STATISTICAL COMPUTING ACTIVITY

4a: Some Continuous Distributions

Purpose: As expected, the main purpose of this activity is to show you how to use R to deal with the three continuous random variables we discussed in class. You will want to know how to calculate cumulative probabilities and quantiles. You will also want to be able to generate random variables based on the distribution. This last will help you better understand the variables and functions of the variables.

R Functions: We will see the following functions in R.

- runif
- punif
- qunif
- rexp
- pexp
- qexp
- rnorm
- pnorm
- qnorm

PROCEDURE

As usual, here is the procedure. Make sure you understand that the goal is not for you to get to the finish line. The goal is for these steps to help you better understand how to achieve the purposes and goals listed above. Racing through these without savoring them wastes your precious time.

PART O: THE START

1. Start R.
2. Open a new script. Make sure you save this script at the end. It will aid you in reviewing these steps in the future.
3. As with the previous SCA, we are generating random values. We will not need to load any external data.
4. You do still need to source the usual file. You will always need to do that:

```
source("http://rfs.kvasaheim.com/stat200.R")
```

PART I: THE UNIFORM RANDOM VARIABLE

Arguably, this is the ancestor of all random variables — both continuous and discrete. It is also the simplest of the continuous random variables, as you discovered this week. Let us start this SCA with the Uniform distribution, using it to examine random variates, estimate population statistics, and calculate probabilities and quantiles.

1. Let us start by drawing a million random values from a Uniform distribution with minimum value 0 and maximum value 10:

   ```
   s = runif(1e6, min=0, max=10)
   ```

2. With that sample from the population, we can estimate the mean, variance, and standard deviation from that population:

   ```
   mean(s)
   var(s)
   sd(s)
   ```

   Of course, we know those values exactly. Check your notes for why we know that $\mu = 5$ and $\sigma^2 = 10^2/12 = 100/12 \approx 8.3333$.

3. We can also estimate the percentiles of the Unif(0, 10) distribution. Here are the quartiles.

   ```
   quantile(s)
   ```

   Here is the fifth percentile.

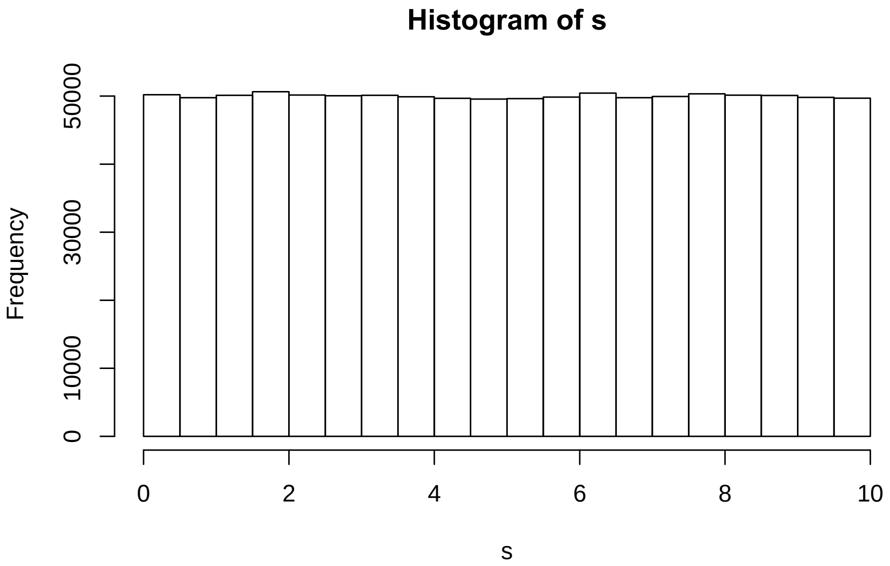
   ```
   quantile(s, 0.05)
   ```

   Again, we know these values exactly in the population. The five population quartiles are exactly 0.0, 2.5, 5.0, 7.5, and 10.0. The fifth percentile is 0.50. Check your notes for how we know this.

   In fact, it may be helpful to look at the cumulative distribution function for the Uniform distribution to refresh your memory about how I calculated those numbers.

   Also, check to see how closely your estimates came to the population values. Since we are using a sample of size one-million, your estimates should be rather close. If you had used a smaller sample size, like n=10, your estimates would most likely be ***far*** off.

4. Here is a histogram of the sample. What code did I use? If you remember back to the previous two SCA activities, this should be easy.



Yours should look very close to this. Since it is based on a random sample, there will be slight differences, however, due to random sampling. Make sure you remember that the histogram is to the sample as the probability density function is to the population.

5. With R, it is very easy to calculate exact probabilities (probabilities based on the population and not on the sample). Remember that generating random variates requires the r* form and calculating probability masses (discrete distributions) or densities (continuous distributions) uses the d* form. Calculating **cumulative probabilities** requires the p* form.

Thus, if X ~ Unif(0, 10), we calculate P[X ≤ 2]:

```
punif(2, min=0, max=10)
```

P[X ≥ 2]:

```
1 - punif(2, min=0, max=10)
```

P[X < 4]:

```
punif(4, min=0, max=10)
```

P[3 ≤ X ≤ 5]:

```
punif(5, min=0, max=10) - punif(3, min=0, max=10)
```

6. Calculating the quantiles (percentiles) is just as easy. The 2.5 percentile is

   ```
   qunif(0.025, min=0, max=10)
   ```

   Compare that exact value to your estimate from the sample above

   ```
   quantile(s, 0.025)
   ```

   They (most likely) agree to two places after the decimal (the hundredths position). That is because the sample size is so large. A smaller sample size would result in an estimate that is farther from the true (population) value.

   Here are two ways of obtaining the five population quantiles. See if you can see why they are the same.

   ```
   qunif(c(0.00,0.25,0.50,0.75,1.00), min=0,max=10)
   qunif(0:4/4, min=0,max=10)
   ```

   Here are the estimates from the sample

   ```
   quantile(s, c(0.00,0.25,0.50,0.75,1.00))
   ```

   How close are your estimates from the sample to the real values from the population? Again, because your sample size is so large, your estimate should be close to the truth. Had we a smaller sample, the estimates would tend to be much farther away.

PART II: THE EXPONENTIAL RANDOM VARIABLE

As discussed in class, the Exponential random variables are used to model or estimate waiting times. These waiting times may be the time waiting for a bus or for a death. If you are thinking about these SCA activities, all I need to do is tell you that the stem for dealing with the Exponential distribution in R is **`exp`**.

Think about how you would generate random variates, calculate cumulative probabilities, and determine quantiles from an Exponential distribution.

In R, the parameter for the Exponential distribution is its mean, **`rate`**. Thus, from your notes, if X ~ Exp($\theta = 5$) it is also true that X ~ Exp($\lambda = 1/5$). So, in R, we can generate a million random values from an Exponential distribution with $\theta = 10$ using

```
s = rexp(1e6, rate=1/10)
```

or using

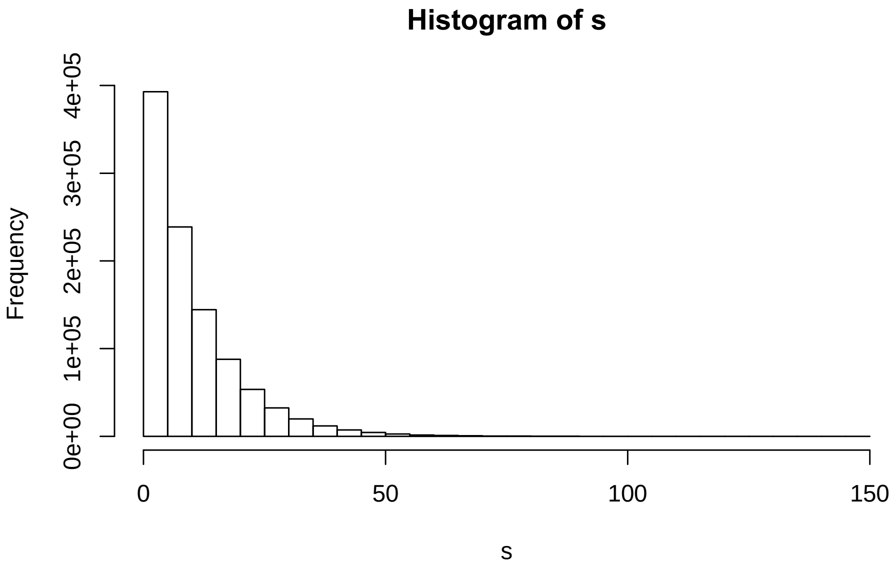
```
s = rexp(1e6, rate=0.10)
```

1. As usual, we can estimate the population statistics using our sample. Thus, the mean, variance, and standard deviation of our sample

```
mean(s)
var(s)
sd(s)
```

approximate the population mean, variance, and standard deviation (10, 100, and 10, respectively).

2. As usual, the histogram of our sample will approximate the probability density function. Here is the histogram.



What code did I use to get this histogram? How close is yours?

3. We can calculate cumulative probabilities using the same p* form. So, if X ~ Exp(λ = 1/10), then we can calculate P[X ≤ 4]:

```
pexp(4, rate=1/10)
```

P[X ≥ 4]:

```
1 - pexp(4, rate=1/10)
```

P[X ≤ 2]:

```
pexp(2, rate=1/10)
```

P[3 ≤ X ≤ 5]:

```
pexp(5, rate=1/10) - pexp(3, rate=1/10)
```

4. Again, we can work with the quantiles using the q* form. So, the exact 95th percentile from the population is

```
qexp(0.95, rate=1/10)
```

whereas the estimate from the population is

```
quantile(s, 0.95)
```

The population median is

```
qexp(0.50, rate=1/10)
```

and the sample median is

```
quantile(s, 0.50)
```

or

```
median(s)
```

Part III: The Normal Random Variable

The third continuous distribution we studied this week is the Normal distribution. Without question, this is the most important distribution of the entire course. Knowing how to work with this distribution will definitely come in handy as we move forward through life.

1. As with all random variables in R, the r* form generates random values. As such, if you want to generate a million random values from a Normal distribution with μ=68 and σ=2, perhaps to learn about heights of adult females in the United States, you could run

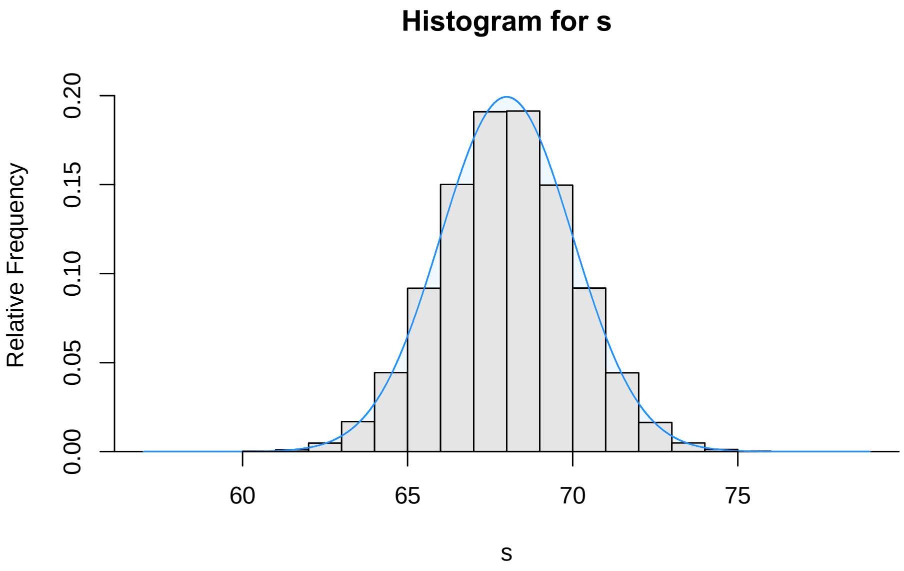
   ```
   s = rnorm(1e6, m=68, s=2)
   ```

   From that, we can estimate the population parameters based on these sample statistics

   ```
   mean(s)
   median(s)
   var(s)
   sd(s)
   ```

2. Of course, we can "see" the distribution of heights using the `hist` function. We can also see the distribution of heights with the Normal density curve added using

   ```
   normoverlay(s)
   ```



What does the curve represent? How closely does the distribution of the sample (histogram) match the distribution of the population (curve)?

3. For those following along, we know how to calculate exact cumulative probabilities. So, $P[X \le 65] =$

   ```
   pnorm(65, m=68, s=2)
   ```

   The probability that someone is taller than 5-foot 5-inches tall is

   ```
   1 - pnorm(65, m=68, s=2)
   ```

4. What rule do the following calculations illustrate?

   ```
   pnorm(68+2, m=68,s=2) - pnorm(68-2, m=68,s=2)
   pnorm(68+4, m=68,s=2) - pnorm(68-4, m=68,s=2)
   pnorm(68+6, m=68,s=2) - pnorm(68-6, m=68,s=2)
   ```

   Here is a hint, these illustrate the same rule, but from the other direction

   ```
   qnorm( c(0.16,0.66), m=68,s=2)
   qnorm( c(0.025,0.975), m=68,s=2)
   qnorm( c(0.0015,0.9985), m=68,s=2)
   ```

   Here is a final hint. How do these illustrate the Empirical Rule?

5. Lastly, we can (as last week) easily learn about functions of random variables. For instance, let T be the total time it takes for me to catch a bus in the morning. I know that it is the sum of two other random variables, $X \sim N(\mu=5, \sigma=1)$ and $Y \sim Exp(\lambda=0.5)$. Think of X as the time it takes for me to walk to the stop (mean 5 minutes and standard deviation 1 minute) and Y as the time I spend waiting at the bus stop (mean 2 minutes). Their sum is the time from when I leave home until I step onto the bus.

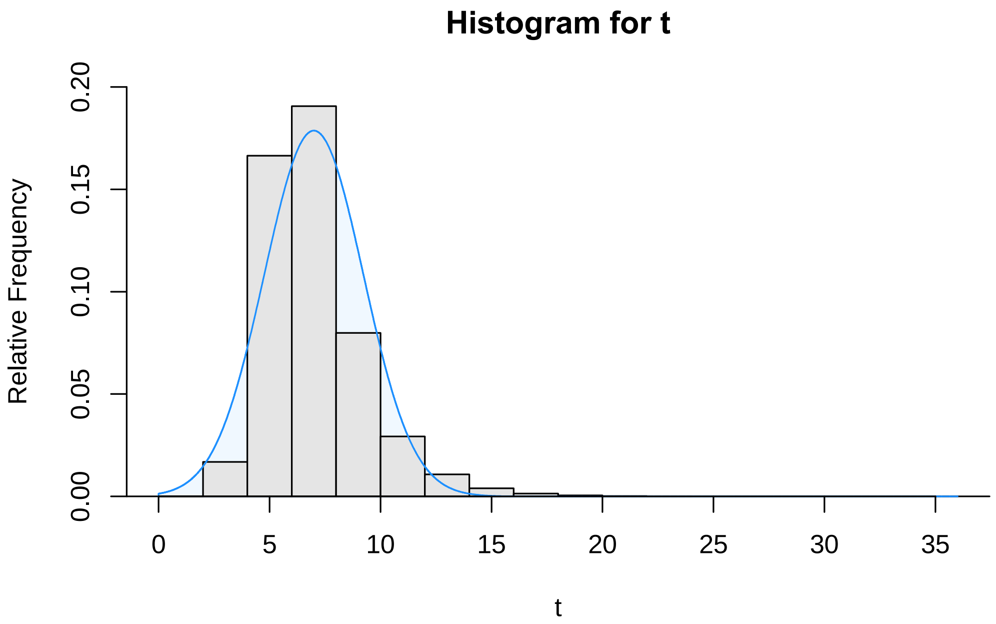
   So, if $T = X + Y$, what is the average time I will wait for the bus? The median time? I am 80% sure that I will wait between what two times? What does the distribution of times look like?

   All of these can be answered using both differential and integral calculus… or using R.

   ```
   x = rnorm(1e6, m=5,s=1)
   y = rexp(1e6, rate=0.5)
   t = x+y

   mean(t)
   median(t)
   quantile(t, c(0.10,0.90))
   hist(t)
   ```

So, we have the following estimates (yours should be close): The mean of the population is close to 7 minutes. The median of the population is approximately 6.57 minutes. Eighty percent of the time, I will wait between 4.67 and 9.85 minutes. The histogram is



Note that this is close to a Normal distribution. However, looking closely suggests that there is a slight right skew to the distribution. It is easier to see this skew in the peak and in the left side. It is slight.

6. One last thing to think about: Why am I using 1e6 in both the **x** and the **y** distribution? Why not 1e6 in one and 352 in the other? Why not 100 in both?