Statistics 251: Lab 3 Pre-reading

Distribution Functions

Objectives:

- Learn to use R to
  - Obtain probability mass functions $f(x)$
  - Obtain cumulative distribution functions $F(x)$
  - Simulate random draws
- Visualize probabilities using R (optional)
- Calculate probabilities using R

Important Reminders and Group Work:

- **Sit with the correct group.**
- If you are slower than the others or are shy, try to speak out. Participate! Watching others work can lead you to believe you understand when in fact you don't.
- If you are faster than the others or tend to dominate, slow down, allow others to work to their potential, and remember to allow others a chance to speak.
- If needed - and wanted - help other group members, but don't do things for them. Everyone must learn to work in the lab. Treat your team-mates with respect.
- If there are problems, try to work them out (respectfully!) within your group. If you find that difficult, please talk to the TA or your instructor.

1 Functions for Calculating $f(x), F(x)$, quartiles, and Simulating Random Draws

R has a bevy of functions available for calculating the probability mass functions $f(x)$, cumulative distribution functions $F(x)$, quartiles, and simulating random draws.

1.1 R Commands

Suppose we wanted to calculate the above for some distribution $<$ myDist $>$. We can calculate each as follows:

```
d<myDist>           f(x)
p<myDist>           F(x) = P(X <= x)
q<myDist>           quartile
r<myDist>           Random draws from <myDist>
```

1.2 Distributions Available that Will be Used in the Course

```
 Binomial:          <d, p, q, r>binom( _, size, prob )
*Geometric:         <d, p, q, r>geom( _, prob )
 Uniform:           <d, p, q, r>unif( _, min, max )
 Exponential:       <d, p, q, r>exp( _, rate )
 Normal:            <d, p, q, r>norm( _, mean, sd )
 T:                 <d, p, q, r>t( _, df )
 F:                 <d, p, q, r>f( _, df1, df2 )
```

* R defines the geometric random variables as **the number of failures** until the kth success, rather than **the number of trials** until the kth success, as we do in the course. Be very careful if you decide to use the geometric distribution in R!

2 Examples

Suppose that you are a Junior Engineer working on a project that involves knowing more information about the daily water discharge at a certain stream in Northern BC. From past records, you observe that the amount of daily water discharge at this stream is **uniformly distributed** with the quantity ranging anywhere from 10 to 25 cm^3/s during the fall, which is the time you will be working at the stream. You are going to calculate different probabilities associated with the daily water discharge to impress your supervisor with the knowledge of probability gained in your STAT 241/251 course. First you need to identify the random variable of interest, its distribution & probability density function and parameters associated with the probability density function. In this case, the random variable of interest is the daily water discharge.

Let X = the daily water discharge (cm^3/s) on any given day during the fall. As X follows a Uniform distribution, its probability density function is

$$f(x) = \begin{cases} \dfrac{1}{b-a} & \text{if } a \leq x \leq b \\ 0 & \text{otherwise} \end{cases}$$

This distribution is defined on the interval $[a, b]$. Any value chosen from the interval $[a, b]$ has an equal chance of being selected in this scenario. Thus, the probability density function is a constant function. Note that for the daily water discharge, $a = 10$ and $b = 25$. Therefore, the Uniform probability density function would be defined by $f(x) = 1/15$. We can visualize the probability density function of X using R (The R commands are given in Section 3).

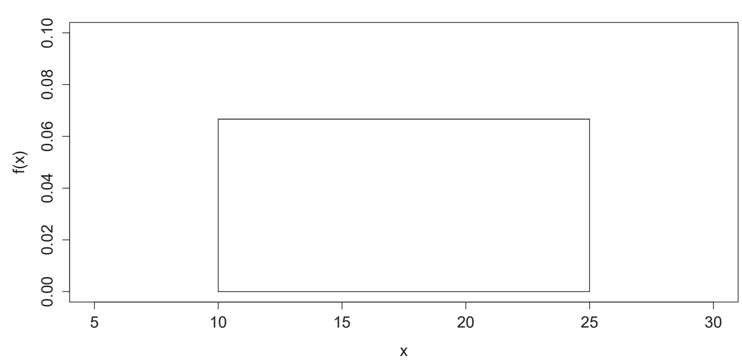

Figure 1: The probability density function of the daily water discharge.

2.1 Calculating Probabilites

Let's calculate some probabilities associated with the daily water discharge.

(a) Suppose you would like to find the probability that on any given day during the fall, the daily water discharge is at most 14 cm^3/s. i.e. as a mathematical expression: $P(X \leq 14)$.

Recall the R commands mentioned above. For the Uniform distribution, the R commands are:

```
<d, p, q, r>unif( x, min, max ),
```

where min = a and max = b.

dunif can be used to produce density values for the uniform distribution instead of calculating it by integrating the probability density function. Let's use this to visualize the probability first. (The R commands are given in the appendix in section 3.)

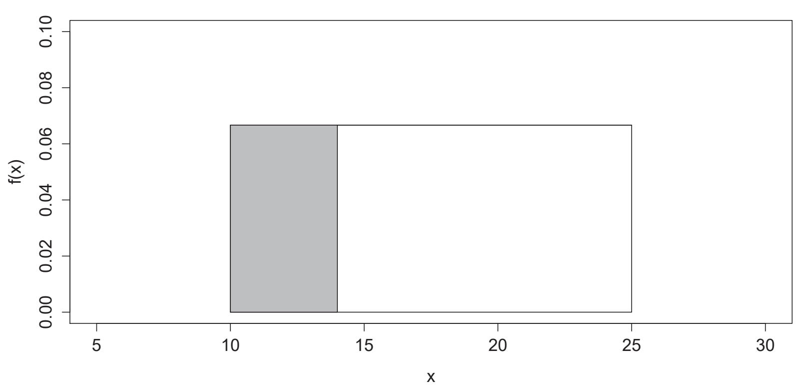

Figure 2: The probability that the daily water discharger is at most 14 cm^3/s, $P(X \leq 14)$.

Now, $P(X \leq 14)$ can be calculated using R as follows:

```
punif(14, min=10, max=25)
```

Are the results you obtain by using integration and using R the same?

Note that since X is a continuous random variable, $P(X \leq 14) = P(X < 14)$.

(b) Suppose that you want to find the probability that on any given day during the fall, the daily water discharge is strictly between 14 and 20 cm^3/s. This can be written as $P(14 < X < 20)$.

Note that since X is a continuous random variable,

$$P(14 < X < 20) = P(14 \leq X \leq 20) = P(14 \leq X < 20) = P(14 < X \leq 20).$$

Let us take a look at the probability visually (The R commands are given in the appendix):

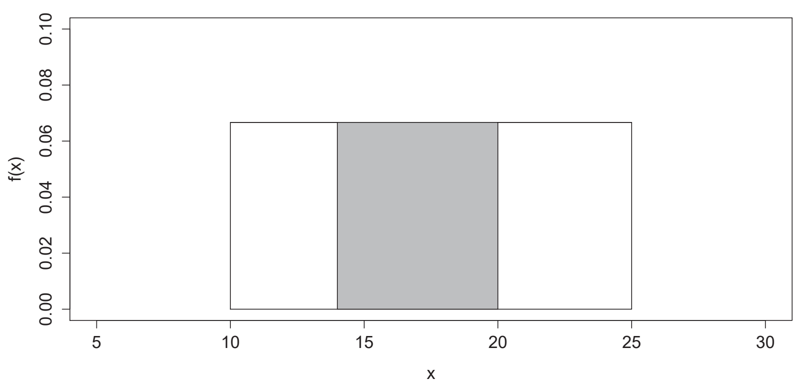

Figure 3: The probability that the daily water discharge is strictly between 14 and 20 cm^3/s, $P(14 < X < 20)$.

Now, $P(14 < X < 20)$ can be calculated using R as follows:

```
punif(20, min=10, max=25)- punif(14, min=10, max=25)
```

Check whether the results you obtain by using integration and using R are the same.

2.2 Find Quartiles

Suppose that you want to find the first quartile, the median (second quartile), and the third quartile of the daily water discharge, i.e., find q_1, q_2, q_3 such that $P(X \leq q_1) = 0.25, P(X \leq q_2) = 0.5$, and $P(X \leq q_3) = 0.75$.

The quartiles can be calculated using the following R commands:

```
qunif(0.25, min=10, max=25)         # q1
qunif(0.5,  min=10, max=25)         # q2
qunif(0.75, min=10, max=25)         # q3
```

2.3 Simulating Random Draws

Suppose you need to take several sample measurements of the daily water discharge during the fall. We can simulate this by drawing random samples from the Uniform distribution on the interval $[10, 25]$.

For example, a random sample of size 10000 can be drawn using the following R command:

```
a <- 10      # minimum
b <- 25      # maximum
x <- runif(n=10000, min = a, max = b)
```

We can represent the probability density of the sample using the following R command:

```
hist(x, prob=T)
curve(dunif(x, min=a, max=b), from=a, to=b, add=T)       # Figure 4
```

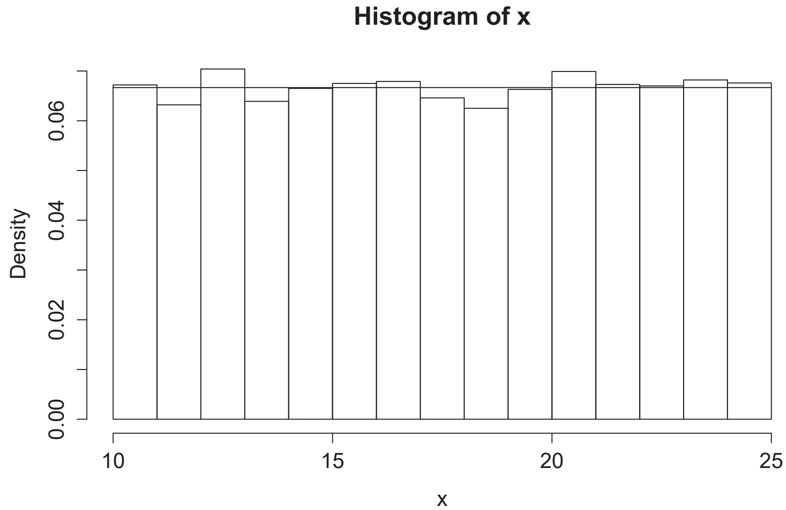

Figure 4: The histogram of the random sample with the true Uniform density curve.

In Figure 4, the pattern of the histogram of the sample is very close to the true uniform density function. Therefore, the sample is very representive of its true distribution.

What if the sample size is very small, say $n = 5$? Is the sample still representative of its true distribution?

3 Appendix of R Commands (Optional Reading)

The R Commands for Creating Figure 1:

```
x <- seq(10, 25, length=200)
prob <- rep(1/(25-10), 200)
plot(x, prob, type="l", xlab="x", ylab="f(x)", xlim=c(5,30),ylim=c(0,0.1))
polygon(c(10, x, 25), c(0, prob, 0))
```

The R Commands for Creating Figure 2:

```
x <- seq(10, 25, length=200)
prob <- dunif(x, min=10, max=25)
plot(x, prob, type="l", xlim=c(5,30), ylim=c(0,0.1), ylab="f(x)")
polygon(c(10, x, 25), c(0, prob, 0))
x <- seq(10, 14, length=100)
prob <- dunif(x, min=10, max=25)
polygon(c(10, x, 14), c(0, prob, 0), col="gray")
```

The R Commands for Creating Figure 3:

```
x <- seq(10, 25, length=200)
prob <- dunif(x, min=10, max=25)
plot(x, prob, type="l", xlim=c(5,30), ylim=c(0,0.1), ylab="f(x)")
polygon(c(10, x, 25), c(0, prob, 0))
x <- seq(14, 20, length=100)
prob <- dunif(x, min=10, max=25)
polygon(c(14, x, 20), c(0, prob, 0), col="gray")
```