ACM 116 Problem Set 1 Solutions

Hannes Helgason

Problem 1

(a) Let p be the probability of any individual having the disease, and whether one individual has a disease or not doesn't depend upon the other individuals. Let $q = 1 - p$. Consider the following indicator variables, where A_1 and A_2 correspond to the 1st and 2nd halves:

$$I_{A_k} = \begin{cases} 1 & \text{if someone in the } k\text{-th half, i.e. in } A_k, \text{ is infected,} \\ 0 & \text{otherwise} \end{cases}$$

Construct I_{B_k} similarly, where B_k corresponds to the k-th quarter. Note that $B_1, B_2 \subset A_1$ and $B_2, B_3 \subset A_2$, meaning that if someone is infected in B_1, then someone is infected in A_1 and so forth. Thus, $I_{B_1} I_{A_1} = I_{B_1 \cap A_1} = I_{B_1}$ and the number of samples taken is

$$\begin{aligned} N &\equiv 1 + 1 + 2 \cdot (I_{A_1} + I_{A_2}) + \frac{n}{4} \cdot (I_{B_1} I_{A_1} + I_{B_2} I_{A_1} + I_{B_3} I_{A_2} + I_{B_4} I_{A_4}) \\ &= 2 + 2 \cdot (I_{A_1} + I_{A_2}) + \frac{n}{4} \cdot (I_{B_1} + I_{B_2} + I_{B_3} + I_{B_4}) \end{aligned}$$

$$E(N) = 2 + 2 \cdot \sum_{k=1}^{2} E(I_{A_k}) + \frac{n}{4} \cdot \sum_{k=1}^{4} E(I_{B_k})$$

Evaluating,

$$\begin{aligned} E(I_{A_k}) &= P(\text{at least one infected in a group of } n/2) \\ &= 1 - P(\text{nobody infected in a group of } n/2) \\ &= 1 - (1-p)^{n/2} = 1 - q^{n/2} \end{aligned}$$

and similarly

$$E(I_{B_k}) = P(\text{at least one infected in a group of } n/4) = 1 - q^{n/4}.$$

Hence,

$$\begin{aligned} E(N) &= 2 + 4 \cdot E(I_{A_1}) + \frac{n}{4} \cdot 4 \cdot E(I_{B_1}) \\ &= -4\ q^{n/2} - n\ q^{n/4} + 6 + n \end{aligned}$$

(b) We wish to find when

$$\begin{gathered} -4\ q^{n/2} - n\ q^{n/4} + 6 + n > n \ \text{ or,} \\ 4\ q^{n/2} + n\ q^{n/4} - 6 < 0 \end{gathered}$$

The quadratic has roots $q^{n/4} = \frac{-n \pm \sqrt{n^2+96}}{8}$. The testing scheme is inferior in the region $q \in [0, (-\frac{n}{8} + \frac{1}{8}\sqrt{n^2+96})^{4/n})$. Alternatively, the test is inferior if

$$p \in \left(1 - \left(\frac{-n + \sqrt{n^2+96}}{8}\right)^{4/n}, 1\right].$$

(c) Where $n = k\ m$, construct analogous indicator functions for each of the m groups.

$$N \equiv m + k\ \sum_{i=1}^{m} I_{i/m}$$

As before, $E(I_{i/m}) = 1 - (1-p)^{n/m} = 1 - q^k$, and we conclude,

$$E(N) = m + n - n\ (1-p)^k.$$

Problem 2

(a) We shall need the following enumerations:

$$\begin{aligned}
\#\{\text{hands}\} &= \binom{52}{13} \\
\#\{\text{hands with no Aces}\} &= \binom{48}{13} \\
\#\{\text{hands with exactly 1 Ace}\} &= \binom{4}{1}\binom{48}{12}
\end{aligned}$$

$$\begin{aligned}
P(\text{2 or more Aces} \mid \text{at least one Ace}) &= \frac{P(\text{ 2 or more Aces})}{P(\text{at least 1 Ace})} \\
&= \frac{\#\{\text{hands w/ 2 or more Aces}\}}{\#\{\text{hands w/ at least 1 Ace}\}} \\
&= \frac{\binom{52}{13} - \binom{48}{13} - \binom{4}{1}\binom{48}{12}}{\binom{52}{13} - \binom{48}{13}} \\
&\approx 0.370
\end{aligned}$$

(b) Similarly,

$$\begin{aligned}
\#\{\text{hands w/ Ace of Hearts}\} &= \binom{51}{12} \\
\#\{\text{hands w/ Ace of Hearts and no other Aces}\} &= \binom{1}{1}\binom{48}{12} \\
P(\text{2 or more Aces} \mid \text{Ace of Hearts}) &= \frac{\#\{\text{2 or more Aces and Ace of Hearts}\}}{\#\{\text{Ace of Hearts}\}} \\
&= \frac{\binom{51}{12} - \binom{48}{12}}{\binom{51}{12}} \\
&\approx 0.561
\end{aligned}$$

It is amusing to note the source of the discrepancy between (a) and (b). Consider a simpler example: choosing two elements of the list $\{A, B, c\}$.

$$P(\text{both capital} \mid \text{at least one capital}) = \frac{1}{3}.$$

But,

$$P(\text{both capital} \mid \text{A is chosen}) = \frac{1}{2}.$$

This difference is caused because of the nonempty intersection of {choices that include A} and {choices that include B}. The information that A is chosen, restricts the unwanted case of $\{B, c\}$ being chosen, but maintains the case of $\{A, B\}$. Hence, the conditional probability of $\{A, B\}$ becomes greater.

A similar effect is at play in parts (a) and (b). Naively, knowing which Ace is in a hand should cut in fourth all relevant quantities. However, hands with multiple Aces remain in consideration after the extra information. Hence, their relative abundance is greater.

Problem 3

Fix a particular $X_1 + X_2 + \cdots + X_n \equiv s$.

$$s = E(X_1 + \cdots + X_n \mid X_1 + \cdots + X_n = s) = \sum_{i=1}^{n} E(X_i \mid s) = n\ E(X_i \mid s)$$

The first equality follows from definition, the second from linearity of expectation, and the third from the symmetry between X_i.

$$E(X_i \mid X_1 + \cdots + X_n = s) = \frac{s}{n}$$

Problem 4

Solution 1: The joint moment generating function of two random variables U and V is defined by

$$\phi_{V,W}(s,t) = E(e^{sV+tW})$$

and V and W are independent if and only if $\phi_{V,W}(s,t) = \phi_V(s)\phi_W(s)$. Set $V = X + Y$, $W = X - Y$. Then V and W are normal random variables with means $E(V) = 2\mu$, $E(W) = 0$ and variances $Var(V) = Var(W) = 2\sigma^2$. Then

$$\begin{aligned}
\phi_{V,W}(s,t) &= E(e^{sV+tW}) = E(e^{(s+t)X}e^{(s-t)Y}) = E(e^{(s+t)X})E(e^{(s-t)Y}) \\
&= \exp\left(\mu(s+t) + \frac{\sigma^2(s+t)^2}{2}\right)\exp\left(\mu(s-t) + \frac{\sigma^2(s-t)^2}{2}\right) \\
&= \exp\left(2\mu s + \frac{2\sigma^2 s^2}{2}\right)\exp\left(\frac{2\sigma^2 t^2}{2}\right) \\
&= \phi_V(s)\phi_W(t)
\end{aligned}$$

which follows from that X and Y are independent and the moment generating function for Gaussians. Thus, $V = X + Y$ and $W = X - Y$ are independent.

Solution 2: Another way to do this problem is to show that the joint density function for $V = X + Y$ and $W = X - Y$ is equal to the product of the density functions for these two random variables V and W. Since X and Y are independent, their joint density function is the product of their distribution function. Then we can use the Jacobian determinant (see Ross p.63-64) to calculate the joint density for V and W. Write $g_1(x,y) = x + y$, $g_2(x,y) = x - y$ so $x = (g_1 + g_2)/2$ and $y = (g_1 - g_2)/2$. Then the Jacobian determinant is $J(x,y) = -2$ so the joint density function for V and W becomes

$$\begin{aligned}
f_{V,W}(v,w) &= f_{X,Y}(x,y) \mid J(x,y) \mid^{-1} = \frac{1}{2} f_{X,Y}\left(\frac{v+w}{2}, \frac{v-w}{2}\right) \\
&= \frac{1}{2}\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\left(\frac{v+w}{2}-\mu\right)^2/(2\sigma^2)} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\left(\frac{v-w}{2}-\mu\right)^2/(2\sigma^2)} \\
&= \frac{1}{\sqrt{2\pi(2\sigma^2)}} e^{-(z-2\mu)^2/(2(2\sigma^2))} \frac{1}{\sqrt{2\pi(2\sigma^2)}} e^{-w^2/(2(2\sigma^2))} \\
&= f_V(v) f_W(w)
\end{aligned}$$

so $V = X + Y$ and $W = X - Y$ are independent.

Remark: If V and W are bivariate normal variables, they are independent iff $Cov(V, W) = 0$, so that gives us yet another approach to solve the problem. Note that this does not hold in general if the random variables are not bivariate normal (just normal is not enough - bivariate is crucial). Take for example $V = Z$ and $W = U|Z|$, where $Z \sim N(0, 1)$ and U is a Bernoulli random variable taking the values 1,−1 with probabilities 1/2 each. Clearly V and W are both distributed like $N(0, 1)$ and $Cov(V, W) = 0$ but V and W are clearly dependent.

Problem 5

The Chebyshev's Inequality goes as follows: If Y is a random variable with mean μ and variance σ^2, then for any $k > 0$,

$$P\{|Y - \mu| \geq k\} \leq \frac{\sigma^2}{k^2}.$$

Now, let $X_1, X_2, \ldots$ be iid (independent and identically distributed) random variables with mean μ and variance σ^2. Define $S_n = \frac{1}{n}\sum_{k=1}^{n} X_k$. Then $E\{S_n\} = \mu$ and

$$Var\{S_n\} = Var\Big\{\frac{1}{n}\sum_{k=1}^{n} X_k\Big\} = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n},$$

since the X's are independent. Let $\epsilon > 0$ be given. Using Chebyshev's Inequality with $Y = S_n$, $k = \epsilon$ we get

$$P\{|S_n - \mu| \geq \epsilon\} \leq \frac{\sigma^2}{n\epsilon^2}$$

for all $n \geq 1$. It immediately follows that

$$P\{|S_n - \mu| > \epsilon\} \to 0 \qquad \text{as } n \to \infty.$$

Problem 6

Assume we only have a routine that samples uniform random variables on the interval $[0, 1]$.

(a) Let $U \sim Unif(0, 1)$ and Y be a random variable that is generated by the process described. Recall that $P(x \in I) = |I|$ for any subinterval I in $[0, 1]$. Then

$$\begin{aligned}
P\Big(Y = k\Big) &= P\Big(U - (p_0 + \ldots + p_{k-1}) < p_k \mid U \geq p_0, U - p_0 \geq p_1 \ldots, U - (p_0 + \ldots + p_{k-2}) \geq p_{k-1}\Big) \\
&= P\Big(U < p_0 + \ldots + p_k \mid U \geq p_0 + \ldots + p_{k-1}\Big) \\
&= P\Big(p_0 + \ldots + p_{k-1} \leq U < p_0 + \ldots + p_k\Big) \\
&= p_0 + \ldots + p_k - (p_0 + \ldots + p_{k-1}) = p_k = P(X = k)
\end{aligned}$$

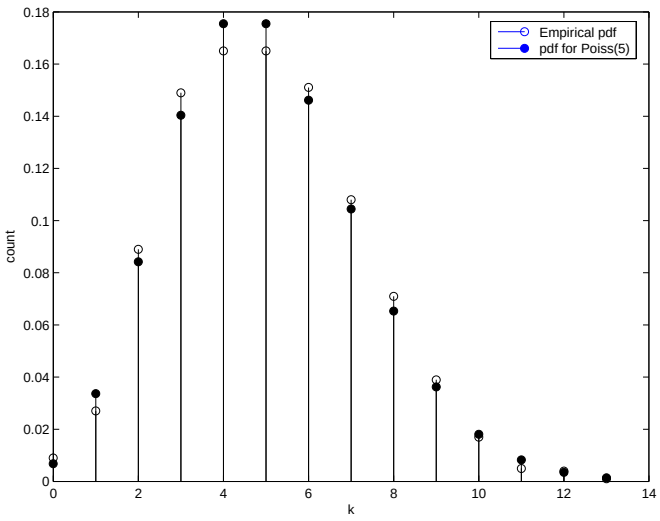
for any positive integer k. If $k = 0$, $P(Y = 0) = P(U < p_0) = p_0 = P(X = 0)$.

(b) If $X \sim Pois(\lambda)$, i.e. Poisson random variable with a parameter λ, then

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!} \quad \text{and} \quad P(X = k + 1) = \frac{\lambda}{k + 1} P(X = k).$$

We will use the latter expression to calculate our p_k's in the Matlab code.

Results from a simulation: The figure below shows the empirical probability distribution based on 1000 simulations done with the process described plotted along with the probability distribution for a Poisson random variable with parameter $\lambda = 5$. We notice that there is not a big difference between the two.



Those that have had an introductory course in statistics remember that the maximum likelihood estimator for the parameter for Poisson distribution, based on a collection of i.i.d. samples, is equal to the mean of those samples. Thus, if everything went well the mean of our generated samples should be close to 5. For the simulations that generated the plot above we got $\hat{\lambda} \approx 5.004$ which is pretty close to 5 as we expected.

Appendix

Below is the `Matlab`-code for Problem 6:

```
% Generate 1000 uniform random variables on the interval
% from 0 to 1.
B = 1000;
unifsamples = unifrnd(0,1,B,1);

X = zeros(1,B); % vector to store simulated rv's

lambda = 5; % rate parameter for Poisson distribution
for s=1:B,
    % loop over simulations
    U = unifsamples(s);
    k = 0;
%    p = poisspdf(k,lambda); % calculate p0
    p = exp(-lambda); % calculate p0
    while (U >= p | k > 1e6),
        % looping until U < p_k,
        % the latter constriction is just put
        % to ensure that the loop is exited sometime
        U = U-p;
        k = k+1;
%        p = poisspdf(k,lambda); % calculate p_k
        p = lambda/k*p; % calculate p_k
    end
    X(s) = k;
```

```
end

epdfX = zeros(1,max(X)+1); % empirical pdf based on X
for m=0:max(X),
   epdfX(m+1) = sum(X==m)/B;
end

figure(1);
stem(0:max(X),epdfX);
xlabel('k');ylabel('count');
hold on;
m = 0:max(X);
stem(m,poisspdf(m,lambda),'filled');
legend('Empirical pdf','pdf for Poiss(5)');
print -deps histogram.eps
```