STAT201A – Sec. 102
Homework #5.

Steven Pollack
24112977

#1.

Proof. If $X_1, X_2, \ldots, X_{30}$ are an SRSample from a population of size $N = 800$, and $G = 200$ of them will vote for candidate A, then $X = \sum_{i=1}^n X_i$ is a Hypergeometric random variable with parameters $N = 800, G = 200, n = 30$. Consequently,

$$\mu_X = 7.5 \text{ and } \sigma_X^2 \approx 5.421$$

We could use Chebyshev's inequality, to find

$$P(X \in [0, 14]) \geq 90\%$$

However, this is clearly too loose of a range. A simple numerical calculation finds that

$$P(4 \leq X \leq 11) \approx 91.74\%$$

Hence, there's at least 90% change that between $2/15$ and $11/30$ of the voters will vote for candidate A. □

#2.

Proof. Let I_i be the indicator random variable which signals 1 on the event that toss i yields a heads. Then, for any $n \in \mathbb{N}$, set $H_n = \sum_{i=1}^n I_i$; Using the assumption that the I_j's are IID Bernoulli with parameter p, we have that $H_n \sim$ Binomial(n,p)[1].

Now, we wish to find

$$\rho(H_n, H_{n+k}) = \frac{\operatorname{cov}(H_n, H_{n+k})}{SD(H_n)SD(H_{n+k})}$$

so the we'll investigate the covariance term in the numerator.

$$\begin{aligned}\operatorname{cov}(H_n, H_{n+k}) &= \operatorname{cov}\left(\sum_{i=1}^n I_i, \sum_{j=1}^k I_j\right) \\ &= \sum_{i=1}^n \sum_{j=1}^{n+k} \operatorname{cov}(I_i, I_j)\end{aligned}$$

But, recall that I_i and I_j were assumed to be iid bernoulli trials with success probability p. Hence,

$$\operatorname{cov}(I_i, I_j) = \begin{cases} \operatorname{var}(I_i) = p(1-p) & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}$$

Thus,

$$\operatorname{cov}(H_n, H_{n+k}) = n \operatorname{var}(I_1) = np(1-p)$$

[1] though, we didn't need to express H_n as this particular sum to see this.

and using the fact that $\mathrm{var}(H_n) = np(1-p)$, it follows that

$$\rho(H_n, H_{n+k}) = \frac{np(1-p)}{\sqrt{n(1-p)p}\sqrt{(n+k)(1-p)p}} = \sqrt{\frac{n}{n+k}} = \frac{1}{\sqrt{1+\frac{k}{n}}}$$

For fixed n, this result claims that the behavior of the correlation is solely a function of k and not p. In particular, the correlation behaves like $k^{-1/2}$. While the speed of this decay isn't necessarily intuitive, the fact that $k \ll n$ implies a high amount of correlation (i.e. dependence) is quite reasonable. For example, if $n = 5$ and $k = 1$, then knowing that $H_5 = 3$, gives us a lot of insight into the possible values of H_6. In generality: given $H_n = h$, we know that $h \leq H_{n+k} \leq h + k$ and more important, we may perform calculations regarding H_{n+k} via the identity

$$P(H_{n+k} = h' \mid H_n = h) = P(H_k = h' - h)$$

It should also be said that the nature of H_n being the sum of IID Bernoulli trials with success rate p makes the fact that ρ is independent of p all the more reasonable: changing p affects both H_n and H_{n+k} in the same way (leaving the dependence relationship intact). □

#3.

Proof. Using the hint, we note that $S, F \sim \text{Binomial}(n, p)$, (where $p = 1/6$), and $S + F \sim \text{Binomial}(n, 2p)$. It then follows that

$$\mathrm{var}(S+F) = \mathrm{var}(S) + \mathrm{var}(F) + 2\,\mathrm{cov}(S,F) \iff n(2p)(1-2p) = 2np(1-p) + 2\,\mathrm{cov}(S,F)$$

Solving for $\mathrm{cov}(S,F)$, we get

$$\mathrm{cov}(S,F) = -np^2$$

Hence,

$$\mathrm{var}(S-F) = \mathrm{var}(S) + \mathrm{var}(F) - 2\,\mathrm{cov}(S,F) = 2np(1-p) + 2np^2 = 2np = n/3$$

That is, $SD(S-F) = \sqrt{n/3}$. From the linearity of expectation, we get

$$E(S-F) = E(S) - E(F) = 0$$

□

#4.

Proof. a) Give each card a numeric value, and reserve the numbers $49, 50, 51, 52$ for the four aces. Then, for a given shuffle, ω, let $I_j(\omega)$ indicate the event that card j precedes all four aces in the ordering of that particular shuffle. It then follows that

$$P(X = k) = P\left(\left\{\omega : \sum_{j=1}^{48} I_j(\omega) = k-1\right\}\right)$$

Furthermore, it follows from a calculation similar to the one performed on homework #1 that $E[I_j] = 1/5$.

Hence,

$$E\left[X\right] = E\left[1 + \sum_{j=1}^{48} I_j\right] = 1 + \sum_{j=1}^{48} E[I_j] = 1 + \frac{48}{5} = 10.6$$

b) Now, using the fact that our shuffle is places cards i and j randomly in the deck, we realize that knowing anything about the position of i reveals little about the position of j relative to any of the aces. Hence $I_i \perp\!\!\!\perp I_j$ for $i \neq j$. This allows us to calculate $SD(X)$ via the root sum of $\mathrm{var}(I_j)$'s. That is,

$$\mathrm{var}(X) = \mathrm{var}\left(1 + \sum_{j=1}^{48} I_j\right) = \mathrm{var}\left(\sum_{j=1}^{48} I_j\right) = \sum_{j=1}^{48} \mathrm{var}(I_j) = 48 \cdot \frac{1}{5} \cdot \frac{4}{5} = 7.68$$

Thus,

$$SD(X) = \sqrt{7.68} \approx 2.77$$

□

#5.

a) Let $I_j = 1$ if student j gets their own homework and 0 otherwise. Then,

$$E[I_j] = P(I_j = 1) = 1/n$$

(since we assume the homework was distributed at random, and thus had equal chance to goto student i or j). Also, if we know that student j got his homework, we know that there's still a chance that student i can receive his. That chance is $P(I_i = 1 \mid I_j = 1) = 1/(n-1)$. Hence,

$$E[I_i I_j] = P(I_i = 1 = I_j) = P(I_i = 1 \mid I_j = 1)P(I_j) = \frac{1}{n-1} \cdot \frac{1}{n}$$

b) The "intuition" is based on the results on the previous assignment. Clearly, $M_n = \sum_{j=1}^{n} I_j$, and $I_j \sim$ Bernoulli(n^{-1}). Hence, $E(M_n) = nE(I_1) = 1$. Now, the intuition that comes into play here, is that while $I_i \not\perp\!\!\!\perp I_j$, they're fairly close to being independent. If they were independent,

$$P(I_i I_j) = P(I_i)P(I_j) = \frac{1}{n \times n} \neq \frac{1}{n(n-1)}$$

But when n is large enough, the difference between n^{-2} and $(n^2 - n)^{-1}$ is negligible. That is, we may as well treat the events as independent and thus M_n looks like a Binomial random variable. However, for $p_n = \Theta(n^{-1})$, $X \sim$ Poisson(np_n) approximates $Y \sim$ Binomial(n, p) incredibly well. Hence, why we can consider M_n to have something similar to a Poisson$(\mu = 1)$ distribution.

c) As demonstrated in part b), $E(M_n) = 1 \to 1$ as $n \to \infty$, so there's no issues here.

$$\begin{aligned}
\mathrm{var}(M_n) &= \mathrm{var}\left(\sum_{j=1}^{n} I_j\right) \\
&= \sum_{j=1}^{n} \mathrm{var}(I_j) + \sum_{i\neq j} \mathrm{cov}(I_i, I_j) \\
&= \sum_{j=1}^{n} \frac{1}{n}\left(1 - \frac{1}{n}\right) + \sum_{i \neq j}\left(\frac{1}{n(n-1)} - \frac{1}{n^2}\right) \\
&= \frac{n-1}{n} + \frac{n(n-1)}{n^2(n-1)} \\
&= \frac{n-1}{n} + \frac{1}{n} \\
&= 1
\end{aligned}$$

So, $SD(M_n) = 1 \to 1$ as $n \to \infty$, and both of the established limits agree with what one would find, should $M_n \xrightarrow{P} M \sim$ Poisson(1).

#6.

Proof. Assuming the following identity for $2 \leq m \leq n-1$:

$$P\left(\bigcup_{i=1}^{m} A_i\right) = \sum_{i=1}^{m} P(A_i) - \sum_{1\leq i<j\leq m} P(A_iA_j) + \sum_{1\leq i<j<k\leq m} P(A_iA_jA_k) - \cdots + (-1)^{m+1}P(A_1A_2\cdots A_m)$$

we preemptively use this to rewrite

$$P\left(\bigcup_{i=1}^{n-1} A_i\right) = \sum_{i=1}^{n-1} P(A_i) - \sum_{1\leq i<j\leq n-1} P(A_iA_j) + \sum_{1\leq i<j<k\leq n-1} P(A_iA_jA_k) - \cdots + (-1)^{n}P(A_1\cdots A_{n-1})$$

and

$$\begin{aligned}
P\left(\bigcup_{i=1}^{n-1} A_nA_i\right) = \sum_{i=1}^{n-1} P(A_nA_i) &- \sum_{1\leq i<j\leq n-1} P(A_nA_iA_j) \\
&+ \sum_{1\leq i<j<k\leq n-1} P(A_nA_iA_jA_k) - \cdots + (-1)^{n}P(A_1\cdots A_n)
\end{aligned}$$

Hence,

$$
\begin{aligned}
P\left(\bigcup_{i=1}^{n} A_i\right) &= P\left(A_n \cup \bigcup_{i=1}^{n-1} A_i\right) \\
&= P\left(\bigcup_{i=1}^{n-1} A_i\right) + P(A_n) - P\left(A_n \cap \bigcup_{i=1}^{n-1} A_i\right) \\
&= \sum_{i=1}^{n-1} P(A_i) + P(A_n) - \sum_{1\le i<j\le n-1} P(A_iA_j) \\
&+ \sum_{1\le i<j<k\le n-1} P(A_iA_jA_k) - \cdots + (-1)^n P(A_1\cdots A_{n-1}) - P\left(\bigcup_{i=1}^{n-1} A_nA_i\right) \\
&= \sum_{i=1}^{n} P(A_i) - \left(\sum_{1\le i<j\le n-1} P(A_iA_j) + \sum_{i=1}^{n-1} P(A_nA_i)\right) \\
&+ \left(\sum_{1\le i<j<k\le n-1} P(A_iA_jA_k) + \sum_{1\le i<j\le n-1} P(A_nA_iA_j)\right) + \cdots + (-1)^{n+1}P(A_1\cdots A_n) \\
&= \sum_{i=1}^{n} P(A_i) - \sum_{1\le i<j\le n} P(A_iA_j) + \sum_{1\le i<j<k\le n} P(A_iA_jA_k) + \cdots + (-1)^{n+1}P(A_1\cdots A_n)
\end{aligned}
$$

□

#7.

Proof. Returning to the notation used in problem #5., my first claim is that

$$P\left(\prod_{i=1}^{m} I_{a_i} = 1\right) = P(I_{a_1} = I_{a_2} = \cdots = I_{a_m} = 1) = \frac{1}{[n]_{m-1}}$$

We already saw that $P(I_{a_1} = I_{a_2} = 1) = 1/[n]_1 = 1/n(n-1)$, so using the base case is valid. Assuming the result holds for $m-1$:

$$
\begin{aligned}
P\left(\prod_{i=1}^{m} I_{a_i} = 1\right) &= P\left(I_{a_m} = 1 \,\middle|\, \prod_{i=1}^{m-1} I_{a_i} = 1\right) P\left(\prod_{i=1}^{m-1} I_{a_i} = 1\right) \\
&= \frac{1}{n-(m-1)}\frac{1}{[n]_{m-2}} \\
&= \frac{1}{[n]_{m-1}}
\end{aligned}
$$

Using this result, and the fact that

$$\sum_{1\le i_1<i_2<\cdots<i_k\le n} 1 = \frac{[n]_{k-1}}{k!}$$

we may apply the principle of inclusion-exclusion to the set $M_n \geq 1$ in the following manner[2]

$$
\begin{aligned}
P(M_n \geq 1) &= P(I_k = 1 \text{ for some } 1 \leq k \leq n) \\
&= P\left(\bigcup_{k=1}^{n} I_k\right) \\
&= \sum_{k=1}^{n} P(I_k) - \sum_{1\leq i<j\leq n} P(I_iI_j) + \sum_{1\leq i<j<k\leq n} P(I_iI_jI_k) - \cdots + (-1)^{n+1}P(I_1\cdots I_n) \\
&= nP(I_1) - \frac{[n]_1}{2}P(I_1I_2) + \frac{[n]_2}{3!}P(I_1I_2I_3) - \cdots + (-1)^{n+1}\frac{1}{[n]_{n-1}} \\
&= 1 - \frac{1}{2!} + \frac{1}{3!} - \frac{1}{4!} + \cdots + \frac{(-1)^{n+1}}{n!} \\
&= \sum_{k=1}^{n} \frac{(-1)^{k+1}}{k!} \\
&= \sum_{k=0}^{\infty} \frac{(-1)^{k+1}}{k!} + 1 - \sum_{k=n+1}^{\infty} \frac{(-1)^{k+1}}{k!} \\
&= -e^{-1} + 1 - \sum_{k=n+1}^{\infty} \frac{(-1)^{k+1}}{k!}
\end{aligned}
$$

This shows that

$$P(M_n = 0) = 1 - P(M_n \geq 1) = \sum_{k=n+1}^{\infty} \frac{(-1)^{k+1}}{k!} + e^{-1} \xrightarrow{n\to\infty} e^{-1}$$

□

#8.

Proof. First, notice that $P(M_n = k)$ is the proportion of hand-backs that successfully return k students' work back. Using the fact that, if k students get the right homework, then $n-k$ must get the wrong work, we see that $[n]_{n-k}P(M_n = k)$ is the number of hand-backs, in a class of size n, which return k students' work back.

Similarly, one may count $\binom{n}{n-k}$ ways to select $n-k$ groups of students from our original class of n. Hence, $\binom{n}{n-k}P(M_{n-k} = 0)$ is the number of ways we can fail to get any particular student their own work back, inside a subclass of size $n-k$ formed from a superclass of size n. That is,

$$
\begin{aligned}
[n]_{n-k}P(M_n = k) = \binom{n}{n-k}P(M_{n-k} = 0) &\iff P(M_n = k) = \binom{n}{n-k}[n]_{n-k}^{-1}P(M_{n-k} = 0) \\
&\iff P(M_n = k) = \binom{n}{k}\frac{(n-k)!}{n!}P(M_{n-k} = 0)
\end{aligned}
$$

[2]please excuse the above of notation: $I_1 \cup I_2$ is the event that I_1 or I_2 is yields 1...

Now, applying this identity to preceding problem's result:

$$\begin{aligned}
\lim_{n\to\infty} P(M_n = k) &= \lim_{n\to\infty} \binom{n}{k} \frac{(n-k)!}{n!} P(M_{n-k} = 0) \\
&= \frac{1}{k!} \lim_{n\to\infty} P(M_{n-k} = 0) \\
&= \frac{e^{-1}}{k!}
\end{aligned}$$

$\square$