Homework 3

Due Feb 12 at 11 pm

Unless stated otherwise, justify any answers you give. You can work in groups, but each student must write their own solution based on their own understanding of the problem.

When uploading your homework to Gradescope you will have to select the relevant pages for each question. Please submit each problem on a separate page (i.e., 1a and 1b can be on the same page but 1 and 2 must be on different pages). We understand that this may be cumbersome but this is the best way for the grading team to grade your homework assignments and provide feedback in a timely manner. Failure to adhere to these guidelines may result in a loss of points. Note that it may take some time to select the pages for your submission. Please plan accordingly. We suggest uploading your assignment at least 30 minutes before the deadline so you will have ample time to select the correct pages for your submission. If you are using LaTeX, consider using the minted or listings packages for typesetting code.

1. (Markov's and Chebyshev's inequalities are tight) In this problem we show that Markov's and Chebyshev's inequalities cannot be improved without further assumptions, because there exist random variables for which they are tight.

   (a) For any $c > 0$ and any $0 < \theta < 1$, build a nonnegative random variable $\tilde{a}$ such that

   $$\mathrm{P}\left(\tilde{a} \geq c\right) = \theta = \frac{\mathrm{E}\left[\tilde{a}\right]}{c}. \tag{1}$$

   - the conditions of Markov's inequality are met so we know that $P(\tilde{a} \geq c) \leq \frac{E[\tilde{a}]}{c}$
   - for any c suppose that we define a non-negative random variable $\tilde{a}$ such that $P(\tilde{a} \in \{0, c\}) = 1$ ie $\tilde{a}$ can only take on values of 0 or c.
   - if this is the case we can write $E[\tilde{a}] = \Sigma_{a\in\mathbb{R}} aP(\tilde{a} = a) = 0P(\tilde{a} = 0) + \Sigma_{a=1}^{c-1} aP(\tilde{a} = a) + cP(\tilde{a} = c) \leq 0P(\tilde{a} = 0) + (c)\Sigma_{a=1}^{c-1} P(\tilde{a} = a) + cP(\tilde{a} = c) = 0 + 0 + P(\tilde{a} = c)$ further since $\tilde{a}$ only takes on values c or 0 we know that $P(\tilde{a} = c) = P(\tilde{a} \geq c)$ thus we have $E[\tilde{a}] \leq cP(\tilde{a} \geq c) \Rightarrow P(\tilde{a} \geq c) \geq \frac{E[\tilde{a}]}{c}$
   - so thus we know$P(\tilde{a} \geq c) \geq \frac{E[\tilde{a}]}{c}$ and $P(\tilde{a} \geq c) \leq \frac{E[\tilde{a}]}{c}$ which implies that $P(\tilde{a} \geq c) = \frac{E[\tilde{a}]}{c}$

   (b) For any $c > 0$, any $0 < \theta < 1$ and any $\mu \in \mathbb{R}$, build a random variable $\tilde{b}$ with mean μ and finite variance, such that

   $$\mathrm{P}\left(|\tilde{b} - \mu| \geq c\right) = \theta = \frac{\mathrm{Var}\tilde{b}}{c^2}. \tag{2}$$

   - here we want to build $\tilde{b}$ such that $(\tilde{b} - \mu)^2$ satisfy $P((\tilde{b} - \mu)^2 \in \{0, c\}) = 1$
   - observe that regardless of the value of μ if $\tilde{b} = \sqrt{\mu}$ we have $(\tilde{b}^2 - \mu)^2 = (\mu - \mu)^2 = 0$
   - similarly regardless of the value of μ if $\tilde{b} = \sqrt{\mu + c^2}$ then $(\tilde{b}^2 - \mu)^2 = (c^2 + \mu - \mu)^2 = c^2$

- so if we design $\tilde{b}$ such that $P(\tilde{b} \in \{\mu, c^2 + \mu\}) = 1$ then the random variable $(\tilde{b} - \mu)^2$ will satisfy our conditions from part one and thus have $P((\tilde{b} - \mu)^2 \geq c) = \frac{E[(\tilde{b}-\mu)^2]}{c}$
- then we can see $P(|\tilde{b} - \mu|) = P((\tilde{b} - \mu)^2 \geq c^2) = \frac{E[\tilde{b}-\mu]^2}{c^2} = \frac{var(\tilde{b})}{c^2}$

2. (Online poll) In online polls, young people are often overrepresented. In this problem we study how to correct for this. When answering the questions use the following notation: α is the proportion of young people (between 18 and 35 years old) in the population, θ_1 the proportion of young people in the population who will vote for the Democratic candidate, θ_2 the proportion of old people in the population who will vote for the Democratic candidate, n_1 the number of young people in the poll, and n_2 the number of old people in the poll. Assume that α is known.

   (a) Derive an estimator of the proportion of voters that will vote for the Democratic candidate, as a function of the number of young people y and the number of old people o in the poll that intend to vote Democrat.
   - we can design an estimator $p = p(o, y) = \alpha(\frac{y}{n_1}) + (1 - \alpha)(\frac{0}{n_2})$

   (b) Evaluate your estimator for a poll with 100 participants where 60 intend to vote for the Democratic candidate. Out of the 100 participants, 70 are young, and 50 of them intend to vote for the Democratic candidate. The fraction of young people among voters in general is 25%.
   - for a poll with 100 participants where 60 intend to vote for the Democratic candidate our estimator would return a proportion of democrat voters equal to 0.4285714285714286

   (c) Under what assumptions is your estimator unbiased? Justify your answer mathematically.
   - we can see that the true proportion of democrat voters in the population is $p_{pop} = \alpha(\theta_1) + (a - \alpha)(\theta_2)$
   - we can find the expectation of our estimator p as $E[p] = E[\alpha(\frac{y}{n_1}) + (1-\alpha)(\frac{0}{n_2})] = E[\alpha(\frac{y}{n_1})] + E[(1 - \alpha)(\frac{0}{n_2})] = \frac{\alpha}{n_1}E[y] + \frac{1-\alpha}{n_2}E[o]$
   - note that we can write $Y = \Sigma_{i=1} n_1 1($ young person i votes democrat)) that is Y (the number of young people in the sample who voted democrat) is the sum of the number of people who chose to vote democrat and
   - similarly for old people we have $o = \Sigma_{i=1} n_2 1($ old person i votes democrat))
   - we assume that the young people y_i and old people o_i are sampled IID from respective populations Y_{pop}, O_{pop} and further that the sampling from Y_{pop} is independent of the sampling from O_{pop}
   - if this is the case we can write $y = \Sigma_{i=1}^{n} 1($ young person i votes democrat)) $= \Sigma_{i=1} n_1(y_i)$
   - thus $E[y] = E[\Sigma_{i=1} n_1(y_i)] = (n_1)_{i=1}^{n_1} E[y_i]$
   - as we assumed y_i is iid we know from lecture $E[y_i] = \frac{1}{N}\Sigma_{i=1}^{N} Y_i P(y_i = Y_i) = \theta_1$
   - thus we have $E[y] = E[\Sigma_{i=1}^{n_1}(y_i)] = (n_1)_{i=1}^{n_1} E[y_i] = (n - 1)(\theta_1)$
   - this allows us to write $\frac{E[y]}{n_1} = \frac{1(n_1)}{n_1} = \theta_1$
   - we can do the same argument for old people to show that under our assumptions $\frac{E[O]}{N_2} = \theta_2$
   - thus $E[p] = E[\alpha(\frac{y}{n_1}) + (1 - \alpha)(\frac{0}{n_2})] = E[\alpha(\frac{y}{n_1})] + E[(1 - \alpha)(\frac{0}{n_2})] = \frac{\alpha}{n_1}E[y] + \frac{1-\alpha}{n_2}E[o] = \alpha(_1) + (1 - \alpha)\theta_2$

- so we can finally see that under our assumptions $p_{pop} - E[p] = \alpha(_1) + (1-\alpha)\theta_2 - \alpha(_1) + (1-\alpha)\theta_2 = 0$ and thus p is an unbiased estimator of p_{pop}

(d) Show that your estimator is consistent as $n_1 \to \infty$ and $n_2 \to \infty$.

- we want to show that as $lim_{n_1,n_2\to\infty} P(|p - p_{pop}| \geq \epsilon = 0)$ for any positive ϵ
- first we showed that P is unbiased under random sampling assumptions above
- now we want to look at the variance of our estimator p.
- $var(p) = var(\alpha(\frac{y}{n_1}) + (1-\alpha)(\frac{0}{n_2}))$
- as we assumed samples are random iid are drawn independently from the young and old populations $var(p) = var(\alpha(\frac{y}{n_1}) + (1-\alpha)(\frac{0}{n_2})) = \frac{\alpha^2}{n_1^2} var(y) + \frac{(1-\alpha)^2}{n_2^2} var(o)$
- from here we can say that $var(p) = \alpha^2 var(\frac{y}{n_1}) + (1-\alpha)^2 var(\frac{o}{n_2})$
- we can see that $var(\frac{y}{n_1})$ is variance of a sample proportion of an unbiased rv thus $var(\frac{y}{n_1}) = var(\frac{1}{n_1}\Sigma_{i=1}^{n_1}\frac{y_i}{n_1} = \frac{1}{n_1^2}\Sigma_{i=1}^{n} var(\frac{y_i}{n-1}) = \frac{1}{n_1^4}\Sigma_{i=1}^{n} var(y_i)$
- $var(y_i) = var((n_1)\Sigma_{i=1}^{n_1} y_i) = n_1^2 \Sigma_{i=1}^{n} var(y_i) = n_1^2 \sigma_y$ where σ_y is the population variance of young people. using the fact that $var(y_i) = \Sigma_{i=1}^{N}(Y_{pop_i} - \theta_1)^2 P(y_i = Y_{pop_i}) = \frac{1}{n}\Sigma_{i=1}^{N}(Y_{pop_i} - \theta_1)^2 = \sigma_y^2$
- so finally we have $var(\frac{y}{n_1}) = var(\frac{1}{n_1}\Sigma_{i=1}^{n_1}\frac{y_i}{n_1} = \frac{1}{n_1^2}\Sigma_{i=1}^{n} var(\frac{y_i}{n-1}) = \frac{1}{n_1^4}\Sigma_{i=1}^{n} var(y_i) = \frac{\sigma_y^2}{n_1^2}$
- we can repeat the $var(\frac{0}{n_2}) = \frac{\sigma_0^2}{n_2^2}$
- this finally yields that $var(p) = \frac{\alpha^2\sigma_y^2}{n_1^2} + \frac{(1-\alpha)^2\sigma_o^2}{n_2^2}$
- from here we can use Chebyshev's inequality and see that $P(|p - \alpha(\theta_1) + (1-\alpha)(\theta_2)| \geq \epsilon) \leq \frac{var(p)}{\epsilon^2} = \frac{\alpha\sigma_y^2}{n_1^2\epsilon^2} + \frac{(1-\alpha)\sigma_o^2}{n_2^2\epsilon^2}$
- we know that ϵ, α are fixed thus as $n_1, n_2 \to \infty$ this quantity will approach zero meaning the estimator is consistent

3. (Consistency of the sample median) Let $\tilde{x}_1, \tilde{x}_2, \ldots$ be a sequence of i.i.d. random variables from a distribution with median γ. We assume that γ is the only point that satisfies $F_{\tilde{x}_i}(\gamma) = 1/2$. Let $\widetilde{md}_n$ denote the median of the n first elements of the sequence. In this problem we establish that for any $\epsilon > 0$

$$\lim_{n\to\infty} \mathrm{P}\left(\left|\widetilde{md}_n - \gamma\right| \geq \epsilon\right) = 0, \tag{3}$$

so the sample median is a consistent estimator of the median. Specifically, the goal is to prove

$$\lim_{n\to\infty} \mathrm{P}\left(\widetilde{md}_n \geq \gamma + \epsilon\right) = 0, \tag{4}$$

because the same argument can be used to prove

$$\lim_{n\to\infty} \mathrm{P}\left(\widetilde{md}_n \leq \gamma - \epsilon\right) = 0, \tag{5}$$

and combining (4) and (5) yields (3).

(a) Let $\tilde{b}$ be the number of elements in $\{\tilde{x}_1, \ldots, \tilde{x}_n\}$ greater or equal to $\gamma + \epsilon$. Explain why

$$\mathrm{P}\left(\widetilde{md}_n \geq \gamma + \epsilon\right) \leq \mathrm{P}\left(\tilde{b} \geq \frac{n+1}{2}\right). \tag{6}$$

- $P(\tilde{m}_{dn} \geq \gamma + \epsilon)$ can be expressed in terms of $\tilde{b}$ such that if n is odd we can write $P(\tilde{m}d_n \geq \gamma + \epsilon) = P(\tilde{b} \geq \frac{n+1}{2})$
- if n is even then it must be the case that both $\frac{n}{2}$ and $\frac{n}{2}+1$ are greater than $\gamma+\epsilon$ so in this case $P(\tilde{m}d_n \geq \gamma + \epsilon) = P(\tilde{b} \geq \frac{n}{2} + 1) \leq P(\tilde{b} \geq \frac{n+1}{2})$
- thus as probabilities are always positive we can conclude $P(\tilde{m}_{dn} \leq \gamma + \epsilon) \leq P(\tilde{b} \geq \frac{n+1}{2})$ which was what we wanted to show.

(b) Use Chebyshev's inequality and (6) to prove (4). (Hint: By the assumption that γ is the only point that satisfies $F_{\tilde{x}_i}(\gamma) = 1/2$, there exists a constant $\epsilon' > 0$ such that for any i the probability that $\tilde{x}_i > \gamma + \epsilon$ is $\mathrm{P}\left(\tilde{x}_i > \gamma + \epsilon\right) = 1/2 - \epsilon' := \theta$.)

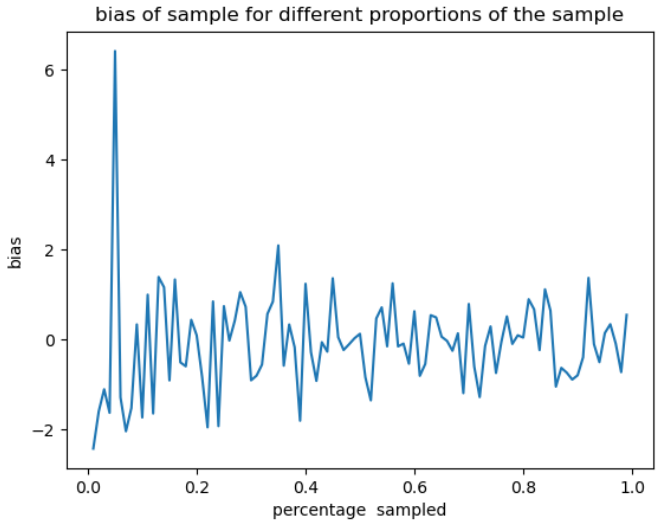
- first note that $E[\tilde{b}] = E[\frac{1}{n}\Sigma_{i=1}^{n} 1(x_i \geq \lambda + \epsilon)] = \frac{1}{n}\Sigma_{i=1}^{n} E[1(x_i \geq \lambda + \epsilon)] = \frac{1}{n}\Sigma_{i=1}^{n} P(x_i \geq \gamma + \epsilon) = 1/2 - \epsilon^*$
- we can use this to write Chebyshev's as $P(|\tilde{b} - 1/2 + \epsilon^*)| \geq \frac{n+1}{2} \leq \frac{var(\tilde{b})}{(\frac{n+1}{2})^2}$
- as we can see $\tilde{b}$ is bernuli we can write that $var(b) = E[\tilde{b}](1 - E[\tilde{b}]) = (1/2 - \epsilon^*)(1/2 + \epsilon^*)$
- and thus $P(|\tilde{b} - 1/2 + \epsilon^*)| \leq \frac{n+1}{2} \leq \frac{var(\tilde{b})}{(\frac{n+1}{2})^2} = \frac{(1/2-\epsilon^*)(1/2+\epsilon^*)}{(\frac{n+1}{2})^2}$ and as $n \to \infty$ we can see $\frac{(1/2-\epsilon^*)(1/2+\epsilon^*)}{(\frac{n+1}{2})^2} \to 0$

- if $\tilde{b}-1/2+\epsilon^* \geq 0$ we can write $P(|\tilde{b}-1/2+\epsilon^*)| \geq \frac{n+1}{2}) = P(\tilde{b}-1/2+\epsilon^* \geq \frac{n+1}{2}) = P(\tilde{b}| \geq \frac{n+1}{2} + 1/2 - \epsilon^*)$ then we can finally write $P(\tilde{m}d_n \geq \lambda + \epsilon + 1/2 - \epsilon^*) = P(\tilde{m}d_n \geq \lambda + \epsilon^{\text{“}}) \leq P(\tilde{b} \geq \frac{n+1}{2} + 1/2 - \epsilon^*) \leq \frac{(1/2-\epsilon^*)(1/2+\epsilon^*)}{(\frac{n+1}{2})^2}$ and thus as n approach's infinity we can see that $\lim_{n\to\infty} \mathrm{P}\left(\widetilde{md}_n \geq \gamma + \epsilon\right) = 0$ holds
-
- if $\tilde{b} - 1/2 + \epsilon^* < 0$ we can write $P(|\tilde{b} - 1/2 + \epsilon^*)| \geq \frac{n+1}{2}) = P(-\tilde{b} + 1/2 - \epsilon^* \geq \frac{n+1}{2}) = P(-\tilde{b}| \geq \frac{n+1}{2} - 1/2 + \epsilon^*) = P(\tilde{b} \leq -\frac{n+1}{2} + 1/2 - \epsilon^*)$ then we can finally write $P(\tilde{m}d_n \geq +1/2 - \epsilon^* - \lambda + \epsilon) = P(\tilde{m}d_n \geq \lambda + \epsilon^{\text{“}}) \leq P(\tilde{b} \geq \frac{n+1}{2} + 1/2 - \epsilon^*) \leq \frac{(1/2-\epsilon^*)(1/2+\epsilon^*)}{(\frac{n+1}{2})^2}$ and thus as n approach's infinity we can see that $\lim_{n\to\infty} \mathrm{P}\left(\widetilde{md}_n \geq \gamma + \epsilon\right) = 0$ holds

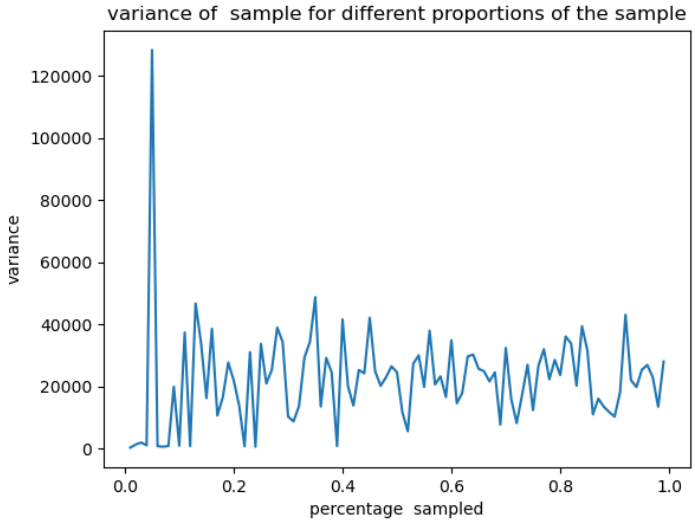
4. (Blood Pressure) The table in `cardio.csv` records the systolic blood pressure (*"ap_hi"*) of patients. Randomly sample subsets consisting of $0.1\%, 0.2\%, \cdots, 99.9\%$ of the full dataset.

   (a) Compute and plot the bias and variance of these subsets as a function of the number of samples. Interpret your findings.

i.



- as can be seen from the above plot sample bias goes down as the proportion of the population used in a sample. further it seems like sample bias is stabilizes with a sample that is around 20% of the population and stops falling

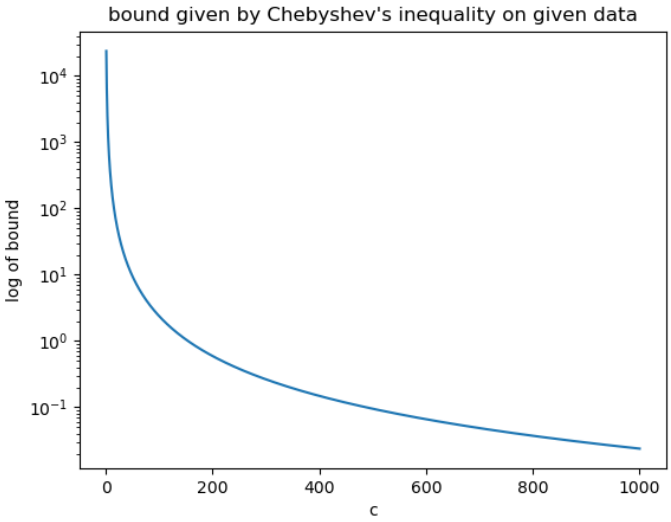


- as can be seen from the above plot sample variance goes down as the proportion of the population used in a sample. further it seems like sample variance is stabilizes with a sample that is around 20% of the population and stops falling

(b) Approximate the probability that *ap_hi* deviates from the corresponding population mean via Monte Carlo simulations, and compare it to the Chebyshev bound that we use to prove the law of large numbers.

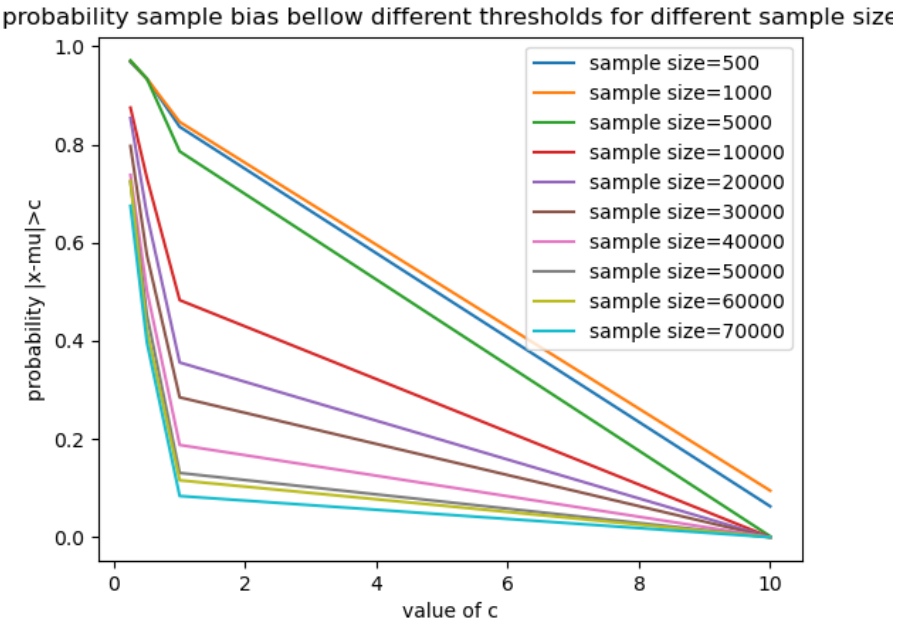
- i was not entirely sure how we were supposed to show this, but here are some figures that i think illustrate the relationship well.

i.



- the variance of our data is 23719.178472621654 which is very high. so thus we would need $\epsilon \geq 154.01031937055924$ to bound $P(|\tilde{x} - \mu| \geq \epsilon) \leq 1$
- so as can be seen from this chart and Chebyshev's inequality vies us a very louse bound for this data

ii.



- this graph shows the empirical probability calculated using Monte Carlo methods that the sample mean is bellow various values of ϵ

- each line shows how the path of this likelihood varies for different sample sizes
- the real take away of this, is we can establish a frailly tight bound using simulation

iii. so in conclusion for samples with high variance and a well defined mean, empirical methods can be better used to bound the probability that our estimator will differ from the population parameter more tightly than Chebyshev's inequality