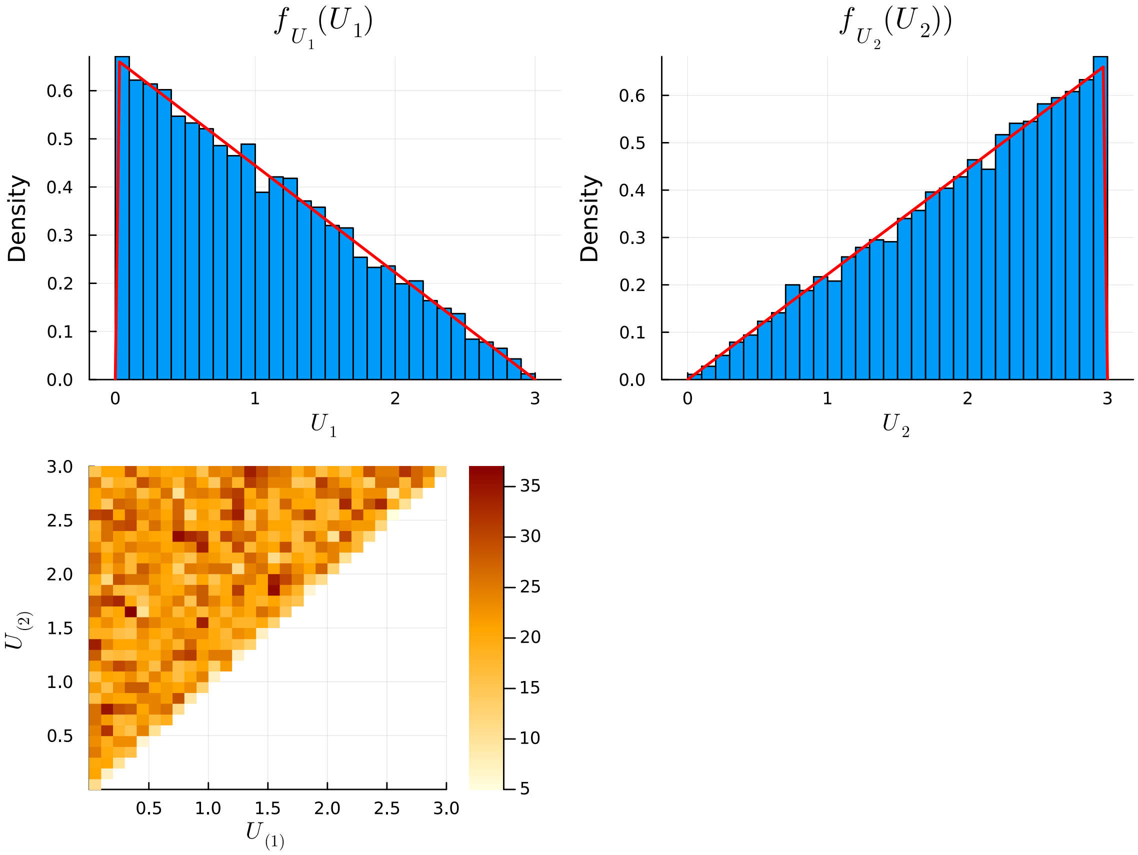

Figure 7: The sampling distributions of the maximum and minimum of two independent uniformly distributed random variables from the interval $(0, t)$. The joint distribution of these two functions of $\mathrm{Uniform}(0, t)$ random variables are shown at the rightmost panel.