QBS 120 - Problem Set 1 Solutions

Rob Frost

Grading: Problems Rice 1.75 and Rice 2.52. 6 points for 1.75, 4 points for 2.52. See grading instructions for each problem for specifics.

Scientists discover the life form in the clouds of Venus that is responsible for unusual levels of phosphine. To their surprise the genetic material is also encoded using DNA but with five possible bases (A,C,T,G,B) rather than 4 for life on Earth (A,C,T,G). Also, this life form uses 2 base codons rather than 3 base codons. Assume the values of the two independent bases are measured. Compute answers to the following using R:

Note that there are many different ways to generate the solutions via R. For these questions, R is not needed (i.e., the answer is obvious by inspection) but computing the solution is good practice.

- **How many unique codons are possible? How does this compare the number of unique 3 base codons for Earth DNA?**

  Because the bases are independent, the number of possible values can be found using a simple application of the multiplication principle:

```
>         (num.alien.codons = 5*5)

[1] 25
```

  A similar calculation can be done for the number of standard DNA:

```
>         (num.earth.codons = 4*4*4)

[1] 64
```

- **What are the sequences of these unique 2 base codons?**

  To generate Ω, I decided to use expand.grid(), a useful function.

```
>         bases = c("A", "C", "T", "G", "B")
>         (omega = expand.grid(b1=bases, b2=bases))

   b1 b2
1   A  A
2   C  A
3   T  A
4   G  A
5   B  A
6   A  C
```

```
7   C  C
8   T  C
9   G  C
10  B  C
11  A  T
12  C  T
13  T  T
14  G  T
15  B  T
16  A  G
17  C  G
18  T  G
19  G  G
20  B  G
21  A  B
22  C  B
23  T  B
24  G  B
25  B  B
```

- **List the elements of the event E_1 that both bases are the same.**

  Easy to compute on the output from expand.grid(); apply() and which() are essential R functions.

```
> omega$same = apply(omega, 1, function(x) {
+         if (x[1] == x[2]) {
+                 return (T)
+         }
+         return (F)
+ })
> in.E.1 = which(omega$same)
> omega[in.E.1,]

   b1 b2 same
1   A  A TRUE
7   C  C TRUE
13  T  T TRUE
19  G  G TRUE
25  B  B TRUE
```

- **List the elements of the event E_2 that both bases are different.**

  Can use the output from the last question:

```
> in.E.2 = which(!omega$same)
> omega[in.E.2,]

   b1 b2  same
2   C  A FALSE
```

```
3   T  A FALSE
4   G  A FALSE
5   B  A FALSE
6   A  C FALSE
8   T  C FALSE
9   G  C FALSE
10  B  C FALSE
11  A  T FALSE
12  C  T FALSE
14  G  T FALSE
15  B  T FALSE
16  A  G FALSE
17  C  G FALSE
18  T  G FALSE
20  B  G FALSE
21  A  B FALSE
22  C  B FALSE
23  T  B FALSE
24  G  B FALSE
```

- **List the elements of the event E_3 that the first base is A.**

  Similar logic:

```
> omega$firstA = apply(omega, 1, function(x) {
+         if (x[1] == "A") {
+                 return (T)
+         }
+         return (F)
+ })
> in.E.3 = which(omega$firstA)
> omega[in.E.3,]

   b1 b2  same firstA
1   A  A  TRUE   TRUE
6   A  C FALSE   TRUE
11  A  T FALSE   TRUE
16  A  G FALSE   TRUE
21  A  B FALSE   TRUE
```

- **List the elements of the event $E_1 \cap E_3$.**

  Use the R set functions:

```
> E.1.intersect.E.3 = intersect(in.E.1, in.E.3)
> omega[E.1.intersect.E.3,]

  b1 b2 same firstA
1  A  A TRUE   TRUE
```

- **List the elements of the event $E_1 \cup E_3$.**

```
> E.1.union.E.3 = union(in.E.1, in.E.3)
> omega[E.1.union.E.3,]

   b1 b2  same firstA
1   A  A  TRUE   TRUE
7   C  C  TRUE  FALSE
13  T  T  TRUE  FALSE
19  G  G  TRUE  FALSE
25  B  B  TRUE  FALSE
6   A  C FALSE   TRUE
11  A  T FALSE   TRUE
16  A  G FALSE   TRUE
21  A  B FALSE   TRUE
```

Five people want to play a card game so they discard the two cards from a 52 card deck and are each dealt 10 cards. How many unique deals are possible? If there were four playing instead and each was dealt 13 cards, how many unique deals are possible? Before computing, which number do you think will be larger?

Need to apply Proposition C from section 1.4.2. Number of possible five person deals = $50!/(10!^5)$. Similarly, the number of possible four person deals = $52!/(13!^4)$. For the 5 person case, the numerator is larger and the denominator should be smaller so my guess is that there will be many more 5 person deals.

Compute using R:

```
> (num.5.person.deals = factorial(50)/(factorial(10)^5))

[1] 4.833478e+31

> (num.4.person.deals = factorial(52)/(factorial(13)^4))

[1] 5.364474e+28

> num.5.person.deals/num.4.person.deals

[1] 901.0162
```

A very large number of possible deals in both cases but substantially more for the 5 person vs. 4 person game!

Rice, Chapter 1, Problem 18: A lot of n items contains k defectives, and m are selected randomly and inspected. How should the value of m be chosen so that the probability that at least one defective item turns up is 0.80? Apply your answer to (a) $n = 1000, k = 10$ and (b) $n = 1000, k = 100$.

We want to approach in a similar manner to Example H in section 1.4.2. The first trick to solving this is to realize that the complement (no defects are discovered) is much easier to compute than the goal event (one or more defects discovered). So, we'll employ Property A from section 1.3 ($P(A) = 1 - P(A^C)$) to solve the problem. To find the probability of zero defects, we'll use fact that if the outcomes are all equally likely, the probability for a given event is the ratio of the # of

ways the event can occur to the total number of possible outcomes. In this case, the number of ways to select no defects is, by the multiplication principle, the product of # of ways to select m items from $n-k$ non-defects and the # of ways to select no items from the k defects. Since this is sampling without replacement, this can be represented as $\binom{n-k}{m}\binom{k}{0}$ which equals $\binom{n-k}{m}$ since $\binom{k}{0}$ is 1 by definition. The total number of outcomes is $\binom{n}{m}$. If a is the number of defects detected, we thus have:

$$P(a \geq 1) = 1 - \frac{\binom{n-k}{m}}{\binom{n}{m}}$$

We are asked to find m such that $P(a \geq 1) \geq 0.8$. We won't attempt to find a closed form solution for m in terms of n and k but will instead compute $P(a \geq 1)$ for a range of m value for the given n and k values and select the first m such that $P(a \geq 1) \geq 0.8$.
First, let's write an R function to compute $P(a \geq 1)$ for a given n, k and m:

```
> probAtLeastOneDefect = function(n, k, m) {
+         prob = 1 - choose(n-k, m)/choose(n, m)
+         return (prob)
+ }
```

a) n=1000, k=10

We'll pick a wide range of possible m values to ensure we find the first m for which the $P(a \geq 1) \geq 0.8$.

```
> m.vals = 1:1000
> probs = probAtLeastOneDefect(n=1000, k=10, m=m.vals)
> (m = which(probs >= 0.8)[1])

[1] 148

> m.and.probs = data.frame(m=m.vals, prob=probs)
> m.and.probs[(m-2):m,]

      m      prob
146 146 0.7952545
147 147 0.7976520
148 148 0.8000241
```

So, $m = 148$.

b) n=1000, k=100

```
> probs = probAtLeastOneDefect(n=1000, k=100, m=m.vals)
> (m = which(probs >= 0.8)[1])

[1] 16

> m.and.probs = data.frame(m=m.vals, prob=probs)
> m.and.probs[(m-2):m,]
```

```
    m      prob
14 14 0.7735555
15 15 0.7965215
16 16 0.8171792
```

In this case, $m = 16$. This makes sense, it should require many fewer samples to ensure at least one defect is detected if there are an order of magnitude more defects.
Note that to get stable numerical values for the original case b), $\frac{\binom{n-k}{m}}{\binom{n}{m}}$ needs to be simplified or approximated to prevent R from generating infinite values via choose().

Rice, Chapter 1, Problem 49: Two fair coins are simultaneously tossed three times.

a. What is the probability of two or more heads given that there was at least one head?

The use of two coins flipped together is a bit of a trick. The flips are all independent so this is no different that one coin flipped six times. We'll solve it using that formulation. Question a) is a conditional probability problem. Let A be the event of two or more heads and let B be the event of at least one head. We are asked to find $P(A|B)$. Although one could solve this by inspection, let's compute using the definition of conditional probability:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Since $A \subset B$ (if we have two or more heads we automatically have at least one head), we know that $P(A \cap B) = P(A)$. So, simplify:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A)}{P(B)}$$

Since each coin toss is independent, we have $2^6 = 64$ total possible outcomes. Since each outcome is equally likely, P(B) and P(A) can be computed as the ratio of the number of outcomes in A or B to 64. For this problem, we could simply enumerate all 64 outcomes and count those in A or B:

```
>         omega = expand.grid(c1=c("H", "T"), c2=c("H", "T"), c3=c("H", "T"), c4=c("H", "T"),
+                 c5=c("H", "T"), c6=c("H", "T"))
>         head(omega)

  c1 c2 c3 c4 c5 c6
1  H  H  H  H  H  H
2  T  H  H  H  H  H
3  H  T  H  H  H  H
4  T  T  H  H  H  H
5  H  H  T  H  H  H
6  T  H  T  H  H  H
```

...
But this is a bit tedious so we'll use properties of the complement and symmetry. To identify the number of outcomes in B, we observe that B^C only includes just a single outcome $\{t, t, t, t, t, t\}$ so

B includes the other 63 outcomes and $P(B) = 63/64$. To identify the number of outcomes in A, we observe that A^C includes all outcomes with at least five tails. This includes $\{t, t, t, t, t, t\}$ and the events with a single head, which can occupy one of 6 spaces, so there are 7 total events in A^C and $P(A) = 1 - P(A^C) = 1 - 7/64 = 57/64$
Thus:

$$P(A|B) = \frac{P(A)}{P(B)} = \frac{57/64}{63/64} = \frac{57}{63}$$

b. What is the probability of two or more heads given that there was at least one tail?

Here, B is the event with at least one tail. Again, we could simply look at all 64 outcomes and enumerate the answer but we'll again use properties of symmetry and complements to determine the size of $A \cap B$ and B by inspection. For B, we know by symmetry that the number of outcomes with at least one tail must be the same as the number with at least one head so again $P(B) = 63/64$. For $A \cap B$, we know that only one outcome in A, $\{h, h, h, h, h, h\}$ is not also in B so the number of outcomes in $A \cap B$ equals the number in A - 1 or 56 and $P(A \cap B) = 56/64$.
Thus:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{56/64}{63/64} = \frac{56}{63}$$

Rice, Chapter 1, Problem 75: A population starts with one member; at a time $t = 1$, it either has 1 progeny with probability p, 2 progeny with probability $2p$, or dies. Also, we know that $3p < 1$. If it successfully reproduces, then its children behave independently with the same alternatives at time $t = 2$. What is the probability that there are no members in the third generation? For what value of p is this probability equal to 0.5?

Let A_1 be the event that the first individual has 1 progeny at $t = 1$, let A_2 be the event that the first individual has 2 progeny at $t = 1$, let A_3 be the event that the first individual dies with no progeny at $t = 1$, and let B be the event that all members of the population die at $t = 2$. We're asked to solve for $P(B)$. Although we don't know $P(B)$ directly, we can compute the conditional probabilities of B given the various A events. So, use the Law of Total Probability to define $P(B)$ in terms of $P(A_1), P(A_2), P(A_3), P(B|A_1), P(B|A_2)$, and $P(B|A_3)$:

$$P(B) = P(B|A_1)P(A_1) + P(B|A_2)P(A_2) + P(B|A_3)P(A_3)$$

(Grading: 1 point for defining appropriate events and associated probabilities)
(Grading: 1 point to use the Law of Total Probability)

Note that this works since A_1, A_2 and A_3 are disjoint and $A_1 \cup A_2 \cup A_3 = \omega$. From the problem statement and inspection we know that:

- $P(A_1) = p$
- $P(A_2) = 2p$

- $P(A_3) = 1 - P(A_1) - P(A_2) = 1 - 3p$ (by Property A since A_3 is the complement of $A_1 \cup A_2$)
- $P(B|A_3) = 1$ (if the initial individual dies at t=1 then there can be no population at t=2)

To find $P(B|A_1)$, and $P(B|A_2)$, recognize that B requires that all progeny of the first individual die. Since each die with probability $1-3p$ and their deaths are independent, the probability that all die is $1-3p$ if A_1 happens and $(1-3p)^2$ if A_2 happens. So, $P(B|A_1) = 1-3p$ and $P(B|A_2) = (1-3p)^2$. Combining all of these gives:

$$\begin{aligned} P(B) &= P(B|A_1)P(A_1) + P(B|A_2)P(A_2) + P(B|A_3)P(A_3) \\ &= (1-3p)p + (1-3p)^2 2p + 1 - 3p \\ &= p - 3p^2 + 2p(1 - 6p + 9p^2) + 1 - 3p \\ &= p - 3p^2 + 2p - 12p^2 + 18p^3 + 1 - 3p \\ &= -15p^2 + 18p^3 + 1 \end{aligned}$$

(Grading: 2 pts to specify the correct polynomial solution. Half credit if there is an error in the form of the polynomial.)

We'll numerically compute the value of p such that $P(B) = 0.5$ using the R function polyroot(). Note that if $P(B) = 0.5$, the equation becomes: $18p^3 - 15p^2 + 0.5 = 0$ and we can then solve for the zero that is between 0 and 1:

```
>         (roots = polyroot(c(0.5, 0, -15, 18)))

[1]  0.2113249+0i -0.1666667+0i  0.7886751-0i
```

Although two of these are valid probabilities, only one satisfies the constraint that $3p < 1$.
Could also just find the approximate value of p by computing P(B) for a range of p values between 0 and 1:

```
>         prob.B = function(x) {
+                 return (1-15*x^2 + 18*x^3)
+         }
>         p.vals = seq(from=0, to=1, by=0.05)
>         p.B = sapply(p.vals, function(x) {
+                 return (prob.B(x))
+         })
>         names(p.B) = p.vals
>         p.B

      0    0.05     0.1    0.15     0.2    0.25     0.3    0.35
1.00000 0.96475 0.86800 0.72325 0.54400 0.34375 0.13600 -0.06575
    0.4    0.45     0.5    0.55     0.6    0.65     0.7    0.75
-0.24800 -0.39725 -0.50000 -0.54275 -0.51200 -0.39425 -0.17600 0.15625
    0.8    0.85     0.9    0.95       1
0.61600 1.21675 1.97200 2.89525 4.00000
```

(Grading: 2 points to identify the exact answer via polyroot or approximate answer through some numerical simulation)

What is a random variable? Define both discrete and continuous RVs in terms of the sample space.

A random variable is a function that maps from the sample space of events to the real numbers. For discrete RVs, the range of the function is countably infinite (typically it is finite). For continuous RVs, the range is infinite.

One of the observed values in an experiment is 13.4, which is modeled as a RV during analysis. How can a fixed number be modeled as a RV? Where is the random component?

The specific value of 13.4 represents one realization of the RV, i.e., the value of the function for an event in the sample space. If the experiment were repeated, the event would change in a random fashion generating a random observed value.

Rice, Chapter 2, Problem 9: For what values of p is a 2 out of 3 majority decoder better than transmission of the message once?

- We can treat each message transmission as a Bernoulli RV with parameter p. Of course, the probability that direct transmission of the message is successful is p.
- For the majority decoder, we have three independent transmissions so the number of successes is binomial with n=3 and p. We'll represent the number of successes for the majority decoder as binomal RV X.
- The probability of successful transmission with the majority decoder, P_D, is:

$$\begin{aligned} P_D &= P(X=2) + P(X=3) \\ &= p_X(2) + p_X(3) \\ &= \binom{3}{2} p^2 (1-p)^{3-2} + \binom{3}{3} p^3 (1-p)^{3-3} \\ &= 3p^2(1-p) + p^3 \end{aligned}$$

- We want to find the value of p such that $p < P_D$ with both p and P_D bound between 0 and 1.

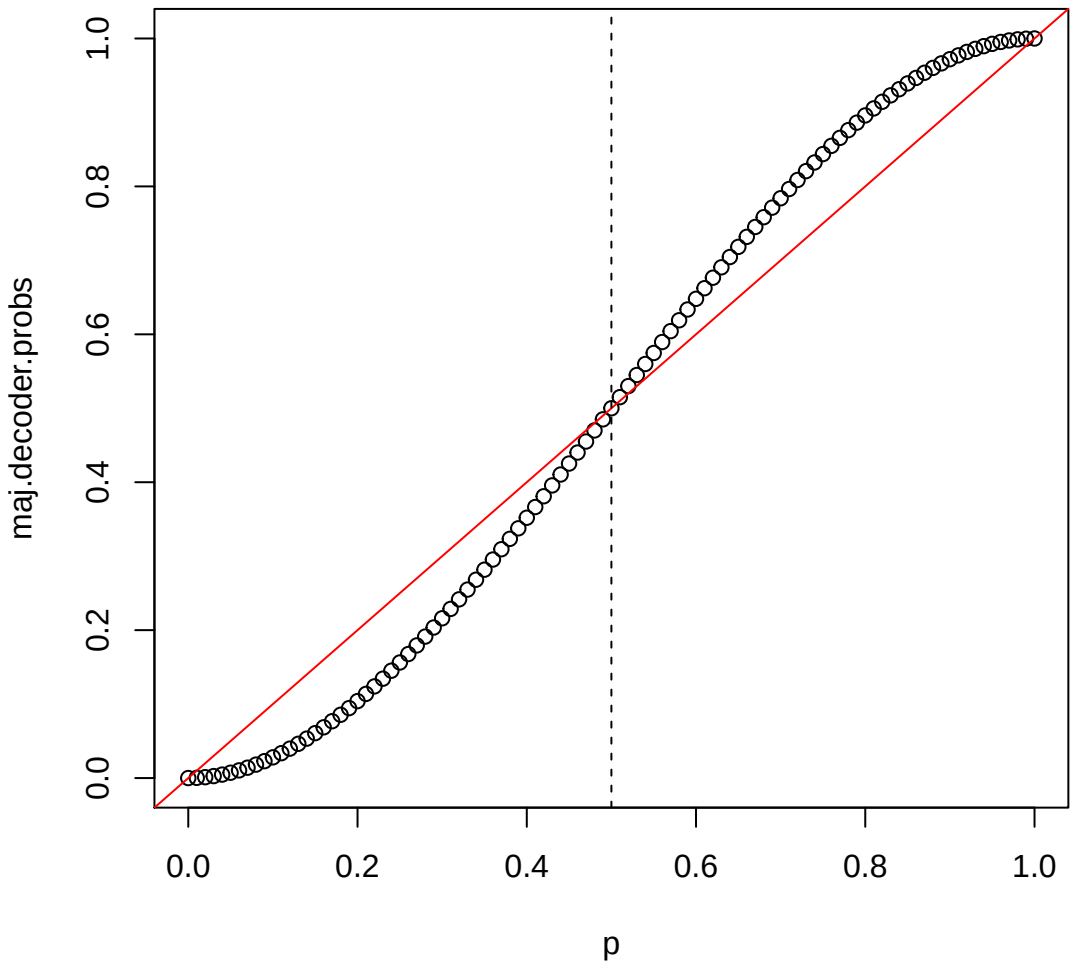
$$\begin{aligned} p &< 3p^2(1-p) + p^3 \\ 0 &< -2p^3 + 3p^2 - p \\ 0 &< -2p^2 + 3p - 1 \\ 0 &< (-2p+1)(p-1) \end{aligned}$$

  The equation is solved exactly for $p = 1$ or $p = 0.5$. Our possible solution ranges are $p < 0.5$ or $0.5 < p < 1$. A check indicates that only $p > 0.5$ satisfies the inequality (if p < 0.5, the first term is positive and the second negative making the entire RHS negative, which violates the inequality).

So, the 2 out of 3 majority decoder is superior to direct transmission for $p > 0.5$.

Confirm through visualization
To confirm, we'll plot the probability of the decoder and direct transmission succeeding vs p.

```
>         p = seq(from=0, to=1, by=0.01)
>         maj.decoder.probs = sapply(p, function(x) {
+                 return (3*x^2*(1-x) + x^3)
+         })
>         plot(p, maj.decoder.probs, type="b")
>         abline(coef = c(0,1), col="red")
>         abline(v=0.5, lty="dashed")
```



The region where the major decoder prob is above the diagonal is the solution region, confirming our answer of $p > 0.5$.

Rice, Chapter 2, Problem 40: Suppose X has the density function $f(x) = cx(x+2)$ for $0 \leq x \leq 2$ and $f(x) = 0$ otherwise.

A **Find c.**

We can find c by realizing that the density must integrate to 1 over $0 \leq x \leq 2$:

$$\begin{aligned}
\int_0^2 cx(x+2)dx &= 1 \\
c\int_0^2 x^2 + 2xdx &= 1 \\
c[x^3/3 + x^2]_0^2 &= 1 \\
c(8/3+4) &= 1 \\
c &= 3/20
\end{aligned}$$

B **Find the cdf.**

By definition, the CDF for X is:

$$\begin{aligned}
F_X(x) &= \int_{-\infty}^{x} f_X(u)du \\
&= \int_0^x 3/20(u^2+2u)du && \text{use range of x with non-zero f(x)} \\
&= 3/20[u^3/3 + u^2]_0^x \\
&= 3/20(x^3/3 + x^2)
\end{aligned}$$

The CDF for $x < 0$ is 0 and for $x > 2$ is 1. So, can express the CDF as:

$$F_X(x) = \begin{cases} 0 & x < 0 \\ 3/20(x^3/3 + x^2) & 0 \leq x \leq 2 \\ 1 & x > 2 \end{cases}$$

C **What is $P(0.1 \leq x \leq 0.5)$?.**

This can be expressed as the difference of CDF values at 0.5 and 0.1:

$$\begin{aligned}
P(0.1 \leq x \leq 0.5) &= F_X(0.5) - F_X(0.1) \\
&= 3/20(0.5^3/3 + 0.5^2 - 0.1^3/3 - 0.1^2)
\end{aligned}$$

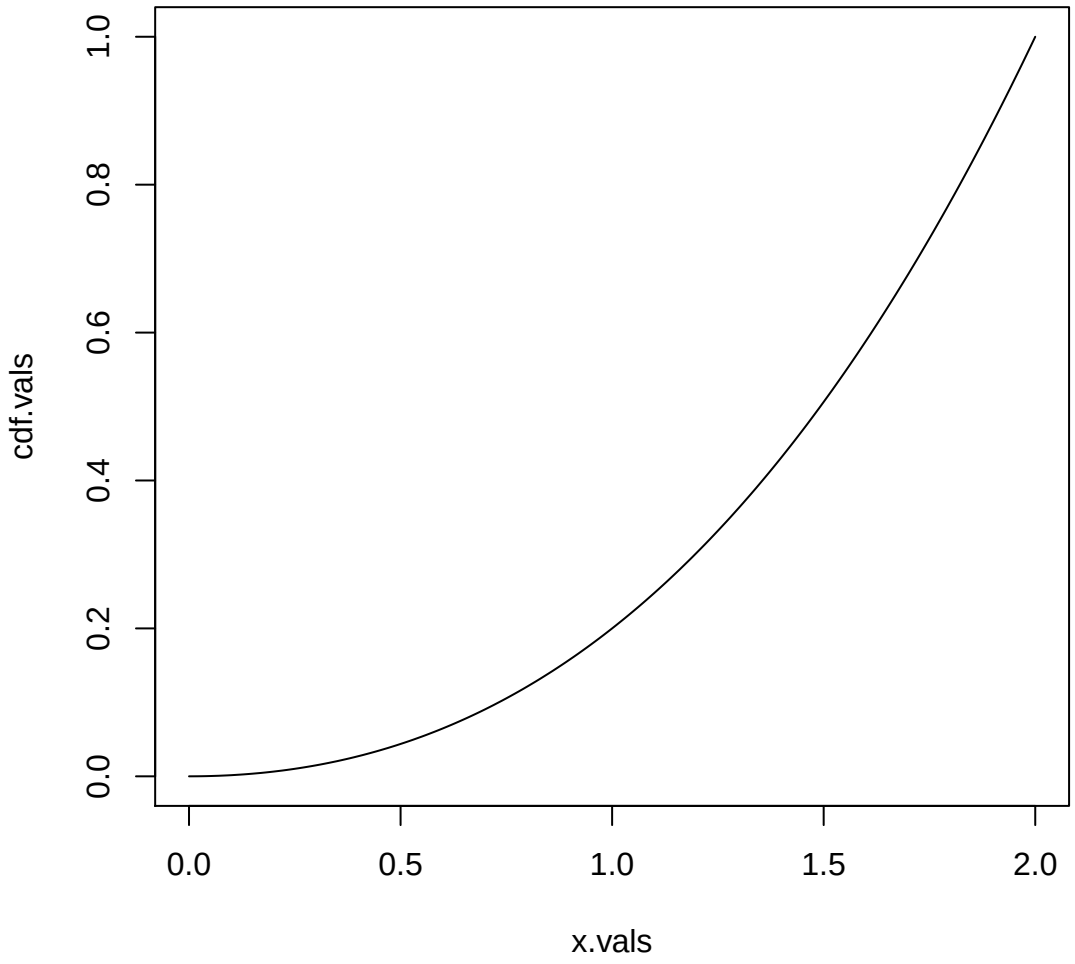
Let's compute via R:

```
> (prob_0.1_0.5 = (3/20)*(0.5^3/3 + 0.5^2 - 0.1^3/3 - 0.1^2))

[1] 0.0422
```

D **Plot the CDF**

```
> x.vals = seq(from=0, to=2, by=0.01)
> cdf.vals = sapply(x.vals, function(x) {
+         return ((3/20)*(x^3/3 + x^2))
+ })
> plot(x.vals, cdf.vals, type="l")
```



Rice, Chapter 2, Problem 52: Suppose that in a certain population, individual's heights are approximately normally distributed with parameters $\mu = 60$ inches and $\sigma = 4$ inches.

A **What proportion of the population is over 6 ft tall?**

Let X be the RV representing height. We're asked to find $P(X > 72)$. If we were free to just call pnorm() with any parameters, this would be easy:

```
> 1- pnorm(72, mean=60, sd=4)
```

```
[1] 0.001349898
```

However, I asked that only functions for standard normal RVs could be used so we must transform into a $\mathcal{N}(0,1)$ RV. We know that for normal RV X with parameters μ and σ, $(X-\mu)/\sigma$ is standard normal. So, we can frame the problem as:

$$\begin{aligned} P_X(X>72) &= P_X((X-60)/4 > (72-60)/4) && \text{Transform both sides} \\ &= P_Z(Z>3) && \text{For Z standard normal} \\ &= 1-\Phi(3) && \text{Where } \Phi \text{ is the standard normal CDF} \end{aligned}$$

(Grading: 1 point for attempting the transformation into standard normal.)

And this we can solve using the default pnorm() (for $\mathcal{N}(0,1)$):

```
> 1- pnorm(3)
```

```
[1] 0.001349898
```

(Grading: 1 point for getting the correct answer.)

B **What is the distribution of heights if they are expressed in centimeters?**

In this case we are making a scaling transformation of X: C = 2.54 X. We know from the text that for Y=bX with $X \sim \mathcal{N}(\mu, \sigma^2)$, Y has the distribution $\mathcal{N}(b\mu, b^2\sigma^2)$. So the height in centimeters has the distribution:

$$\mathcal{N}(60 * 2.54, 2.54^2 * 4^2) = \mathcal{N}(152.4, 103.2256)$$

(Grading: 1 point.)

C **In meters?**

For meters, M = 0.0254 X and the distribution is:

$$\mathcal{N}(60 * 0.0254, 0.0254^2 * 4^2) = \mathcal{N}(1.524, 0.01032256)$$

(Grading: 1 point.)