Statistical Genomics Homework 1

Enrique Jimenez Schwarzkopf

January 31, 2017

1

Start from random variables with standard normal distribution, define your own random variable that is function of the normal distributed variables. Name the random variable as your last name and develop a R function to generate the random variable. The input of your R function should include n, which is number variables to be generated, and parameters for the distribution of the random variable you defined. Note: try not to be the same as the known distributions such as Chi-square, F and t.

We created a function that operates similarly to the `rchisq`, where we take a sample size (`n`, 10 by default) and number of degrees of freedom (`df`, 2 by default) and produce a variable that is the sum of the natural logarithms of `df` standard normal distributions.

```
rJimenez<-function(n=10, df=2){
 y<-replicate(n,{
   x=rnorm(df,0,1)
   y=sum(log(abs(x)))
 })
 return(y)
}
```

2

Sample ten thousand observations from the distribution you defined. Graph their properties and describe the potential application of your distribution in nature.

We set `n` to 10,000 and `df` to 10, and run `rJimenez` to obtain the desired sample.

```
Jimenez<-rJimenez(10000, 10)
```

We used the following to graph our sample from the Jimenez distribution:

```
par(mfrow=c(2,2), mar=c(3,4,1,1))
plot(Jimenez)
hist(Jimenez)
plot(density(Jimenez))
plot(ecdf(Jimenez))
```

And we obtain:

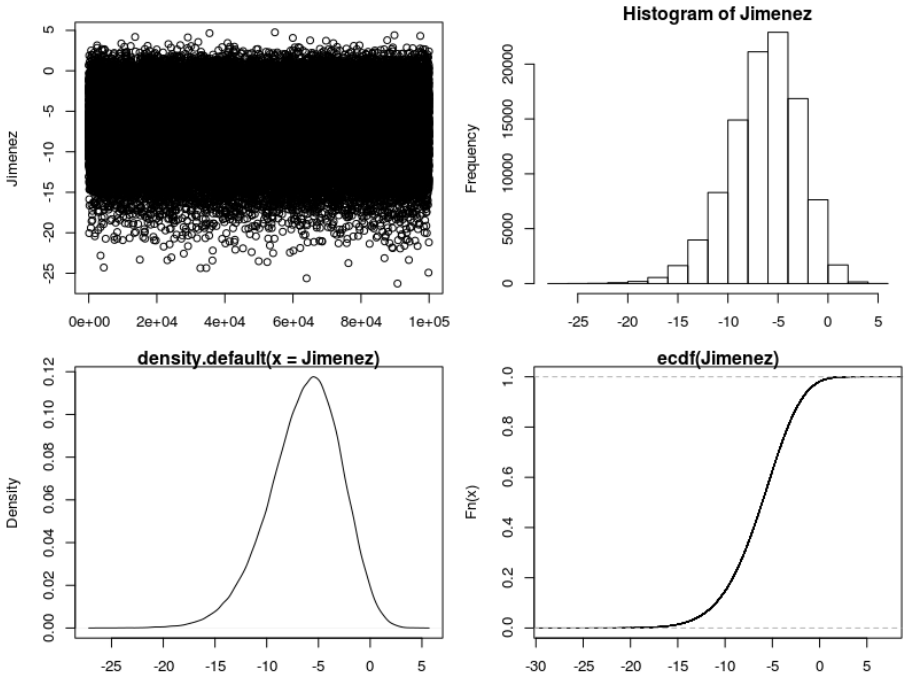

Figure 1: Plots showing the 10000 samples of the Jimenez distribution

These figures indicate that the Jimenez distribution has a long left tail, which could be useful in cases of near normality, but with more extremely low values. This might be the case with population counts near the edge of a species range.

3

There is an R function, rf(n, df1, df2), to generate n random variables that follow F distribution with degree freedom of df1 and df2. Develop your own R function to do the same thing. Your function should start with normal distribution by using rnorm function in R, followed by derivation of Chi-square distribution variables.

We developed the following function to generate `n` random variables that follow the F distribution:

```
rfJimenez<-function(n=10, df1=1, df2=2){
 w<-replicate(n,{
   x=rnorm(df1,0,1)
   chi1=sum(x^2)
   z=rnorm(df2,0,1)
   chi2=sum(z^2)
   y=(chi1/df1)/(chi2/df2)
 })
 return(w)
}
```

4

Sample 10, 100, 1000 and 100,000 F distributed variables by using the R function you developed. Calculate means of these samples and test them on the null hypothesis that the samples have expected mean of $\frac{df_2}{(df_2-2)}$

We begin by generating 10, 100, 1000, and 100,000 F distributions using `rfJimenez`:

```
x10<-rfJimenez(10,10,20)
x100<-rfJimenez(100,10,20)
x1000<-rfJimenez(1000,10,20)
x100000<-rfJimenez(100000,10,20)
```

We created the `Mtest` function to empirically test the hypothesis that the means of these samples have the expected mean of: $\frac{df_2}{(df_2-2)}$.

```
Mtest<-function(x, n=10, k=10000, df1=10, df2=20){
 xt<-mean(x)-(df2/(df2-2))
 y=replicate(k,{y=mean(rf(n,df1,df2))-(df2/(df2-2))})
 P=length(y[y>xt])/k
 return(P)
}
```

And we apply `Mtest` to all of our samples:

```
Mtest(x10)
[1] 0.5
Mtest(x100,n=100)
[1] 0.0816
Mtest(x1000,n=1000)
[1] 0.6701
Mtest(x100000,n=100000)
[1] 0.6798
```

We find that none of the samples show a mean that is significantly ($\alpha = 0.05$) different from the expected $\frac{df_2}{(df_2-2)}$.

5

Sample 10, 100, 1000 and 100,000 F distributed variables by using the R function you developed. Calculate variance of these samples and test them on the null hypothesis that the samples have expected variance of $\frac{2df_2^2(df_1+df_2-2)}{df_1(df_2-2)^2(df_2-4)}$.

We use a similar function to `Mtest`, called `Vtest` to empirically test the hypothesis that the variance of each sample is $\frac{2df_2^2(df_1+df_2-2)}{df_1(df_2-2)^2(df_2-4)}$.

```
Vtest<-function(x, n=10, k=10000, df1=10, df2=20){
 V=((2*(df2^2)*(df1+df2-2))/(df1*((df2-2)^2)*(df2-4)))
 xt<-var(x)-V
 y=replicate(k,{y=var(rf(n,df1,df2))-V})
 P=length(y[y>xt])/k
 return(P)
}

Vtest(x10)
[1] 0.6468
Vtest(x100,n=100)
[1] 0.2554
Vtest(x1000,n=1000)
[1] 0.6779
Vtest(x100000,n=100000)
[1] 0.7063
```

As we found with the hypothesis tests for the mean, none of the samples show a variance that is significantly ($\alpha = 0.05$) different from the expected $\frac{2df_2^2(df_1+df_2-2)}{df_1(df_2-2)^2(df_2-4)}$.