Homework 5 Solutions

1. (10 points) *Roulette*

   a. We model the bets as an iid random sequence $X_1, X_2, \ldots$ where

$$X_i := \begin{cases} 35 & \text{if Bob wins the } i\text{th bet,} \\ -1 & \text{if Bob loses the } i\text{th bet.} \end{cases} \tag{1}$$

   The mean and variance of X_i are

$$\mathrm{E}\left(X_i\right) = \sum_{x \in \{-1,1\}} x\, p_{X_i}\left(x\right) = 35 \cdot \frac{1}{38} - 1 \cdot \frac{37}{38} = -\frac{1}{19} \tag{2}$$

$$\mathrm{E}\left(X_i^2\right) = \sum_{x \in \{-1,1\}} x^2\, p_{X_i}\left(x\right) = 35^2 \cdot \frac{1}{38} + 1 \cdot \frac{37}{38} = \frac{631}{19} \tag{3}$$

$$\mathrm{Var}\left(X_i\right) = \mathrm{E}\left(X_i^2\right) - \mathrm{E}^2\left(X_i\right) = 33.2. \tag{4}$$

   Bob's gain after n bets is equal to

$$G_n^{\text{number}} = \sum_{i=1}^{n} X_i. \tag{5}$$

   The central limit theorem tells us that the distribution of G_n^{number}/n converges to a Gaussian distribution with mean $\mathrm{E}\left(X_i\right)$ and variance $\mathrm{Var}\left(X_i\right)/n$. This implies that the distribution of G_n^{number} converges to a Gaussian distribution with mean $n\,\mathrm{E}\left(X_i\right)$ and variance $n\mathrm{Var}\left(X_i\right)$ (linear transformations of Gaussians are Gaussian). We can approximate the probability of making money by

$$\mathrm{P}\left(G_n^{\text{number}} > 0\right) = P\left(\frac{G_n^{\text{number}} - \mathrm{E}\left(G_n^{\text{number}}\right)}{\sigma_{G_n^{\text{number}}}} > -\frac{\mathrm{E}\left(G_n^{\text{number}}\right)}{\sigma_{G_n^{\text{number}}}}\right) \tag{6}$$

$$= P\left(\frac{G_n^{\text{number}} - n\,\mathrm{E}\left(X_i\right)}{\sqrt{n\mathrm{Var}\left(X_i\right)}} > -\frac{\sqrt{n}\,\mathrm{E}\left(X_i\right)}{\sqrt{\mathrm{Var}\left(X_i\right)}}\right) \tag{7}$$

$$= P\left(\frac{G_n^{\text{number}} - \mathrm{E}\left(G_n^{\text{number}}\right)}{\sigma_{G_n^{\text{number}}}} > 0.09\right) \tag{8}$$

$$= Q\left(0.09\right) \approx 0.46. \tag{9}$$

   b. We model the bets as an iid random sequence $Y_1, Y_2, \ldots$ where

$$Y_i := \begin{cases} 1 & \text{if Bob wins the } i\text{th bet,} \\ -1 & \text{if Bob loses the } i\text{th bet.} \end{cases} \tag{10}$$

The mean and variance of Y_i are

$$\mathrm{E}\left(Y_i\right) = \sum_{x\in\{-1,1\}} x\, p_{Y_i}\left(x\right) = 1\cdot\frac{18}{38} - 1\cdot\frac{20}{38} = -\frac{1}{19} \tag{11}$$

$$\mathrm{E}\left(Y_i^2\right) = \sum_{x\in\{-1,1\}} x^2\, p_{Y_i}\left(x\right) = 1 \tag{12}$$

$$\mathrm{Var}\left(Y_i\right) = \mathrm{E}\left(Y_i^2\right) - \mathrm{E}^2\left(Y_i\right) = \frac{360}{361}. \tag{13}$$

Following the same exact reasoning as in the first question,

$$\mathrm{P}\left(G_n^{\text{color}} > 0\right) = P\left(\frac{G_n^{\text{color}} - \mathrm{E}\left(G_n^{\text{color}}\right)}{\sigma_{G_n^{\text{color}}}} > -\frac{\mathrm{E}\left(G_n^{\text{color}}\right)}{\sigma_{G_n^{\text{color}}}}\right) \tag{14}$$

$$= P\left(\frac{G_n^{\text{color}} - n\,\mathrm{E}\left(Y_i\right)}{\sqrt{n\mathrm{Var}\left(Y_i\right)}} > -\frac{\sqrt{n}\,\mathrm{E}\left(Y_i\right)}{\sqrt{\mathrm{Var}\left(Y_i\right)}}\right) \tag{15}$$

$$= P\left(\frac{G_n^{\text{color}} - \mathrm{E}\left(G_n^{\text{color}}\right)}{\sigma_{G_n^{\text{color}}}} > 0.527\right) \tag{16}$$

$$= Q\left(0.527\right) \approx 0.30. \tag{17}$$

c. According to the Law of Large Numbers the gains per bet $G_n^{\text{number}}/n \to \mathrm{E}\left(X_i\right) = -1/19$ and $G_n^{\text{color}}/n \to \mathrm{E}\left(Y_i\right) = -1/19$ as $n \to \infty$ so they are equivalent asymptotically.

d. Bob will choose to bet for a number. Recall that by the Central Limit Theorem the pdf of G^{number} is approximately Gaussian with mean $n\,\mathrm{E}\left(X_i\right) = -100/19 \approx -5.26$ and variance $n\mathrm{Var}\left(X_i\right) \approx 3320$. First we compute the cdf conditioned on the event $G^{\text{number}} > 0$ for $a > 0$

$$F_{G^{\text{number}}|\text{Bob makes money}}\left(a\right) = \frac{P\left(G^{\text{number}} \le a\right)}{P\left(\text{Bob makes money}\right)} \tag{18}$$

$$= \frac{F_{G^{\text{number}}}\left(a\right)}{0.46}. \tag{19}$$

Differentiating we obtain that

$$f_{G^{\text{number}}|\text{Bob makes money}}\left(a\right) = \begin{cases} \frac{1}{0.46\cdot\sqrt{2\pi\cdot 3320}} e^{-\frac{(a+5.26)^2}{2\cdot 3320}}, & \text{for } a > 0 \\ 0, & \text{otherwise.} \end{cases} \tag{20}$$

The mean conditioned on the event *Bob makes money* is

$$E\left(G^{\text{number}}|\text{Bob makes money}\right) = \int_{a=0}^{\infty} a\, f_{G^{\text{number}}|\text{Bob makes money}}\left(a\right)\,\mathrm{d}a \tag{21}$$

$$= \frac{1}{0.46\cdot\sqrt{2\pi}} \int_{u=5.26/\sqrt{3320}}^{\infty} \left(\sqrt{3320}u - 5.26\right) e^{-\frac{u^2}{2}}\,\mathrm{d}u \tag{22}$$

$$= -\sqrt{\frac{3320}{0.46\cdot 2\pi}} e^{-\frac{u^2}{2}}]_{5.26/\sqrt{3320}}^{\infty} - \frac{5.26}{0.46}\,Q(5.26/\sqrt{3320}) \tag{23}$$

$$= \sqrt{\frac{3320}{0.46\cdot 2\pi}} e^{-\frac{5.26^2}{6640}} - \frac{5.26}{0.46}\,Q\left(0.09\right) \tag{24}$$

$$\approx 44.50 \text{ dollars.} \tag{25}$$

2. (10 points) *Weight*

   a. By Corollary 1.21 in Lecture Notes 4, the width of a conservative $1-\alpha$ confidence interval is $2b/\sqrt{\alpha n}$ where b is a bound on the standard deviation. We bound the variance by the mean square, which in turn can be bounded by the maximum possible of the weight squared, which we set to 1400^2, so $b = 1400$. Fixing $\alpha = 0.05$ we find that if the width is 5 then

$$n = \lceil \left(\frac{2b}{5\sqrt{\alpha}} \right)^2 \rceil = 6272000. \tag{26}$$

   b. By Theorem 1.23 in Lecture Notes 4, the width of an approximate $1-\alpha$ confidence interval is

$$\frac{2\sigma}{\sqrt{n}} Q^{-1}\left(\frac{\alpha}{2}\right) \leq \frac{24}{\sqrt{n}} Q^{-1}\left(\frac{\alpha}{2}\right). \tag{27}$$

   Fixing $\alpha = 0.05$ we have that if the width is 5 then

$$n = \lceil \left(\frac{12}{5} Q^{-1}\left(\frac{\alpha}{2}\right) \right)^2 \rceil = 88. \tag{28}$$

   c. The number of confidence intervals that do not contain the mean is very close to 5% for $n = 1000$, but not so much for $n = 20$. From the plots it seems that a plausible reason is that the sample variance has not converged to the real variance.

3. (10 points) *Convergence in probability implies convergence in distribution.*

   a. If $A_n > a$ and $|A_n - A| \leq \epsilon$ then $A > a - \epsilon$, so that

$$\{A_n > a\} \cap \{|A_n - A| \leq \epsilon\} \subseteq \{A > a - \epsilon\}.$$

   For any sets S_1 and S_2, $S_1 \subseteq S_2$ implies $S_2^c \subseteq S_1^c$, so the above equation becomes

$$\{A \leq a - \epsilon\} \subseteq \{A_n \leq a\} \cup \{|A_n - A| > \epsilon\},$$

   after applying the De Morgan's laws. Finally, applying the union bound we obtain

$$\mathrm{P}(A \leq a - \epsilon) \leq \mathrm{P}(A_n \leq a) + \mathrm{P}(|A_n - A| > \epsilon).$$

   This proves the left inequality. For the right inequality, note that if $A > a+\epsilon$ and $|A_n - A| \leq \epsilon$ then $A_n > a$. Following the same reasoning as before, this implies

$$\{A_n \leq a\} \subseteq \{A \leq a + \epsilon\} \cup \{|A_n - A| > \epsilon\},$$

   which again by the union bound allows us to conclude

$$\mathrm{P}(A_n \leq a) \leq \mathrm{P}(A \leq a + \epsilon) + \mathrm{P}(|A_n - A| > \epsilon).$$

   b. Taking the limit as $n \to \infty$ in the expression obtained in (a) and using the fact that since $A_n \to A$ in probability $\lim_{n\to\infty} \mathrm{P}(|A_n - A| > \epsilon) = 0$ yields

$$\mathrm{P}(A \leq a - \epsilon) \leq \lim_{n\to\infty} \mathrm{P}(A_n \leq a) \leq \mathrm{P}(A \leq a + \epsilon).$$

Now, we take the limit as ϵ tends to zero and assume that F_A is continuous at a to obtain

$$F_A(a) = \mathrm{P}\left(A \leq a\right) \leq \lim_{n\to\infty} F_{A_n}\left(a\right) \leq \mathrm{P}\left(A \leq a\right) = F_A(a).$$

This implies that for any a such that F_A is continuous at a $\lim_{n\to\infty} F_{A_n}\left(a\right) = F_A(a)$, which establishes convergence in distribution.

4. (30 points) *Radioactive sample.*

   a. The sample running average is plotted in Figure 1. It converges to 3.5. If we assume that the expected value of X is finite then it is equal to zero, since f_X is an even function and consequently $xf_X(x)$ is odd so that

   $$\mathrm{E}(X) = \int xf_X(x)\mathrm{d}x = 0.$$

   By the Law of Large Numbers we should estimate c to equal 3.5 since $\bar{S}_n \to c$ in mean square and in probability. This works under the assumption that the X has finite mean and variance.

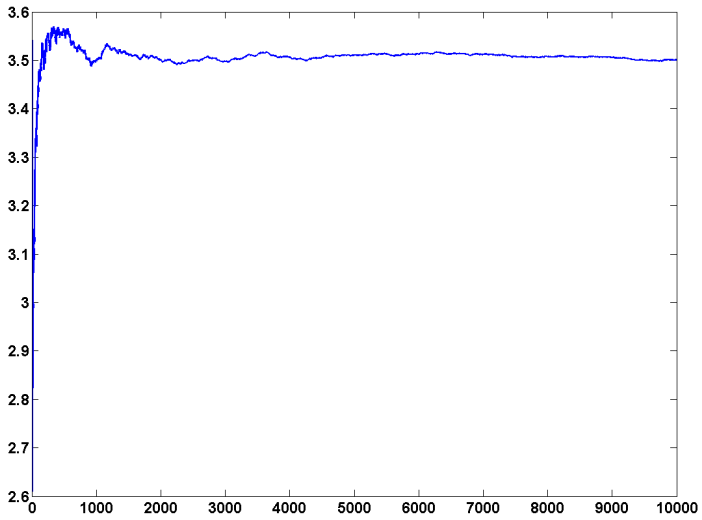

Figure 1: Sample running average of the data in *radioactive_sample_1.txt*.

   b. The sample running average is plotted in Figure 2. It does not seem to converge to any value, so the method does not work.

   c. The cdf of X is equal to

$$\begin{aligned}
F_X(x) &= \mathrm{P}(X \leq x) \\
&= \mathrm{P}(\tan A \leq x) \\
&= \mathrm{P}(A \leq \arctan x) \quad \text{by monotonicity of the tangent between } -\pi/2 \text{ and } \pi/2 \\
&= \frac{1}{\pi}\int_{-\pi/2}^{\arctan x} \mathrm{d}a \\
&= \frac{1}{2} + \frac{\arctan x}{\pi},
\end{aligned}$$

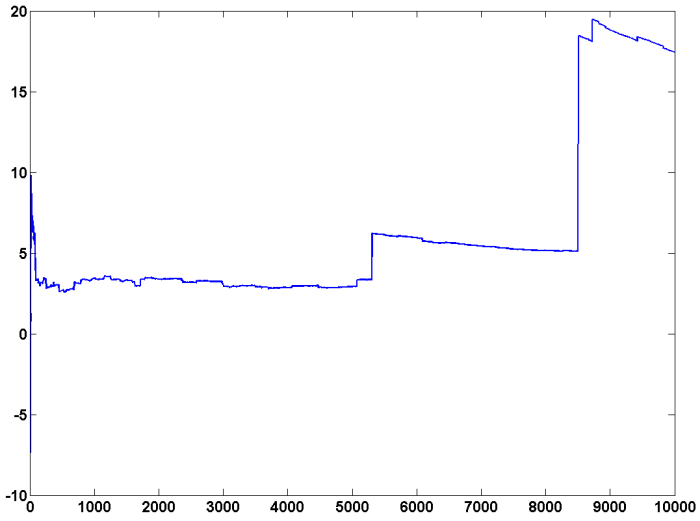

Figure 2: Sample running average of the data in *radioactive_sample_2.txt*.

so the pdf is equal to

$$f_X(x) = \frac{1}{\pi(1+x^2)}.$$

X is a Cauchy random value. As we saw in the notes, $\mathrm{E}(X)$ does not exist, as it is the difference of two limits that tend to infinity. The condition in (a) does not hold for this distribution, which is the reason that we cannot estimate c from the sample average. As we can see in Figure 2 the probability of having samples that deviate very significantly from the mean is relatively high, so that the running average jumps up and down without converging.

d. Figure 3 shows the plot, which further illustrates the point made in (c).

5. (10 points) *Election poll*

a. If we assume uniform random sampling, $R_1, R_2, \ldots$ is an iid sequence such that

$$\mathrm{E}\left(R_i\right) = \frac{m_R}{m_R + m_D}, \tag{29}$$

$$\mathrm{Var}\left(R_i\right) = \frac{m_R}{m_R + m_D}\left(1 - \frac{m_R}{m_R + m_D}\right) \leq 1. \tag{30}$$

where m_R is the number of people that vote Republican and m_D is the number of people that vote Democrat. By the Law of Large Numbers, the sample mean will converge to $\frac{m_R}{m_R+m_D}$, so if the sample mean is larger than 0.5 then we should predict that the Republican candidate will win the election.

b. The sample mean is

$$\mathrm{E}\left(\bar{R}_n\right) = \frac{45000}{94000} \approx 0.479 < 0.5, \tag{31}$$

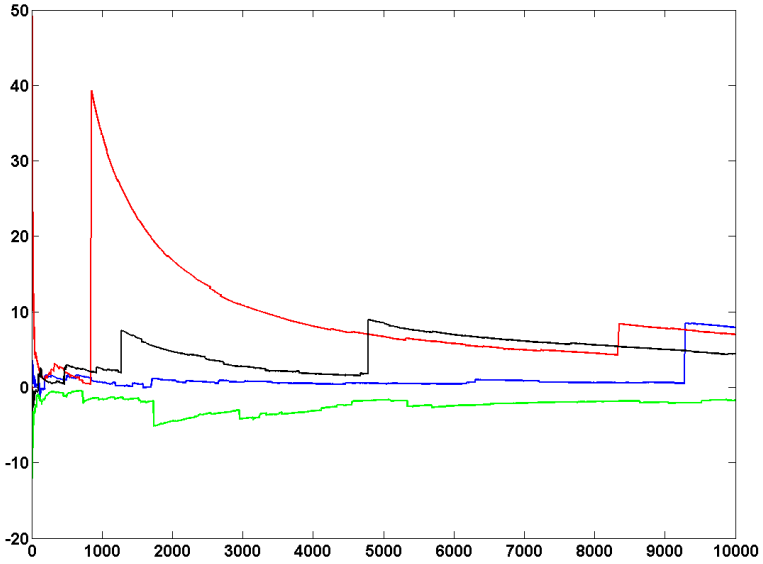

Figure 3: Sample running average of four different i.i.d. vectors of length 10^4 from the distribution of X derived in (c).

so we should predict that the Democrat will win. The sample mean is 0.021 away from 0.5. We can bound the probability that the sample mean is that distance away from the true mean by applying Chebyshev's inequality and bounding the variance by 1 (see (30)),

$$\mathrm{P}\left(\left|\bar{R}_n - \frac{m_R}{m_R + m_D}\right| > 0.021\right) \leq \frac{\mathrm{Var}\left(R_i\right)}{n \cdot 0.021^2} \leq \frac{1}{n \cdot 0.021^2} = 2.41\%. \tag{32}$$

This is the probability that I am right *before* computing the sample mean. Once we have computed it, we are either right or wrong.

c. If Y_i and O_i are uniformly sampled from the population of young people and old people respectively,

$$\mathrm{E}\left(Y_i\right) = \frac{y_R}{y_R + y_D}, \tag{33}$$

$$\mathrm{Var}\left(Y_i\right) = \frac{y_R}{y_R + y_D}\left(1 - \frac{y_R}{y_R + y_D}\right) \leq 1, \tag{34}$$

$$\mathrm{E}\left(O_i\right) = \frac{o_R}{o_R + o_D}, \tag{35}$$

$$\mathrm{Var}\left(O_i\right) = \frac{o_R}{o_R + o_D}\left(1 - \frac{o_R}{o_R + o_D}\right) \leq 1. \tag{36}$$

where y_R/o_R is the number of young/old people that vote Republican, y_D/o_D the number

of young/old people that vote Democrat. We want

$$\frac{y_R + o_R}{y_R + y_D + o_R + o_D} = \mathrm{E}\left(X_{n_1,n_2}\right) \tag{37}$$

$$= \mathrm{E}\left(a \sum_{i=1}^{n_1} Y_i + b \sum_{j=1}^{n_2} O_i\right) \tag{38}$$

$$= \frac{a\, n_1\, y_R}{y_R + y_D} + \frac{b\, n_2\, o_R}{o_R + o_D}, \tag{39}$$

so

$$a = \frac{y_R + y_D}{n_1\left(y_R + y_D + o_R + o_D\right)}, \tag{40}$$

$$b = \frac{o_R + o_D}{n_2\left(y_R + y_D + o_R + o_D\right)}. \tag{41}$$

d. According to the information given,

$$a = \frac{0.2}{0.75\, n_1}, \tag{42}$$

$$b = \frac{0.55}{0.75\, n_2}. \tag{43}$$

The realization of the estimator X_{n_1,n_2} is equal to

$$\frac{0.2 \cdot 24000}{0.75 \cdot 59000} + \frac{0.55 \cdot 21000}{0.75 \cdot 35000} = 0.548. \tag{44}$$

So according to the model, the Republican candidate will win the election.

To bound the probability that the realization of the estimator is 0.048 away from the outcome, we compute

$$\mathrm{Var}\left(X_{n_1,n_2}\right) = a^2 n_1 \mathrm{Var}\left(Y_i\right) + b^2 n_2 \mathrm{Var}\left(O_i\right) \leq \frac{(0.2)^2}{(0.75)^2\, 59,000} + \frac{(0.55)^2}{(0.75)^2\, 35,000} \leq 0.000017 \tag{45}$$

By Chebyshev's inequality

$$\mathrm{P}\left(\left|X_{n_1,n_2} - \frac{m_R}{m_R + m_D}\right| > 0.048\right) \leq \frac{\mathrm{Var}\left(X_{n_1,n_2}\right)}{0.048^2} \leq \frac{0.000017}{0.048^2} = 0.734\%. \tag{46}$$