Nikayla Strauss

CROPS 545

Homework 1

1) Define a random variable that is a function of random variables with known distributions such as uniform, binomial, Poison, normal, Chi square, F, or t distributions. Name the distribution of your new random variable as your last name and develop a R function to generate random variables. The input of your R function should include n, which is number of variables to be generated, and parameters for the distribution you defined.

Function: strauss=function(n,df1), h=rbinom(n,5,0.5), f=rt(n,5), y=h-f

The function above is a function of two random variables, h and f. H is a variable with a binomial distribution with n observations, 5 trials, and a 0.5 probability of success. F is a variable with a t distribution, n observations, and 5 degrees of freedom. Degrees of freedom is the defined parameter, labeled df1.

2) Sample 10,000 observations from the distribution you defined. Make scatter, histogram, density, and accumulative density plots.

Hypothesis: Based on the high number of observations, I hypothesize that the distribution will appear normal.

Method: The distribution was defined as the strauss function with 10,000 observations and the degrees of freedom as 100. The plots were then visualized with standard R code. For example: hist(dist) would output a histogram of the distribution. The ecdf function was used to calculate and plot the cumulative density.

Results: Below are the scatter, histogram, density, and cumulative density plots.

Figure 2

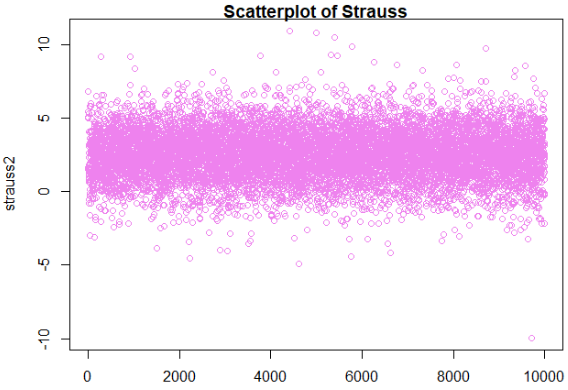


Figure 1

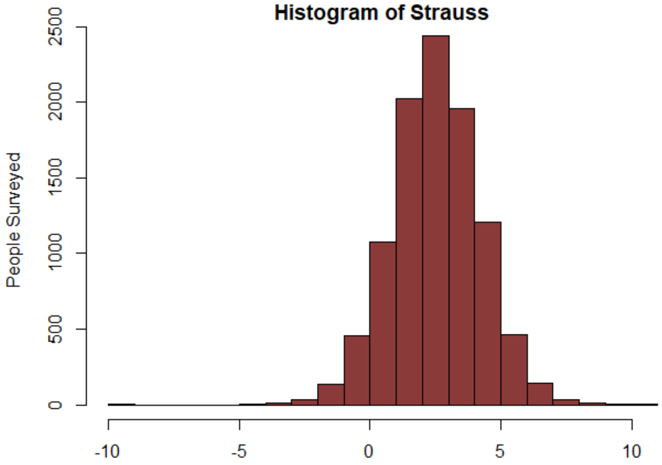


Figure 4

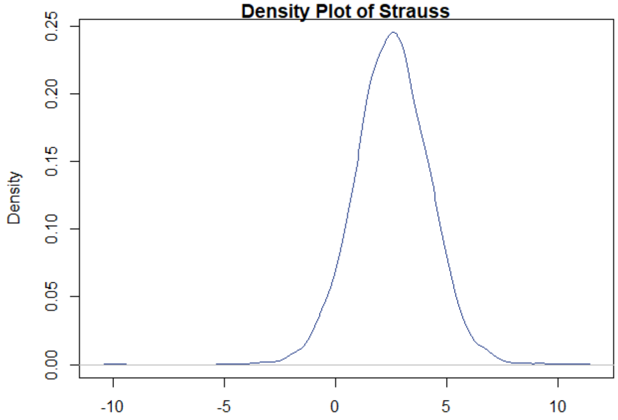
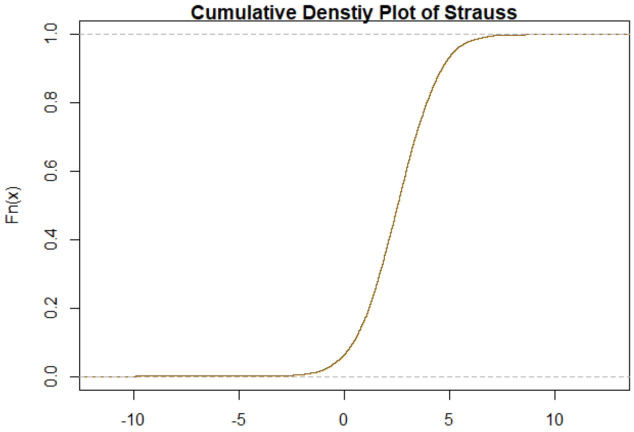


Figure 3



Discussion: Each plot supports the hypothesis that with 10,000 observations, the distribution becomes close to normal. It is also evident that the distribution is shifted to the right and centered around 2.5. Another interesting observation of this distribution is that it has two tails. The tails stretch to -10 and 10, but the left tail is not continuous between -5 and -10.

3) Create tables for your variable at different percentile (1%, 5%, 10%, 50%, 90%, 95% and 99%), and describe the impact of the parameters of your distribution.

Hypothesis: My parameter, degrees of freedom, will have a substantial impact on the shape of the distribution.

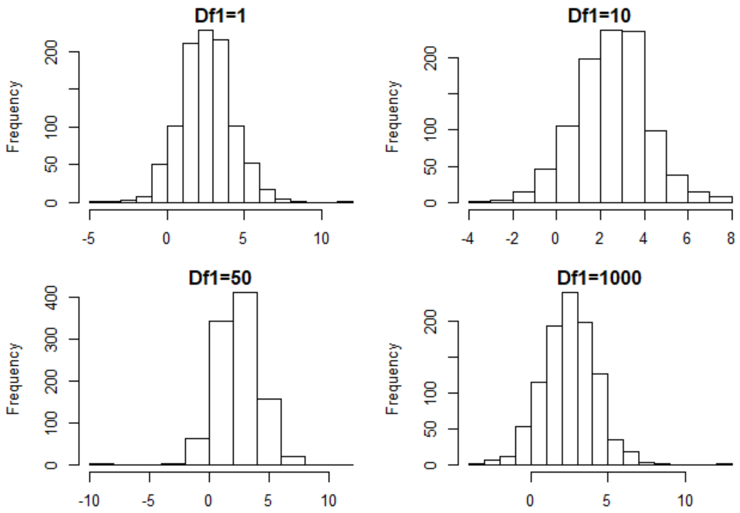
Method: To create the table of my variable at different percentiles I used the quantile function in R, which allows for specific percentiles to be included in a single line of code. A table was then generated in Microsoft Word. To determine the impact of the degrees of freedom on the distribution, I ran the function several times, each time changing the value of the degrees of freedom and observing the impact of the change on the histogram.

Results:

Table 1

Percentile	Variable
1	-1.526
5	-0.224
10	0.433
50	2.522
90	4.628
95	5.210
99	6.575

Figure 5



Discussion: The percentile table (Table 1) gives information about the distribution of the data at given probabilities. For example, 10% of the data is below 0.433 and 95% of the data is below 5.210. No matter what the value of the degrees of freedom are, the mean stays centered around 2.5, and the min and max stay the same (Figure 5). While the hypothesis can be considered true, the difference is not as dramatic as was expected. As degrees of freedom increase, it can be observed that the distribution becomes more normal and gradual. The exception is when df1=50. When df1=50, the distribution is less normal. An explanation for this might be because of the properties of the binomial distribution and how it is defined in the strauss function.

4) Give an example to make sense out of your distribution.

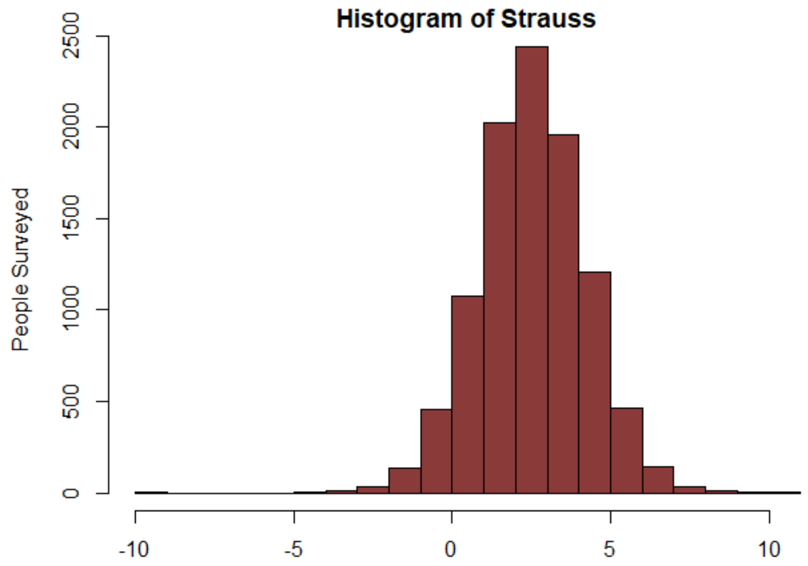
Hypothesis: On average, Americans gain weight over the course of a year.

Method: A random sample of 10,000 Americans were surveyed with questions about how much weight they have gained or lost this year. This data was analyzed using a model of a t-distribution subtracted from a binomial distribution. All plots and statistics were generated in R.

Results:

Table 2

Statistic	Value
Mean	2.494
Min	-9.717
Max	11.046
Variance	2.901



As shown above, and average American gains approximately 2.5 pounds per year. Of the individuals sampled, the maximum weight loss was 9.7 pounds, and the maximum weight gain was 11.05 pounds (Table 2).

Discussion: It is logical that this data appears normal because of the random individuals sampled and the nature of the data. Weight gain or loss is widely varied and can be any value. The mean, however, is not centered around 0. This confirms the hypothesis that on average, Americans gain weight over the course of a year. While 2.5 pounds is not significant for most people, this number reflects a larger trend in American society. Future research should include comparison studies between America and other 1st world countries and a study that factors age into the analysis.

5) Generate one or multiple samples with sizes of your choices from the distribution you defined, and define a statistics from your samples.

As shown in the code, labeled Part 5, the strauss function was defined as having degrees of freedom equal to 50 and a sample size of 100. The statistic to be defined was variance, which in this part is equal to 3.061.

6) Create ten thousand replicates of your statistics and make the same plots in Part 2.

Hypothesis: The distribution of the variance will also be normal.

Method: The 10,000 replicates were executed with the replicate function in R. A distribution was then made with the results of the replicated function. The plots were then generated with the same code as in Part 2.

Results:

Figure 9

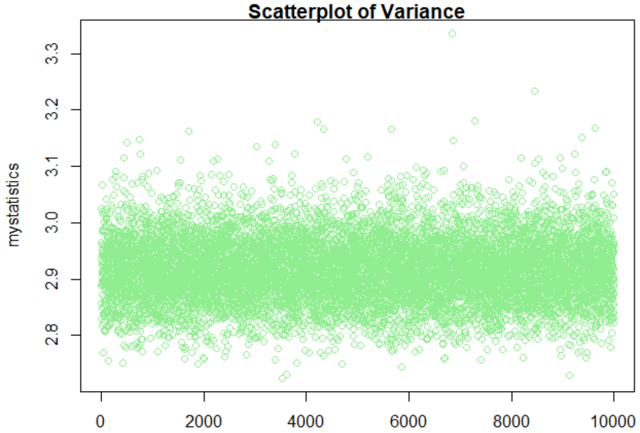


Figure 8

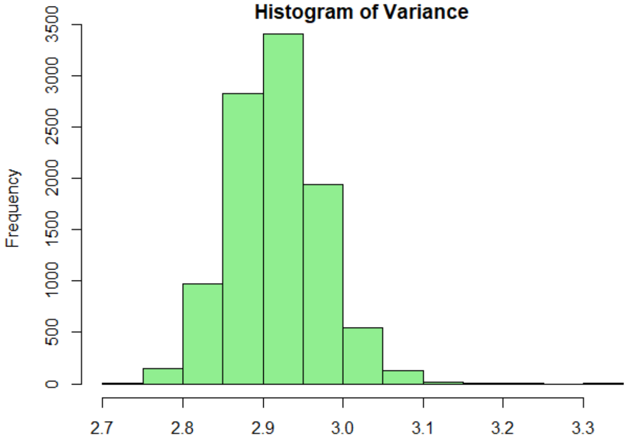


Figure 7

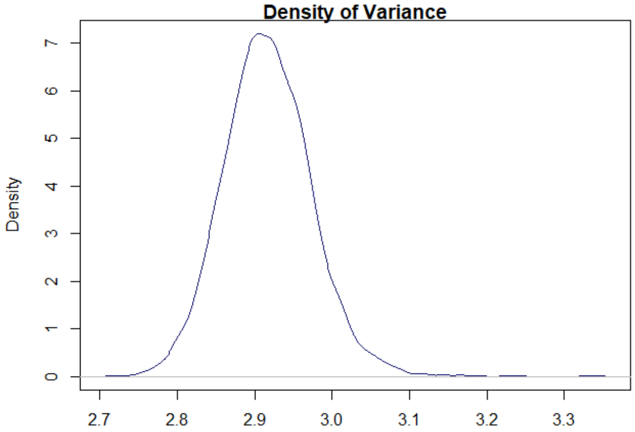
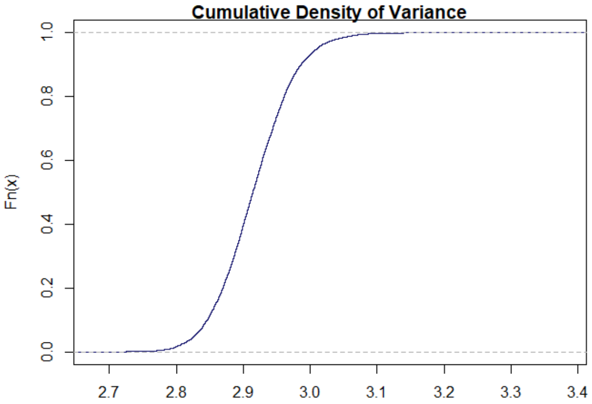


Figure 6



The variance is approximately normally distributed (Figure 7), but with a smaller range and a higher variance than the strauss function. If a random sample of 10,000 is replicated 10,000 times, the average variance is 2.916, with a min and max of 2.725 and 3.336 respectively. The variance of the variance in the strauss function is 0.003.

Discussion: The results of this test both confirm the hypothesis and show that there is limited variance in the original function. This indicates that most values do not differ greatly from the mean.