All students should **review** §6.1–6.1.1, **read** §6.2–6.2.1 and §6.3.2, and scan through §6.5 paying particular attention to algorithms 6.5.1, 6.5.4–6.5.6 before answering their assigned questions in this assignment. **Keep in mind:** although the text does not use the term, algorithm 6.5.6 is colliquilly known as *"reservior sampling"*. It is an **important algorithm** in data science and simulation, and I'll refer to 6.5.6 as simply *reservior sampling* in lecture, assessments (quizzes and exams) and other written material such as learning goals.

The following numbered questions should be split across your group and the solutions discussed during the next lecture period. Students should review the learning goals for the day, determine which are applicable to their questions and provide answers or commentary to their group members.

When using the Internet to formulate answers (some questions may require this), keep track of **where** you find your information on the web. You may be asked for, and are expected to have (in Email-able form), URLs supporting your investigations.

1. **(For aficionados of math!)**

   (a) Question 6.2.3 (§6.2.4). Be prepared to explain to your group how you determined the supporting set (aka the domain) of this discrete probability distribution.

   (b) Question 6.2.5 (§6.2.4); **Hint:** attack (b) like $f(x)$ was a continuous random variable, then use floor ($\lfloor\cdot\rfloor$) or ceiling ($\lceil\cdot\rceil$) to convert the result to integers. Show numerically or graphically that your solution works for $n = 5$ and u of $\frac{1}{10}$, $\frac{2}{5}$, and $\frac{5}{6}$.

2. (a) Suppose $\mathcal{X}$ is a discrete random variable with possible values $\{a, a+1, \ldots, b\}$ (a, b finite). In your favorite language, design and implement a *binary search* algorithm for $F^*(u)$ for all $u \in (0,1)$.

   (b) Write a wrapper utility that reads a cdf data file with an x_i and $F(x_i)$ on each line:

   ```
   a    F(a)
   a+1  F(a+1)
   a+2  F(a+2)
   :
   b    1
   ```

   Using `binomial-100-3.cdf` , demonstrate that your code works by generating 1000 values and generating two plots:

   i. A **discrete data histogram** (which should look like Figure 6.3.2 of the text),

   ii. A CDF of the values, plotted alongside the data of `binomial-100-3.cdf` .

   Notice that by design, the contents of `binomial-100-3.cdf` defines a discrete CDF.

3. §6.3.2 of the text describes the technique of **constrained inversion** for truncated random variates; complete the following questions and tasks with your insights from the reading.

x_i	$f(x_i)$
3	0.05
4	0.11
5	0.15
6	0.16
7	0.13
8	0.10
9	0.09
10	0.09
11	0.05
12	0.02
13	0.02
14	0.00
15	0.02
16	0.00
17	0.00
18	0.00
19	0.01

   (a) Why are *truncated probability distributions* needed in computer simulations?

   (b) Given the discrete probability distribution at the left, work through the details of using constrained inversion to create a random variate for the distribution over the supporting set $\mathcal{X}_t = \{5,6,\ldots,12\}$.

   (c) Draw (by hand or computer, but make them good!) several graphs showing the geometric interpretation of the algorithmic steps in your answer to part b.

   (d) Be prepared to review and explain your work when your learning group reconvenes next lecture.

4. Suppose you have T items and would like an unbiased random sample of size $T-1$. An easy way to do this is to simply throw one element out in an unbiased manner (choose the loser with $Equilikely(1,T)$).

   Consider another way to achieve a similar result:

   1. Collect the first $T-1$ items and put them aside (call them $\mathcal{A}$); call the remaining T^{th} item Z.
   2. Flip a $\frac{T-1}{T}$ bias coin (biased towards success, "heads").
   3. If the flip is **tails**, then the random sample is simply $\mathcal{A}$ set aside in step 1, and you are done..
   4. If the flip is **heads**, then choose an item from $\mathcal{A}$ in an unbiased manner ($Equilikely(1,T-1)$ springs to mind), and **swap** Z with this item. Now Z is in $\mathcal{A}$, and the size of $\mathcal{A}$ is still $T-1$. Call $\mathcal{A}$ the sought after random sample, and you are done.

   Using basic laws of probability (**not theorems or algorithms from the book**), show that the probability of having any particular sample is the same between the first and second methods; and in fact all possible samples have the same probability of occurring regardless of the method used. **Hint:** Equivalently, and arguably easier, you can show that each item 1 through T has the same probability of **not being chosen for the sample**, regardless of method used.