First homework Big Data Computing

Andrea Morelli

October 2022

1

2

2.1 a

We have to compute $Var(S)$ with S being $S = \frac{1}{n}\sum_{i=1}^{n} X_i$.
We start with $Var(\frac{1}{n}\sum_{i=1}^{n} X_i)$
$Var\left(\frac{X_1}{n} + \frac{X_2}{n} + \cdots \frac{X_n}{n}\right)$
Then being the variance a squared function we can use the property of the variance $Var(aX) = a^2 Var(X)$
and derive $\frac{1}{n^2}Var\left(X_1\right) + \frac{1}{n^2}Var\left(X_2\right) + \cdots + \frac{1}{n^2}Var\left(X_n\right)$
and being $Var(X_i) = \sigma^2$ we obtain
$Var(S) = \frac{n\sigma^2}{n^2} = \frac{\sigma}{n}$

2.2 b

Given $Z = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - S\right)^2$i have to compute $E[Z]$ as a function of n and σ^2.
So i have to compute: $E\left(Z\right) = E\left[\frac{\sum_{i=1}^{n}(X_i - S)^2}{n}\right]$.
I develop the square of a binomial term

$$E\left[\frac{\sum_{i=1}^{n}\left(X_i^2 - 2X_iS + S^2\right)}{n}\right]$$

Then distribute the sum

$$E\left[\frac{\sum_{i=1}^{n} X_i^2 \ - \sum_{i=1}^{n} 2X_iS + \sum_{i=1}^{n} S^2}{n}\right]$$

Then using $S = \frac{1}{n}\sum_{i=1}^{n} X_i$

$$E\left[\frac{(\sum_{i=1}^{n} X_i^2) \ - 2nS^2 + nS^2}{n}\right]$$

$$E\left[\frac{(\sum_{i=1}^{n} X_i^2) \ - nS^2}{n}\right]$$

Then using the linearity of the expectation

$$E[Z] = \frac{(\sum_{i=1}^{n} E[X_i^2]) - nE[S^2]}{n} \tag{1}$$

Then using the property of the variance $Var(X) = E\left[X^2\right] - (E[X])^2$ we have

$$E[X^2] = \sigma^2 + \mu^2$$

If applied on S and using the result of the "a" point we derive

$$E[S^2] = \frac{\sigma^2}{n} + \mu^2$$

Now applying those on 1 we obtain

$$E[Z] = \frac{(\sum_{i=1}^{n} \sigma^2 + \mu^2) - n(\frac{\sigma^2}{n} + \mu^2)}{n}$$

Developing everything

$$\frac{n\sigma^2 + n\mu^2 - \sigma^2 - n\mu^2}{n}$$

Obtaining

$$E[Z] = \frac{(n-1)\sigma^2}{n}$$

3

In order to solve the third assigment i used the chernoff bound. We want to estimate the fraction p of people that have been infected by a virus in a given population. So we perform n tests sampled uniformly and independently at random from the population and each test will return the correct answer every time. Given $0 < \varepsilon < 1$ *and* $0 < \delta < 1$ We want to find $P(|\hat{p} - p| > \varepsilon p) < \delta$, whenever $p > \theta$, where $\hat{p}$ is our estimator of p.

i start by defining $X_i = \begin{cases} 1 & if\ infected\ with\ proability\ p \\ 0 & if\ not\ infected\ with\ probability\ 1-p \end{cases}$

and $X = \sum_{i=1}^{n} X_i$.$\hat{p}$ that is our estimator will be $\frac{1}{n}X$ (sample mean)and will have a mean equal to p.

Using those we can use also use the chernoff bound:

For $\varepsilon > 0 : \mathbf{P}[X \geq (1+\varepsilon)\mu] < e^{-\frac{\varepsilon^2}{3}\mu}$ where μ is the mean of X and X is $\sum_{i=1}^{n} X_i$ with X_i being an indipendent poisson trial.

Using $\hat{p}$ and p that is its mean we can write:

$$P(|\hat{p} - p| > \varepsilon p) < 2e^{-\frac{\varepsilon^2}{3}p}$$

Use the 2 to take into account the module so both the tails of the distribution and not just one.
To use the form used by the chernoff bound we transform in the follwing way by just scaling everything with n:

$$P(|X - np| > \varepsilon np) < 2e^{-\frac{\varepsilon^2}{3}np}$$

Being this the maximum value it can assume we can write that :

$$2e^{-\frac{\varepsilon^2}{3}np} < \delta$$

and solving for n we can obtain the minimum number of n such that the initial condition holds.

$$n > \ln\left(\frac{\delta}{2}\right) \cdot \frac{3}{\varepsilon^2 \rho}$$

4

We can start by defining the null hyphothesis $H_0 = P(|\hat{\mu} - \mu| = 0)$ with μ being the mean of the supposed uniform distribution 175 and $\hat{\mu}$ being the sample mean $\hat{\mu} = \frac{\sum_{i=0}^{n} X_i}{n} = 180$ with n being 200 and X_i assuming values between 160 and 190 with probability $\frac{1}{30}$ and 0 outside this range and having a mean of 175.
We can then use the theorem 1.1 linked in the assignment 3.
Let $X = \sum_{i=0}^{n} X_i$ where $X_i \in [n]$ are indipendetly distributed in $[0, 1]$.
Then for $0 < \varepsilon < 1$

$$Pr[X > (1+\epsilon)\mathrm{E}[X]] \leq \exp\left(-\frac{\epsilon^2}{3}\mathrm{E}[X]\right)$$

$$Pr[X < (1-\epsilon)\mathrm{E}[X]] \leq \exp\left(-\frac{\epsilon^2}{2}\mathrm{E}[X]\right)$$

We can rescale the variables X_i in order to have values between $[0, 1]$ applying this simple normalization $Z_i = \frac{X_i - minValue}{maxValue - minValue}$ with $minValue = 160$ and $maxValue = 190$ obtaining $Z = \sum_{i=0}^{n} Z_i$ with $E[Z_i] = \frac{1}{2}$ and $E[Z] = n\frac{1}{2}$.
$\hat{\mu}$ rescaled becomes $\frac{180-160}{30} = \frac{2}{3}$ and ε in order for $(1+\varepsilon)\frac{1}{2}$to become $\frac{2}{3}$ has to be $\frac{1}{3}$ Now we can apply the theorem on Z.

$$P(Z > (1+\varepsilon)E[Z]) <= e^{-\frac{\varepsilon^2 E[Z]}{3}}$$

That becomes

$$P(Z > (1+\frac{1}{3})200\frac{1}{2}) <= e^{-\frac{\frac{1}{3}^2 200\frac{1}{2}}{3}}$$

And

$$P(Z < (1-\varepsilon)E[Z]) <= e^{-\frac{\varepsilon^2 E[Z]}{2}}$$

That becomes

$$P(Z < (1 - \frac{1}{3})200\frac{1}{2}) <= e^{-\frac{\frac{1}{3}^2 200\frac{1}{2}}{2}}$$

Summing those 2 results together we can derive

$$P(|Z - E[Z]| >= \varepsilon E[Z]) <= e^{-\frac{\varepsilon^2 E[Z]}{2}} + e^{-\frac{\varepsilon^2 E[Z]}{3}}$$

that is 0.03
Now having the null hypothesis being that the mean and the sample mean are the same we can find that we observe a case in which we are in the most extreme 3% of the distribution(3% including both the tails) so with a confidence interval of 95% we can reject the null hyphothesis and accept the alternative one that is the mean and the sample mean can't be the same.Now noticing that the uniform distribution has the mean at the center in 175 the heights of the people can't follow this distribution but have to be of a different type.So my argument is in favor of Professor Wolowoitz and rejects the claim of Professor Cooper.