Introduction to Probability and Information Theory
Homework #6

1. (a) $H(X) = -4 \times \left(\frac{1}{4} \times \log_2 \frac{1}{4}\right) = \log_2 4 = 2$ bits
   (b) $H(X) = 2$ bits
   (c) $H(X) \approx 1.997$ bits
   (d) $H(X) = \log_2 8 = 3$ bits
2. (a) $H(X) = log_2 6 = 2.59$ bits
   (b) If X is the result of one roll, then $H(X) = \log_2 8 = 3$. Since Y is the joint distribution of four independent rolls:

$$\begin{aligned} H(Y) &= H(X) + H(X) + H(X) + H(X) \\ &= 4H(X) \\ &= 12 \text{ bits} \end{aligned}$$

3. Starting with the definition of KL divergence:

$$\begin{aligned} D(P(X,Y)||P(X)P(Y)) &= \sum_{x \in X, y \in Y} p(x,y) \log_2 \frac{p(x,y)}{p(x)p(y)} \\ &= I(P(X,Y); P(X)P(Y)) \end{aligned}$$

4. This one is tricky, but a typo in the original assignment makes it even harder. The correct series identities are:

$$\begin{aligned} \sum_{n=0}^{\infty} r^n &= \frac{1}{1-r} \\ \sum_{n=0}^{\infty} n r^n &= \frac{r}{(1-r)^2} \end{aligned}$$

   Sorry about that.

   Okay, now we can get going on the problem. X is geometrically distributed, so $P(X = x) =$

$(1-\frac{1}{2})^x \times \frac{1}{2} = (\frac{1}{2})^{x+1}$. That means that:

$$
\begin{aligned}
H(X) &= -\sum_{x=0}^{\infty} \left(\frac{1}{2}\right)^{x+1} \log_2 \left(\frac{1}{2}\right)^{x+1} \\
&= -\sum_{x=0}^{\infty} \left(\frac{1}{2}\right)^{x+1} (x+1) \log_2 \frac{1}{2} \\
&= \sum_{x=0}^{\infty} \left(\frac{1}{2}\right)^{x+1} (x+1) \\
&= \frac{1}{2} \sum_{x=0}^{\infty} \left(\frac{1}{2}\right)^{x} (x+1) \\
&= \frac{1}{2} \left(\sum_{x=0}^{\infty} x \left(\frac{1}{2}\right)^{x} + \sum_{x=0}^{\infty} \left(\frac{1}{2}\right)^{x} \right)
\end{aligned}
$$

Now we can use the (correct) identities that were given as a hint:

$$
\begin{aligned}
H(X) &= \frac{1}{2} \left(\frac{\frac{1}{2}}{(1-\frac{1}{2})^2} + \frac{1}{1-\frac{1}{2}} \right) \\
&= \frac{1}{2} \left(\frac{4}{2} + \frac{2}{1} \right) \\
&= 2 \text{ bits}
\end{aligned}
$$

Since X can take on any non-negative value, we can't verify this by numerically evaluating the formula for entropy directly. But, since the series converges (because $p(x)$ gets very small as x gets very big), we can get a good estimate of the entropy by summing the first, say, 100 values using R:

```
> -sum(dgeom(0:100,.5)*log(dgeom(0:100,.5),2))
[1] 2
```

Looks good!