CS265/CME309, Problem Set 3

SUNet ID(s):
Name(s):

By turning in this assignment, I agree by the Stanford honor code and declare that all of the writing is the work of my partner and I (discussion in larger groups is permissible).

Due by 11:59 PM on **Tuesday**, October 15th.

4. Suppose a class has n students. On a given day, assume that if the ith student is asked "What is the temperature today", their response is drawn from a Gaussian (Normal) distribution whose mean is the actual temperature, and variance is σ_i^2 (and each student's guess is independent from the guesses of the other students).

   (a) (4 points) If $\sigma_i = 1$ for all i, how accurate will the average of the n guesses be? [Feel free to use the fact that the sum of independent Gaussians is Gaussian: namely for X_i distributed according to $N(\mu_i, \sigma_i^2)$, if the X_i are independent, then $\sum_i X_i$ is distributed according to $N(\sum_i \mu_i, \sum_i \sigma_i^2)$. This fact can be proved by looking at the moment generating function of X, as in the previous problem.] In this part, and subsequent parts, it is up to you to quantify what notion of accuracy you want—reasonable options include 1) bounding the expected error, 2) bounding the probability that the returned answer deviates from the truth by more than a certain amount.

   SOLUTION: The average $X = \sum X_i/n$ is distributed according to $N(T, 1/n)$, where T denotes the true temperature, and the variance $1/n$ comes from the fact that $Var[X] = \frac{1}{n^2}\sum Var[X_i] = \frac{n}{n^2}$. Hence $\Pr[|X - T| \geq 1/\sqrt{n}] \approx 0.3$, since this is the probability that a Gaussian deviates from its expectation by at least a standard deviation.

   (b) (4 points) If $\sigma_i = 1$ for all i, how accurate can we expect the *median* of the guesses to be? Specifically, find some function $d(n)$ for which with probability at least 0.9, the median will be within distance $d(n)$ from the true temperature.

   SOLUTION: Define $\delta_n = \alpha/\sqrt{n}$ for some constant α that we will choose later. We wish to show that the median is within δ_n of the truth, T, with probability close to 1. We first consider the probability that the median is larger than $T + \delta_n$. Let Z_i denote the 0/1 random variable that is 1 if the ith guess, $X_i \geq T + \delta_n$. The Z_i's are independent, and $\Pr[Z_i = 1] \leq 1/2 - 0.2\delta_n$. To see where this bound comes from, note that $\Pr[X_i \in [T, T + c] \geq 0.2c$, for any $c \leq 1$ since the Gaussian density function of a standard Gaussian evaluates to $\frac{1}{\sqrt{2\pi}}e^{-1/2} \approx 0.24 > 0.2$ at 1 standard deviation.
   To finish the proof, $\mathbf{E}[\sum Z_i] \leq n(1/2 - 0.2\delta_n)$, and from the standard Chernoff bound from class (Corollary 5 and 6 in Lecture NOtes 6):

$$\begin{aligned}\Pr[median > T + \delta_n] &= \Pr[\sum Z_i \geq n/2] \leq \Pr[\sum Z_i \geq (1 + 0.4\delta_n)\,(n(1/2 - 0.2\delta_n))] \\ &\leq e^{-\frac{n(\frac{1}{2} - 0.2\delta_n)(0.4\delta_n)^2}{3}} = e^{-O(\alpha^2) + O(1/\sqrt{n})},\end{aligned}$$

   where both big Oh's hide only constant factors. A symmetric argument applies to the probability that the median is less than $T - \delta_n$, and hence for constant α, the probability the median is *not* within $\alpha/\sqrt{n}$ of the truth decays inverse exponentially with α^2.

   (c) (4 points) Suppose $\sigma_i^2 = 1$ for $i \leq n/2$, and $\sigma_i^2 = n$ for $i > n/2$. Are we better-off returning the mean of the guesses, or the median? Support your claim with a Chernoff bound or two and a brief discussion.

   SOLUTION: The variance of the mean will be $\frac{1}{n^2}(n/2 + n^2/2)) > 1/2$, and hence the mean of the guesses will be distributed according to $N(T, \sigma^2)$ for $\sigma^2 > 1/2$. Since the probability that a Gaussian is NOT within a standard deviation of its mean is roughly $1 - 0.68 \approx 0.3$, there will be a ≈ 0.3 probability that the mean is more than 1/4 from the true temperature.

We now analyze the median using the same approach as in the previous part. For this part, our analysis will hold as long as the second set of $n/2$ guesses is drawn from *any distribution that is symmetric about the true mean, even one of INFINITE variance!!!!* Let Z_i be the 0/1 random variable that is 1 if the ith duess is greater than $T + \delta_n$ for $\delta_n = \alpha/\sqrt{n}$. For $i \in \{1, \ldots, n/2\}$, we have $\Pr[Z_i = 1] \leq 1/2 - 0.2\delta_n$ as in the previous part. For $i \in \{n/2 + 1, \ldots, n\}$, we have that $\Pr[Z_i = 1] \leq 1/2$, by the symmetry of the distribution of the guesses about the true temperature. So, $Z = \sum Z_i$ is a sum of independent 0/1 random variables, and $\mathbf{E}[Z] \leq (n/2)(1/2 - 0.2\delta_n) + (n/2)(1/2) = \frac{n}{2} - 0.1\delta_n$, and we are trying to bound the probability that $Z \geq n/2$. An identical argument to the one in the previous part applies, yielding the same high-level punchline that the probability that the median deviates from the truth by more than $\alpha/\sqrt{n}$ decays inverse exponentially with α^2, and hence we would still expect that the median, say, is within $10/\sqrt{n}$ from the truth.

In the setting of this part of the problem, we would be MUCH better off relying on the median, versus the mean, as we expect it to have accuracy $O(1/\sqrt{n})$, versus the mean which has accuracy only constant (not vanishing as n gets large).

(d) BONUS (2 points): Suppose you are given the list of σ_i^2's, and hence you know the variance of each of the n guesses. What is the "best" estimate of the temperature, where, for example "best" could be defined as minimizing the expected square of the error of the guess.

(e) DOUBLE BONUS: Suppose you are given the *set* of σ_i^2's, but are not told which student has each variance. What is the "best" estimate of the temperature? [This is more of a research problem/food for thought, and we will not grade it.]