Assignment 2

Section a

We have n independent samples $X_1, ..., X_n$ from a distribution with expectation μ and variance σ^2, and the relative sample mean $S = \frac{1}{n}\sum_{i=1}^{n} X_i$. In order to compute $\mathrm{Var}(S)$, leveraging on the independence of the samples, we can apply the following basic properties

$$\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$$

$$\mathrm{Var}(aX) = a^2 \,\mathrm{Var}(X)$$

Thus we have

$$\begin{aligned}
\mathrm{Var}(S) &= \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \\
&= \frac{1}{n^2}\,\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) \\
&= \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{Var}(X_i) \\
&= \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}
\end{aligned}$$

Section b

Now, considering the sample variance $Z = \frac{1}{n}\sum_{i=1}^{n}(X_i - S)^2$, we must compute $\mathrm{E}[Z]$. The steps of the computation are based on the linearity of expectation, the values of $\mathrm{E}[S]$ and $\mathrm{Var}[S]$ and the following consideration:

$$\begin{aligned}
\mathrm{Var}(X) &= \mathrm{E}[X^2] - (\mathrm{E}[X])^2 \\
\mathrm{E}[X^2] &= \mathrm{Var}(X) + (\mathrm{E}[X])^2
\end{aligned}$$

The derivation of $\mathrm{E}[Z]$ is the following

$$
\begin{aligned}
\mathrm{E}[Z] &= \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - S)^2\right] \\
&= \frac{1}{n}\,\mathrm{E}\left[\sum_{i=1}^{n}(X_i - S)^2\right] \\
&= \frac{1}{n}\,\mathrm{E}\left[\sum_{i=1}^{n}(X_i^2 - 2X_iS + S^2)\right] \\
&= \frac{1}{n}\,\mathrm{E}\left[\sum_{i=1}^{n}X_i^2 - \sum_{i=1}^{n}2X_iS + \sum_{i=1}^{n}S^2\right] \\
&= \frac{1}{n}\,\mathrm{E}\left[\sum_{i=1}^{n}X_i^2 - 2S\sum_{i=1}^{n}X_i + nS^2\right] \\
&= \frac{1}{n}\,\mathrm{E}\left[\sum_{i=1}^{n}X_i^2 - 2nS^2 + nS^2\right] \\
&= \frac{1}{n}\,\mathrm{E}\left[\sum_{i=1}^{n}X_i^2 - nS^2\right] \\
&= \frac{1}{n}\left(\mathrm{E}\left[\sum_{i=1}^{n}X_i^2\right] - \mathrm{E}\left[nS^2\right]\right) \\
&= \frac{1}{n}\left(n\,\mathrm{E}\left[X_1^2\right] - n\,\mathrm{E}\left[S^2\right]\right) \\
&= \mathrm{E}[X_1^2] - \mathrm{E}[S^2] \\
&= \mathrm{Var}[X_1] + (\mathrm{E}[X_1])^2 - \mathrm{Var}[S] - (\mathrm{E}[S])^2 \\
&= \sigma^2 + \mu^2 - \frac{\sigma^2}{n} - \mu^2 = \frac{n-1}{n}\sigma^2
\end{aligned}
$$

Assignment 3

Let T_i be the variable that corresponds to the blood test of the i-th subject, and it is equal to 1 if the subject is positive and 0 otherwise. Hence we have n independent random variables, from which we can define their sum and the estimated fraction $\hat{p}$, given respectively by $T = \sum_{i=1}^{n} T_i$ and $\hat{p} = (\sum_{i=1}^{n} T_i)/n = T/n$.

Knowing that $P(T_i = 1) = p$, we have $\mathrm{E}[T_i] = p$ and, by linearity of expectation, we obtain $\mathrm{E}[T] = np$ and $\mathrm{E}[\hat{p}] = np/n = p$.

We can now apply the Chernoff bound. The general version of Chernoff bounds for the upper and the lower tails when $0 < \epsilon < 1$ are given by

$$P(\hat{p} \geq (1+\epsilon)p) \leq e^{-\frac{\epsilon^2}{3}p}$$
$$P(\hat{p} \leq (1-\epsilon)p) \leq e^{-\frac{\epsilon^2}{2}p}$$

and we can easily combine them into one simple formula

$$P(|\hat{p} - p| \geq \epsilon p) \leq 2e^{-\frac{\epsilon^2}{3}p}$$

In order to solve the question, we must make explicit the dependence on n

$$\begin{aligned} P(|\hat{p} - p| \geq \epsilon p) &= P\left(\left|\frac{T}{n} - p\right| \geq \epsilon p\right) \\ &= P(|T - pn| \geq \epsilon pn) \\ &\leq 2e^{-\frac{\epsilon^2}{3}pn} \end{aligned}$$

Finally, we can derive the minimum value of n

$$\begin{aligned} 2e^{-\frac{\epsilon^2}{3}pn} &< \delta \\ e^{-\frac{\epsilon^2}{3}pn} &< \frac{\delta}{2} \\ e^{\frac{\epsilon^2}{3}pn} &> \frac{2}{\delta} \\ \frac{\epsilon^2}{3}pn &> \ln\frac{2}{\delta} \\ n &> \frac{3}{\epsilon^2 p}\ln\frac{2}{\delta} \end{aligned}$$

If we consider θ as the accuracy of our estimation and we take $\theta = \epsilon p$, the constraint $p > \theta$ holds and we obtain

$$n > \frac{3}{\epsilon\theta}\ln\frac{2}{\delta} = \frac{3p}{\theta^2}\ln\frac{2}{\delta}$$

Assignment 4

Section a

The information obtained by the 200 tested subjects is the interval $[160, 190]$ inside which the heights vary and their average value is equal to 180. To settle the argument between Professor Cooper and Professor Wolowitz we can set up a statistical hypothesis test. In particular, Professor Cooper supports the hypothesis that the heights are uniformly distributed in the interval, and it defines the null hypothesis H_0 of our test. On the contrary, Professor Wolowoitz states that the results are inconsistent with the former idea, and it defines the alternative hypothesis H_1. In other words,

$H_0 : X \sim U(160, 190)$
$H_1 : X \nsim U(160, 190)$

These hypotheses can be translated into a simpler form. If the distribution of the heights were truly uniform, the theoretical mean would be $\mu = (160 + 190)/2 = 175$. It means that we can rewrite the test in this new form

$H_0 : \mu = 175$
$H_1 : \mu \neq 175$

Section b

Before actually performing the test we must define the significance level α, used to reject or not reject the null hypothesis, and we set it to $\alpha = 0.05$. Now we can use a two-tailed Z-test

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

For our uniform distribution the value of the standard deviation is $\sigma = \sqrt{(190 - 160)^2/12}$, so we have

$$Z = \frac{180 - 175}{\sqrt{75}/\sqrt{200}} = 8.165$$

Finally, we can compare the result with the critical values for the chosen confidence level. In this case, we are using a two-tailed test, so we have to consider $\alpha/2 = 0.025$. The corresponding critical values are ± 1.96.

Since $8.165 > 1.96$, we reject the null hypothesis H_0.

Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08
-2.1	0.01786	0.01743	0.01700	0.01659	0.01618	0.01578	0.01539	0.01500	0.01463
-2.0	0.02275	0.02222	0.02169	0.02118	0.02068	0.02018	0.01970	0.01923	0.01876
-1.9	0.02872	0.02807	0.02743	0.02680	0.02619	0.02559	0.02500	0.02442	0.02385
-1.8	0.03593	0.03515	0.03438	0.03362	0.03288	0.03216	0.03144	0.03074	0.03005
-1.7	0.04457	0.04363	0.04272	0.04182	0.04093	0.04006	0.03920	0.03836	0.03754

Figure 1: Z-table

As an addendum to the hypothesis test, here are a few lines of code that reproduce the computation of the Z-test and empirically show how, taking 200 samples uniformly at random from $[160, 190]$ multiple times, we don't come up with a sample mean of 180.