Homework 1

Ercoli, Fabio Massimo
802397

Mai Mihai, Cristian Andrei
1942925

November 28^{st} 2024

1 Assignment 1

1.1 Setup the problem

Sample space We start from a sample space of 200 subjects. All measured heights vary in the interval [160, 190]. The average height in the group is 180 cm.

Null Hypothesis The null hypothesis *H_0* should be a precise and complete statement about the population from which the sample comes.

In this case we can use the Professor Cooper's thesis as *H_0* , and the Professor Hofstadter's thesis as the alternative hypothesis *H_1*.

The *H_1*, on the other hand, must not define a precise alternative statement. As in the case of *Testing ESP* presented in the course, we may know very little about the statistical properties of *H_1*.

H_0 := the heights of the population are distributed uniformly in the interval [160, 190].

H_1 := the heights of the population are (simply) **not** distributed uniformly in the interval [160, 190].

P-value and significance level We define the p-value as the conditional probability of having the evidence returned from the experiment E, assuming true the null hypothesis:

$$p = P(E \big| H_0 = true) \tag{1}$$

Before we start to calculate this probability we need to define the significance level α we're going to apply to this experiment. The value denotes the threshold of the p-value below which we reject the null hypothesis. In this case we set α to 0.05.

So all the analysis consists in estimating the probability of having the sample space we got, assuming the null hypothesis a ground true.

1.2 Apply the Hoeffding's inequality

To apply the *Chernoff bound*, probably we need a series of mutually independent random variables assuming values exactly in the set $\{0, 1\}$. Scaling the range of the heights [160, 190], we would obtain rational values in the range $[0, 1]$, so the authors are not really sure that this can be done.

For this reason we decided to apply the *Hoeffding's inequality*:

$$P(X - E[X] \geq t) \leq e^{-\frac{2t^2}{\sum_{i=1}^{n}(b_\mathrm{i} - a_\mathrm{i})^2}} \tag{2}$$

That is similar to the *Chernoff bound*, and according to its Wikipedia page[1]:

- The independent random variables X_i may take values in the ranges $[\mathrm{a_i}, \mathrm{b_i}]$.
- We don't need to scale the range to $[0, 1]$.
- The inequalities also hold when the X_i have been obtained using sampling without replacement. It is nice to have this property since we don't have this information from the null hypothesis.

Define the random variables We can define the X_i as the i-th subject of the sample space.

Applying the null hypothesis, we can say that the population are distributed uniformly at random in the interval [160, 190]. So we can assume that they are mutually independent.

The size of the sample space n is equal to 200. The expected value of each X_i, $E[X_\mathrm{i}]$, is equal to 175.

Now we define a new random variable named X defined as the sum of all 200 X_i:

$$X = \sum_{i=1}^{200} X_\mathrm{i} \tag{3}$$

Compute the p-value upper-bound Applying the linearity of expectation we can deduce that its expected value ($E[X]$) is equal to $175 \cdot 200$. The value of X we got from the sample space is $180 \cdot 200$. This means that the value of t to use in the inequality is $5 \cdot 200$.

Computing the right side of the inequality we have:

$$P(X - (175 \cdot 200) \geq (5 \cdot 200)) \leq e^{-\frac{2 \cdot 5^2 \cdot 200^2}{200 \cdot 30^2}} \tag{4}$$

$$P(X - (175 \cdot 200) \geq (5 \cdot 200)) \leq e^{-\frac{100}{9}} = 0.000015 \tag{5}$$

[1] https://en.wikipedia.org/wiki/Hoeffding

1.3 Conclusions

Since:

$$P(E \big| H_0 = true) \leq P(X - E[X] \geq t) \leq 0.000015 \tag{6}$$

The p-value is is much less that the significance level we defined at the beginning of the experiment (0.05). So we can reject the null hypothesis, claiming that the Professor Hofstadter was right, with very high probability (1 - p-value), the type 1 error seems to be very unlikely.

2 Assignment 2

2.1 Bernoulli random variables

The first idea is to define n Bernoulli random variables, one for each execution of the algorithm.

$$X_{\text{i}} := \begin{cases} 1, & \text{if i}^{\text{th}} \text{ execution returns the correct answer} \\ 0, & \text{otherwise} \end{cases} \tag{7}$$

According the assignment we can say that:

$$P[X_{\text{i}} = 1] = \frac{1}{2} + \epsilon \tag{8}$$

$$P[X_{\text{i}} = 0] = \frac{1}{2} - \epsilon \tag{9}$$

We also define the random variable X as:

$$X := \begin{cases} 1, & \text{if the full execution returns the correct answer} \\ 0, & \text{otherwise} \end{cases} \tag{10}$$

Notice that X = 1 if and only if a majority of X_{i} returns 1.

2.2 Define the aggregation function

Let's define a function **f** returning the random variable that is the sum of all the X_{i}:

$$f(X_1, ..., X_{\text{n}}) := \sum_{i=1}^{n} X_{\text{i}} \tag{11}$$

The expected value of **f**, can be easily computed, applying the linearity of expectations:

$$E[f(X_1, ..., X_{\text{n}})] := \frac{n}{2} + n \cdot \epsilon \tag{12}$$

Notice that it is possible to express the variable X in terms of **f**:

$$X := \begin{cases} 1, & f(X_1, ..., X_{\text{n}}) > n/2 \\ 0, & \text{otherwise} \end{cases} \tag{13}$$

2.3 Apply McDiarmid's inequality

Furthermore, the function **f** satisfies the *bounded differences property* with bound equals to 1 for all variables X_1, ... , X_n, since substituting the value of any single coordinate may change the returned value by **f** at most by 1 or -1.

The *bounded differences property* is the first requirement to apply the McDiarmid's inequality to the function **f**. The second requirement is that the parameter of **f** should be independent random variables, in our case we know that, since we defined X_1, ... , X_n as Bernoulli random variable.

This is the McDiarmid's inequality we could use:

$$P[f(X_1, ..., X_\mathrm{n}) - E[f(X_1, ..., X_\mathrm{n})] \leq -\alpha] \leq e^{-\frac{2\alpha^2}{\sum_{i=1}^{n} c_\mathrm{i}{}^2}} \tag{14}$$

Applying $\alpha = n \cdot \epsilon$, substituting $E[f()] := \frac{n}{2} + n \cdot \epsilon$, knowing that all the bounds for *bounded differences property* (c_i) are 1, we have:

$$P[f(X_1, ..., X_\mathrm{n}) - \frac{n}{2} - n \cdot \epsilon \leq -n \cdot \epsilon] \leq e^{-\frac{2n^2 \cdot \epsilon^2}{n}} \tag{15}$$

$$P[f(X_1, ..., X_\mathrm{n}) \leq \frac{n}{2}] \leq e^{-2n \cdot \epsilon^2} \tag{16}$$

Then we have an upper bound of the probability that the result produced by the entire process is wrong:

$$P[X = 0] \leq e^{-2n \cdot \epsilon^2} \tag{17}$$

2.4 Conclusions

Given any, arbitrary, $\delta \in$ (0, 1) if we want that the answer is correct with probability 1 - δ, we can impose that the answer is not correct with probability at most δ.

The condition is satisfied for:

$$\delta \geq e^{-2n \cdot \epsilon^2} \tag{18}$$

$$\frac{1}{\delta} \leq e^{2n \cdot \epsilon^2} \tag{19}$$

$$\ln(\frac{1}{\delta}) \leq 2n \cdot \epsilon^2 \tag{20}$$

$$n \geq \frac{1}{2 \cdot \epsilon^2} \ln(\frac{1}{\delta}) \tag{21}$$

The C we found (using our bound) is 1/2.

3 Assignment 3

3.1 Define the unbiased estimator

We can use the hash function used to define the locality sensitive function for the cosine distance that we have seen at the lecture, to introduce n random variables X_1, X_2, ... , X_n, defined as:

$$X_{\mathrm{i}}(x,y) := \begin{cases} 1, & \text{if } sign(u_{\mathrm{i}} \cdot x) \neq sign(u_{\mathrm{i}} \cdot y) \\ 0, & \text{otherwise} \end{cases} \tag{22}$$

Where x and y any two vectors in R^{d}. u_i is the i^{th} vector chosen uniformly at random to generate the i^{th} signature of each vector.

We know that:

$$\forall i \; P[X_{\mathrm{i}} = 1] = P[sign(u_{\mathrm{i}} \cdot x) \neq sign(u_{\mathrm{i}} \cdot y)] = \frac{\theta(x,y)}{\pi} \tag{23}$$

Where $\theta(x,y)$ is the exact (the real one - not the estimated one) cosine distance between the vectors x and y.

According to the definition of expected value we can say that:

$$\forall i \; E[X_{\mathrm{i}} = 1] = \frac{\theta(x,y)}{\pi} \tag{24}$$

Notice that, since the vectors u are chosen uniformly at random in the sphere of length 1 of R^{d}, the $\mathbf{X_1}$, X_2, ... , X_n match the definition of *Bernoulli random variables* where p in this case is exactly equal to $\frac{\theta(x,y)}{\pi}$.

We define the cumulative function X:

$$X = \sum_{i=1}^{n} X_{\mathrm{i}}(x,y) \tag{25}$$

Applying the linearity of expectations, we have:

$$E[X] = \frac{n \cdot \theta(x,y)}{\pi} \tag{26}$$

Finally, we define our estimator $\hat{\theta}(x,y)$:

$$\hat{\theta}(x,y) = \frac{\pi \cdot X}{n} \tag{27}$$

Since π and n are constants we can say that:

$$E[\hat{\theta}(x,y)] = \theta(x,y) \tag{28}$$

That is the exact definition of unbiased estimator [2].

[2] See https://en.wikipedia.org/wiki/Bias_of_an_estimator

3.2 General problem of estimating θ with $\hat{\theta}$

We want to know find some bound of the probability that our estimation goes beyond a given threshold as function of n. In particular we expect that the higher is n, the better will be the estimation. But we want some numbers.

There are several bounds that can be used. In particular, here we decided to apply *the Chernoff-Hoeffding inequality*[3] to the cumulative function X we defined in the previous chapter.

Since our cumulative function X is the sum of n Bernoulli random variables, we can apply the bound used to estimate confidence intervals:

$$P(X(x,y) - p \cdot n \geq \epsilon \cdot n) \leq e^{-2\epsilon^2 n} \tag{29}$$

In our case we have:

$$P(X(x,y) - \frac{n \cdot \theta(x,y)}{\pi} \geq \epsilon \cdot n) \leq e^{-2\epsilon^2 n} \tag{30}$$

Finally, multiplying both sides of the inner inequality by $\frac{\pi}{n}$ we get a bound for our estimator:

$$P(\hat{\theta}(x,y) - \theta(x,y) \geq \pi \cdot \epsilon) \leq e^{-2\epsilon^2 n} \tag{31}$$

3.3 Solve the point (b)

The problem is about correctly classifying θ whether:

- $\theta \leq \frac{\pi}{2} - \phi$: the cosine distance is below a given threshold. We want to classify this distance as **near**.
- $\theta \geq \frac{\pi}{2} + \phi$: the cosine distance is above another given threshold. We want to classify this distance as **far**.

With ϕ as arbitrary small value grater then 0. We want to find the value of n such that the probability of misclassifying θ is at most δ.

We can notice that, independently by the value of ϕ, the problem is centered on the median value of δ of $\frac{\pi}{2}$, so we can expect the same probability of misclassifying the two cases, we can say that false positives and false negatives are equal (in this case) in expectation. This means that we can bound δ considering just one of the two possible errors (δ would be equal to the symmetric case).

Usually the Locality Sensitive Hashing (LSH) technique can improve a lot the result of the classification, if the goal is the one to classify and not to estimate the value of $\theta(x,y)$. Anyway here the hints of the problem seems to suggest to use instead the estimator we've defined in the previous chapter ($\hat{\theta}(x,y)$) to solve this classification problem.

[3] https://en.wikipedia.org/wiki/Hoeffding%27s_inequality

A possible way to use the estimator as classifier is to classify the distance as:

- **near** if $\hat{\theta}(x,y) \leq \frac{\pi}{2}$
- **far** if $\hat{\theta}(x,y) > \frac{\pi}{2}$ [4]

Estimate a bound for misclassifying far distances This is the probability of classifying $\theta(x,y)$, when instead the real value of $\theta(x,y)$ is above the threshold is:

$$P[\hat{\theta}(x,y) \leq \frac{\pi}{2} \mid \theta(x,y) \geq \frac{\pi}{2} + \phi] \tag{32}$$

This value is at least:

$$P[\hat{\theta}(x,y) \leq \frac{\pi}{2} \mid \theta(x,y) = \frac{\pi}{2} + \phi] \tag{33}$$

Because when $\theta(x,y)$ is exactly $\frac{\pi}{2} + \phi$ this is the worst case possible to correctly classify a **far** distance. Using the result obtained in the last chapter we know that:

$$P[\frac{\pi}{2} - \theta(x,y) \geq \phi] \leq e^{-\frac{2 \cdot \phi^2 \cdot n}{\pi^2}} \tag{34}$$

Considering the symmetric case:

$$P[\hat{\theta}(x,y) \geq \frac{\pi}{2} \mid \theta(x,y) \leq \frac{\pi}{2} + \phi] \tag{35}$$

We get a similar result:

$$P[\theta(x,y) - \frac{\pi}{2} \geq \phi] \leq e^{-\frac{2 \cdot \phi^2 \cdot n}{\pi^2}} \tag{36}$$

The two bounds are conditioned on two disjointed events:

- $\theta(x,y) \geq \frac{\pi}{2} + \phi$
- $\theta(x,y) < \frac{\pi}{2} - \phi$

$$\delta \geq e^{-\frac{2 \cdot \phi^2 \cdot n}{\pi^2}} + e^{-\frac{2 \cdot \phi^2 \cdot n}{\pi^2}} = 2e^{-\frac{2 \cdot \phi^2 \cdot n}{\pi^2}} \tag{37}$$

$$\frac{2}{\delta} \leq e^{\frac{2 \cdot \phi^2 \cdot n}{\pi^2}} \tag{38}$$

[4] Notice that in case is equivalent to use the equal for the first or the second inequalities

$$ln(\frac{2}{\delta}) \leq \frac{2 \cdot \phi^2 \cdot n}{\pi^2} \tag{39}$$

$$n \geq \frac{\pi^2}{2 \cdot \phi^2} ln(\frac{2}{\delta}) \tag{40}$$

3.4 Solve the point (c)

If the average cosine similarity between any pair of points in S is below 0.01 in absolute value, it means that given any pair of points in S we expect them to be almost orthogonal. This phenomenon seems to related to what we call the curse of the dimensionality[5].

This means that is very difficult in general to classify correctly the distances between them as far or near. In particular, in other to have a good classification, in general, we need n to be very high.

In order to have an idea of the value of n required to be use, we can use the result from the previous chapter in which δ we can use 0.05 and ϕ can be derived imposing the average cosine similarity less or equal to 0.01:

$$cos(\frac{\pi}{2} + \phi) \leq 0.01 \tag{41}$$

For $\pi < \frac{\pi}{2}$:

$$sin(\phi) \leq 0.01 \tag{42}$$

$$\phi \leq arcsin(0.01) \simeq 0.01 \tag{43}$$

So we can impose n to be:

$$n \geq \frac{\pi^2}{2 \cdot 0.01^2} ln(\frac{2}{0.05}) \simeq 182038.90 \tag{44}$$

That is an enormous value.

[5] https://en.wikipedia.org/wiki/Curse_of_dimensionality