Homework 2

1. (book #2.2.) A monkey types on a 26-letter keyboard that has lowercase letters only. Each letter is chosen independently and uniformly at random from the alphabet. If the monkey types $n = 1,000,004$ letters, what is the expected number of times the sequence "proof" appears?

   The expected number of times is equal to (up to 2 decimals is fine): $\boxed{}$

2. (book #2.4) Prove that $\underbrace{E[X^n]}_{L} \geq \underbrace{(EX)^n}_{R}$ for $n = 2k$.

   Which of the following arguments proves the claim?

   (a) $f''(x) = n(n-1)x^{2(k-1)} \geq 0 \implies L \geq R$ by Jensen's inequ.
   (b) by the mean value theorem we have $L \geq R$
   (c) $L \geq (E[X^n])^{1/n} = E(X) \geq [E(X)]^n = R$
   (d) $L \geq \mathrm{Var}(X^k) \geq [\mathrm{Var}(X)]^k \geq R$
   (e) none of these

3. (book # 2.14) The geometric distribution arises as the distribution of the number of times we flip a coin until it comes up heads. Consider now the distribution of the number of flips X until the kth head appears, where each coin flip comes up heads independently with probability p. This is known as the negative binomial distribution, and we also write $X \sim NBin(k, p)$. Find $p_X(n)$.
   Hint: Here X counts the number of flips *including* the flip with the k-th head.

   Mark the correct expression for $p_X(n)$, below.
   (a) $\binom{n-1}{k-1} p^k (1-p)^{n-k}$ (b) $\binom{n}{k} p^k (1-p)^{n-k}$ (c) $p^k (1-p)^{n-k}$ (d) $\binom{n}{k} p^{k-1} (1-p)^{n-k}$ (e) none of these

4. (book # 2.15) For a coin that comes up heads independently with probability p on each flip, what is the expected number (X) of flips until the kth heads?

   Hint: here x counts the number of flips *including* the flip with the k-th head.

   Mark the correct expression for $E(X)$:

   (a) $1/p$ (b) kp (c) k/p (d) np/k (e) none of these

5. (book #2.24) We roll a standard fair die over and over. What is the expected number of rolls (N) until the first pair of consecutive sixes appears?

   Hint: Let X_i denote the face on the i-th roll. Set up Lemma 2.5 twice, once for EN, conditioning on X_1; and a second time for $E(N \mid X_1 = 6)$, now (additionally) conditioning on X_2. For the second application of Lemma 2.5, a version of the lemma as: $E(X \mid Y = y, Z = z) = \sum_z E(X \mid Y = y, Z = z) \Pr(Z = z \mid Y = y)$ (that is, use the lemma for the probability function $q(N) = \Pr(N \mid X_1 = 6)$). Solve the two equations in $a = EN$ and $b = E(N \mid X_1 = 6)$.

   $\boxed{EN = }$

6. (book # 2.25) A blood test is being performed for $n = N \cdot k$ individuals by pooling the samples for k people, and analyzing them together. If the test is negative, this pooled test suffices for all k individuals.
   If the test is positive, we test each of the k persons separately, i.e., we carry out a total of $k + 1$ tests.

Suppose that we create N disjoint groups of k people and use the pooling method. Let A_i = i-th person has a positive result, and assume $\Pr(A_i) = p$, independently, $i = 1, \ldots, n$.

6a. Let B_j denote the event of a positive pooled sample, for the j-th pooled sample consisting of samples from a batch of k individuals.

For $p = 0.01$, $k = 4$ and $n = 40$, find $\Pr(B_j)$.

$\boxed{\Pr(B_j) = \quad\quad}$

6b. Let $X = \#$ of tests. What is the expected number of tests necessary, $E(X)$?

For $p = 0.01$, $k = 4$ and $n = 40$, find $E(X)$.

$\boxed{E(X) = \quad\quad}$

6c. Describe how to find the best value of k.

Nothing to turn in for this problem.

7. Prove the Poi approximation of binomial probabilities.

Claim: If $X \sim \text{Bin}(n, p)$ with $np = \lambda$ (or $p = \lambda/n$) and large n, then $p_X(i) \approx e^{-\lambda}\frac{\lambda^i}{i!}$, i.e., approximated by a Poisson probability with rate λ.

Hint: Use $(1 - \lambda/n)^{\,i} \to 1$; $(1 - \lambda/n)^n \to e^{-\lambda}$; and $(n - i + 1)/n \to 1$ as $n \to \infty$. The last assumption implies we use the approximation for moderate i only.

Which of the following arguments justifies the Poisson approximation? All limits are as $n \to infty$.

(a) $\lim np = \lambda$, by assumption

(b) $\lim \binom{n}{i} p^i (1-p)^{(n-i)} = \lim (1 + i/n)^n = e^{-\lambda}\frac{\lambda^i}{i!}$

(c) $\lim \binom{n}{i} p^i (1-p)^{(n-i)} = \lim \frac{n(n-1)\cdots(n-i+1)}{n^i} = e^{-\lambda}\frac{\lambda^i}{i!}$

(d) $\lim \binom{n}{i} p^i (1-p)^{(n-i)} = \lim \frac{n(n-1)\cdots(n-i+1)}{n^i} \frac{\lambda^i}{i!} \frac{(1-\lambda/n)^n}{(1-\lambda/n)^i} = e^{-\lambda}\frac{\lambda^i}{i!}$

(e) none of these

8. Recall the birthday problem which we discussed in Chapter 1. Assume that each of n people is equally likely to have any of the 365 days of the year as his or her birthday. Let A_2 = "no two persons share the same birthday". In Chapter 1 we used a combinatorial argument to find $\Pr(A_2)$. Alternatively we could use a Poisson approximation as follows. Consider all $L = \binom{n}{2}$ possible (unordered) pairs $\{i, j\}$. Let $E_{ij} = 1$ if i and j have the same birthdays and $E_{ij} = 0$ otherwise. Then $p = \Pr(E_{ij} = 1) = 1/365$. Although the E_{ij} are not mutually independent, they are only weakly dependent. In fact, they are pairwise independent. Let $X = \sum_{i<j} E_{ij}$ denote the number of pairs with shared birthdays. Let $Y \sim \text{Bin}(L, p)$. Then $p_X(0) \approx p_Y(0)$, and $p_Y(0)$ in turn we can approximate by a Poisson approximation.

Use a similar approximation to find the probability of A_3 = "no 3 same birthdays among n people". That is, letting $X = \#$ of triples $\{i, j, k\}$ with same birthday, approximate $p_X(0)$.

Hint: Let $K = \binom{n}{3}$ and let $p = (1/365)^2$ and proceed similar to before.

For $n = 50$, find the described approximation for $\Pr(A_3)$:

$\boxed{\Pr(A_3) = \quad\quad}$

9. Prove Lemma 5.2 for two Poi r.v.'s. That is, assuming $X \sim \text{Poi}(\mu_1)$ and $Y \sim \text{Poi}(\mu_2)$, independently, show that $S = X + Y \sim \text{Poi}(\mu_1 + \mu_2)$.

Hint: Use the law of total probability with $E_k = \{X_1 = k\}$ to find $p_S(j) = \sum_{k=0}^{j} \ldots$.

Starting with the law of total probability for $p_S(j)$, and then moving all factors that do not involve j outside the sum we get

$$p_S(j) = \sum_{k=0}^{j} p_{X_1}(k)\, p_{X_2}(j-k) = \sum_{k=0}^{j} \frac{e^{-\mu_1}\mu_1^k}{k!} \cdot \frac{e^{-\mu_2}\mu_2^{j-k}}{(j-k)!} = \frac{e^{-(\mu_1+\mu_2)}}{j!} \sum_{k=0}^{j} \binom{j}{k} \mu_1^k \mu_2^{j-k}$$

Let RHS denote the final right hand side above. Completing the proof from here, which of the following argument proves the claim.

(a) $RHS = \frac{e^{-(\mu_1+\mu_2)}(\mu_1+\mu_2)^j}{j!} \sum_{k=0}^{j} \binom{j}{k} p^k (1-p)^{j-k}$ with $p = \mu_1/(\mu_1+\mu_2)$ and $\sum_k \ldots = 1$.

(b) $RHS = \frac{e^{-(\mu_1+\mu_2)}}{j!}$ since $\sum_{k=0}^{j} \binom{j}{k} \mu_1^k \mu_2^{j-k} = 1$

(c) Use $np = \lambda = \mu_1 + \mu_2$ for $p = \mu_1/(\mu_1+\mu_2)$

(d) $\lim_{j\to\infty} RHS = \frac{e^{-(\mu_1+\mu_2)}(\mu_1+\mu_2)^j}{j!}$

(e) none of these

Simulation

The following problem requires some programming using R, or any comparable programming language

10. Kullback-Leibler divergence (KL) of two distributions p_X and p_Y is defined as

$$D(p_X, p_Y) = \sum_x p_X(x) \log\left(\frac{p_X(x)}{p_Y(y)}\right).$$

KL divergence is a measure of discrepancy between two distributions with $D = 0$ for $p_X = p_Y$ and $D > 0$ for $p_X \neq p_Y$. Using KL divergence, we then define mutual information (MI) for two jointly distributed r.v.'s as

$$I(X, Y) = D(p_{X,Y},\ p_X p_Y),$$

that is KL divergence between the joint distribution $p_{X,Y}$ and the product of the marginals p_X and p_Y. The latter is the hypothetical joint distribution under independence.

In this example we use MI to judge whether two variables X, Y are independent or not. We record n pairs $(X_i, Y_i) \sim p_{X,Y}$, $i = 1, \ldots, n$, independently (data). Here $X_i \in \{1, 2, 3, 4\}$ and $Y_i \in \{1, 2, 3\}$. The following (4×3) table summarizes the data by reporting $n_{X,Y}(x, y) = \#\{(X_i, Y_i) = (x, y)\}$, the count of observations with $X_i = x$ and $Y_i = y$.

Table 1. counts $n_{X,Y}(x, y)$ (center block), $n_X(x)$ (right column) and $n_Y(y)$ (bottom row).

	y			
x	1	2	3	
1	10	9	2	21
2	10	7	2	19
3	17	11	5	33
4	22	18	7	47
	59	45	16	

The row totals are $n_X(x) = \#\{X_i = x\}$ and similarly the column totals report $n_Y(y) = \#\{Y_i = y\}$.

Let $f_{X,Y}(x, y) = n_{X,Y}(x, y)/n$ denote the (relative) frequencies, and simiarly for $f_X(x)$ and $f_Y(y)$. We use $f_X \approx p_X$ as an estimate for p_X, and $f_Y \approx p_Y$ as an estimate for p_Y. Then

$$\hat{I} = \sum_{x,y} f_{X,Y}(x, y) \log\left(\frac{f_{X,Y}(x, y)}{f_X(x) f_Y(y)}\right),$$

serves as estimate for $I(X,Y)$. [1] In the following questions we implement a possible approach to decide whether to report that $X \perp Y$, or $X \not\perp Y$. The logic is

(a) If $X \perp Y$ were true, then $I(X,Y) = 0$.
(b) Instead of $I(X,Y)$ we can only evaluate $\hat{I} \approx I(X,Y)$. It is okay for $\hat{I} > 0$, but it should not be "too large" if $X \perp Y$ were true;
(c) To judge how much is "too large", we carry out a small simulation: We generate a hypothetical repeat of the experiment, generating $X_i' \sim f_X$, $Y_i' \sim f_Y$, independently, $i = 1, \ldots, n$, and evaluate $\hat{I}'$. Repeat this simulation $M = 100$ times and record the M evaluations of $\hat{I}'$ (see the footnote about zero counts). Let $\hat{I}_m'$, $m = 1, \ldots, M$, denote the *ordered* list of those M evaluations. We use $\hat{I}^\star \equiv \hat{I}_{95}'$ to draw the line and decide what is "too large."
(d) If $\hat{I}$ is "too large", i.e., $\hat{I} > \hat{I}^\star$ report $X \not\perp Y$. Otherwise we report $X \perp Y$.

The logic of our algorithm is an indirect argument: if in fact $X \perp Y$ were true, then $\hat{I} > \hat{I}^\star$ is unlikely. Therefore we accept $\hat{I} > \hat{I}^\star$ as evidence "beyond reasonable doubt" against $X \perp Y$.

In this setup, answer the following questions:

10a. Evaluate $\hat{I}$ (step 2, above). You can use the R macro `Ih()` below this problem (you need not use them).

$\boxed{\hat{I} = \quad\quad\quad\quad}$

10b. Carry out the simulation described in step 3. Find $\hat{I}^\star$. See the R code fragments `sim()` and `Ihmstar()` shown below this problem (you need not use them).

$\boxed{\hat{I}^\star = \quad\quad\quad\quad}$

[1] We lucked out here with all counts being non-zero. If any count were 0, we could just replace it by 0.5.

10c. If $X \perp Y$ is true, then what is $\Pr(\hat{I} > \hat{I}^\star)$? An approximate argument is okay.

$\boxed{\Pr(\hat{I} > \hat{I}^\star) = \quad}$

10d. In the light of (b) and (c), and following steps 1-4 above, what do you report?

(a) $X \perp Y$ (b) $X \not\perp Y$ (c) can not decide (d) need more data (e) none of these

Note: In this problem we carried out a test of the hypothesis $H_0 : X \perp Y$. We used $\hat{I}(X,Y)$ as a test statistic. The decision rule to reject H_0 when $\hat{I} > \hat{I}^\star$ defined the rejection region. We will talk much more about the concept of hypothesis tests later in the course.

Below are three R scripts to evaluate $\hat{I}$ and to carry out the simulation in **10b.**. The argument of `Ih(nxy)` is the (4×3) table of counts n_{XY}, and function returns $\hat{I}(X,Y)$. The arguments of `sim(M,fxh,fyh)` are the simulation size M, f_X and f_Y, and the function returns an $(M \times 1)$ vector $\hat{I}'$ of $\hat{I}(X,Y)$ for the M simulations.

```
Ih <- function(nxy)
{ # input: f= nx x ny table of counts
  # output: MI I(X,Y)
    n <- sum(nxy)
    f <- nxy/n
    fx <- apply(f,1,sum)
    fy <- apply(f,2,sum)
    fx[fx==0] <- 0.01
    fy[fy==0] <- 0.01
    f[f==0] <- 0.01
    fxfy <- fx %*% t(fy)
    I  <- sum(f*log(f/fxfy))
    return(I)
}

Ihmstar  <- function(Ihm)
{ ## input: list Ihm[1..M] of Ih(X,Y)
  ## output: threshold Ih* for test
    M <- length(Ihm)
    istar <- round(0.95*M)
    Is  <- sort(Ihm)[istar]
    return(Is)
}
```

```
sim  <- function(M,fxh,fyh)
{ ## input: M=#simulations,
  ##        fxh = marginal on X, fyh = .. on Y
  ## output: Ihm = (M x 1) vector of Ih(X,Y)
    n <- 120  # to match the table
    nx <- length(fxh)
    ny <- length(fyh)
    Ihm  <- rep(0,M) # initialize
    for(m in 1:M){
        xm <- sample(1:nx,n,prob=fxh,replace=T)
        ym <- sample(1:ny,n,prob=fyh,replace=T)
        nxym <- table(xm,ym)
        Ihm[m]  <- Ih(nxym)
    }
    return(Ihm)
}
```