CS 747 (Autumn 2025) Week 3 Test (Batch 1)

5.35 p.m. – 6.00 p.m., August 26, 2025, LA 001

Name: ________________________________ Roll number: ________________

Note. There is one question (with two parts) in this test. You can use the space on both pages for your answer. Draw a line (either vertical or horizontal) and do all your rough work on one side of it.

Question 1. X is a Bernoulli random variable with mean $p \in (0, 1)$, and Y is a Bernoulli random variable with mean $q \in (0, p)$. Notice that $0 < q < p < 1$. For $n \geq 1$, we obtain $X_1, X_2, \ldots, X_n$ as i.i.d. samples of X, and $Y_1, Y_2, \ldots, Y_n$ as i.i.d. samples of Y. Also, X_i and Y_j are independent for $i, j \in \{1, 2, \ldots, n\}$. We are interested in the probability that the sum of the "X" variables exceeds the sum of the "Y" variables: that is.

$$\alpha \stackrel{\text{def}}{=} \mathbb{P}\left\{\sum_{i=1}^{n} X_i > \sum_{j=1}^{n} Y_j\right\}.$$

1a. Give an *exact expression* for α in terms of p, q, and n. You are free to define and use helper variables. [1 mark]

1b. Obtain a simpler expression (again in terms of p, q, and n) for an *upper bound* on α by suitably applying Hoeffding's Inequality. [2 marks]

Answer 1a. $\sum_{i=1}^n X_i$ exceeds $\sum_{i=1}^n Y_i$ when the former is some $\ell \in \{1, 2, \ldots, n\}$ and the latter is some $m \in \{0, 2, \ldots, \ell - 1\}$. The probability of such an occurrence (considering all qualifying choices of ℓ and m) is

$$\alpha = \sum_{\ell=1}^{n} \sum_{m=0}^{\ell-1} \binom{n}{\ell} p^\ell (1-p)^{n-\ell} \binom{n}{m} q^m (1-q)^{n-m}.$$

Answer 1b. It is useful to define $Z_i \stackrel{\text{def}}{=} X_i - Y_i$ for $i \in \{1, 2, \ldots, n\}$, since then, the condition $\sum_{i=1}^n X_i > \sum_{j=1}^n Y_j$ can equivalently be written as $\sum_{i=1}^n Z_i > 0$. Now, Z_i takes values from the set $\{-1, 0, 1\}$. We scale Z_i so we can deal with a random variable with support in $[0, 1]$. In particular, define $\bar{Z}_i \stackrel{\text{def}}{=} \frac{Z_i+1}{2}$, and also define $\bar{Z} = \frac{1}{n}\sum_{i=1}^n \bar{Z}_i$. The new variables $\bar{Z}_i$ take values from the set $\{0, \frac{1}{2}, 1\}$, which is a subset of $[0, 1]$. We observe that the condition $\sum_{i=1}^n Z_i > 0$ is equivalent to the condition $\bar{Z} > \frac{1}{2}$. Collecting our sequence of steps thus far, we have established that

$$\alpha = \mathbb{P}\left\{\bar{Z} > \frac{1}{2}\right\},$$

which is set up for applying Hoeffding's inequality since $\bar{Z}$ is the average of n i.i.d. random variables with support in $[0, 1]$. The expected value of $\bar{Z}$ is $\frac{1}{2} + \frac{p-q}{2}$. Hence, we are looking to upper-bound $\mathbb{P}\{\bar{Z} > \mathbb{E}[\bar{Z}] + \frac{q-p}{2}\}$. Notice that $q - p$ is *negative.* Hence, we cannot use Hoeffding's Inequality to upper-bound α. However, we can use it to lower-bound α. Indeed

$$\begin{aligned}
\alpha &= \mathbb{P}\{\bar{Z} > \mathbb{E}[\bar{Z}] + \frac{q-p}{2}\} \\
&= 1 - \mathbb{P}\{\bar{Z} \leq \mathbb{E}[\bar{Z}] - \frac{p-q}{2}\} \\
&\geq 1 - \exp\left(-\frac{n}{2}(p-q)^2\right)
\end{aligned}$$

As such, the only "natural" upper bound for α is 1, since α is a probability.

Note. The question should ideally have had α defined as $\mathbb{P}\left\{\sum_{i=1}^n X_i \leq \sum_{j=1}^n Y_j\right\}$ rather than as $\mathbb{P}\left\{\sum_{i=1}^n X_i > \sum_{j=1}^n Y_j\right\}$, since such a definition would have facilitated a genuine exponentially-decreasing upper bound for part 1b. The instructor regrets the typo. Solutions will be graded based on the student's attempt to interpret α as the probability that the average of i.i.d. random variables exceeds a threshold.

CS 747 (Autumn 2025) Week 3 Test (Batch 2)

6.15 p.m. – 6.40 p.m., August 26, 2025, LA 001

Name: ________________________________ Roll number: ________________

Note. There is one question (with two parts) in this test. You can use the space on both pages for your answer. Draw a line (either vertical or horizontal) and do all your rough work on one side of it.

Question 1. You are given a Bernoulli random variable X with mean $p \neq \frac{1}{2}$. You must write an algorithm to determine whether $p > \frac{1}{2}$ or $p < \frac{1}{2}$. The algorithm does not have knowledge of p; however it is allowed to repeatedly sample X for information. The algorithm is provided a parameter $\delta \in (0, 0.1)$ as input. It is required that

- if indeed $p > \frac{1}{2}$, then with probability at least $1 - \delta$, the algorithm stop and print out "High";
- if indeed $p < \frac{1}{2}$, then with probability at least $1 - \delta$, the algorithm stop and print out "Low".

1a. Describe an algorithm that satisfies the requirement specified above. Keep in mind that p can be arbitrarily close to $\frac{1}{2}$, although it cannot be exactly $\frac{1}{2}$. [1 mark]

1b. Explain why your algorithm will terminate with probability at least $1 - \delta$. Also explain why its answer is the correct one with probability at least $1 - \delta$. [2 marks]

Answer 1a. We ensure that X is sampled at least once. In general, consider the situation that X has been sampled $t \geq 1$ times. We define $\alpha_t \stackrel{\text{def}}{=} \sqrt{\frac{1}{2t} \ln \frac{2t^2}{\delta}}$. Let $\hat{p}^t$ be the empirical mean of the samples from the first $t-1$ samples. Define

$$\text{ucb}^t \stackrel{\text{def}}{=} \hat{p}^t + \alpha_t \text{ and } \text{lcb}^t \stackrel{\text{def}}{=} \hat{p}^t + \alpha_t.$$

Our algorithm is as follows:

1. if the interval $[\text{lcb}^t, \text{ucb}^t]$ contains $\frac{1}{2}$, obtain a new sample from X and proceed to round $t+1$;
2. if $\text{ucb}^t < \frac{1}{2}$, print "Low" and terminate, whereas if $\text{lcb}^t > \frac{1}{2}$, print "High" and terminate.

Answer 1b. We shall provide our answers for the case that the true mean p exceeds $\frac{1}{2}$. A similar argument shall hold by symmetry for the case $p < \frac{1}{2}$.

Let B be the ("bad") event that there exists $t \geq 1$ such that $\text{ucb}^t < p$. By a union bound over t, the probability of B is upper-bounded as follows:

$$\mathbb{P}\{B\} \leq \sum_{t=1}^{\infty} \mathbb{P}\{\text{ucb}^t < p\} = \sum_{t=1}^{\infty} \mathbb{P}\{\hat{p}^t < p - \alpha_t\} \leq \sum_{t=1}^{\infty} \exp(-2t(\alpha_t)^2) = \sum_{t=1}^{\infty} \frac{\delta}{2t^2} < \delta.$$

The width of the interval $[\text{lcb}^t, \text{ucb}^t]$ is $2\alpha_t$, which monotonically decreases to 0 as t increases. Hence there exists some $t^\star$ such that for all $t \geq t^\star$, the size of the interval is smaller than $p - \frac{1}{2}$. If B does not occur, it follows that the algorithm must terminate with lcb^t exceeding $\frac{1}{2}$—which would lead to printing the correct answer "High". Hence, with probability at least $1 - \delta$, our algorithm terminates and prints out the correct answer.