Chapter 2 Solutions

1. A coffee shop buys roasted coffee from a supplier. In order to assess the quality of the supplied coffee, the manager of the shop conducts a tasting experiment where she selects a small portion of coffee beans from different batches and tastes the coffee from each portion. For each portion she gives a score in the scale $1, 2, \ldots, 10$ with 10 corresponding to coffee of the best taste and uses the results to assess the quality of the coffee.

   Identify the population, parameter, and statistic.

   Population: In this case we want to draw conclusions about all roasted coffee delivered from the supplier.

   Parameter: In this tasting experiment we give a score to the taste from each batch. The parameter could be the average score from all possible batches even those which we didn't taste. Because each tasting experiment is a score from 1 to 10, the parameter space is the interval $[1, 10]$.

   Statistic: The statistic that helps us estimate the value of the parameter is the sample average. The possible values of the statistic are in the range of 1 to 10. It is not easy to come up with the distribution of the test statistic. If we assume that the central limit theorem applies, then the distribution of the sample average is approximately the normal distribution.

2. Read the abstract of the article: "Dietary Intake of Marine n-3 Fatty Acids, Fish Intake, and the Risk of Coronary Disease among Men" by Ascherio and others published in *The New England Journal of Medicine* on April 13, 1995
   `http://www.nejm.org/doi/full/10.1056/NEJM199504133321501`

   Identify the population, parameter, sample, and statistic.

   Population: All men (according to the title).

   Parameter: There are two parameters of interest: (1) The probability of having a coronary event (death from coronary disease, nonfatal myocardial infarction, and coronary-artery bypass or angioplasty procedures). (2) The reduction, if any, of the risk of coronary disease caused by the intake of marine n-3 fatty acids.

Sample: The sample consists of the 44,895 health professionals, who completed the questionnaire.

Statistic: The number of participants in the study who experienced a coronary event.

3. Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathrm{N}(\mu, \sigma^2)$. Derive the sampling distribution of $\bar{X}$ given in Example 2.4

Using the result in Problem 7 in Exercise sheet 2, the moment generating function of a Normal distribution with mean μ and variance σ^2 is given by:

$$M_X(t) = e^{\mu t + \frac{\sigma^2 t^2}{2}}$$

Therefore the MGF of the average of n independent Normal distributions is:

$$M_{\bar{X}}(t) = \left(M_X\left(\frac{t}{n}\right)\right)^n = \left(e^{\mu\frac{t}{n} + \frac{\sigma^2\left(\frac{t}{n}\right)^2}{2}}\right)^n = e^{\mu t + \frac{\sigma^2 t^2}{2n}}$$

This is the MGF of a Normal distribution with mean μ and variance $\frac{\sigma^2}{n}$. Thus, the average of n independent Normal distributions with mean μ and variance σ^2 is also Normally distributed with mean μ and variance $\frac{\sigma^2}{n}$. To find the mean and variance, we write $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$ so

$$\begin{aligned}
\mathsf{E}\,\bar{X} &= \mathsf{E}\left(\frac{1}{n}\sum_{i=1}^n X_i\right) & \mathsf{Var}\,\bar{X} &= \mathsf{Var}\left(\frac{1}{n}\sum_{i=1}^n X_i\right) \\
&= \frac{1}{n}\,\mathsf{E}\left(\sum_{i=1}^n X_i\right) & &= \frac{1}{n^2}\,\mathsf{Var}\left(\sum_{i=1}^n X_i\right) \\
&= \frac{1}{n}\sum_{i=1}^n \mathsf{E}\,X_i & &= \frac{1}{n^2}\sum_{i=1}^n \mathsf{Var}\,X_i, \text{ by independence} \\
&= \frac{1}{n}\sum_{i=1}^n \mu & &= \frac{1}{n^2}\sum_{i=1}^n \sigma^2 \\
&= \frac{1}{n} n\mu = \mu & &= \frac{1}{n^2} n\sigma^2 = \sigma^2/n.
\end{aligned}$$

4. Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Bernoulli}(p)$.

(a) Derive the sampling distribution of $\bar{X}$.

(b) Derive the asymptotic distribution of $\bar{X}$ from the central limit theorem.

(c) Draw a graph of the exact and approximate CDFs when $n = 20$ and $p = 0.4$.

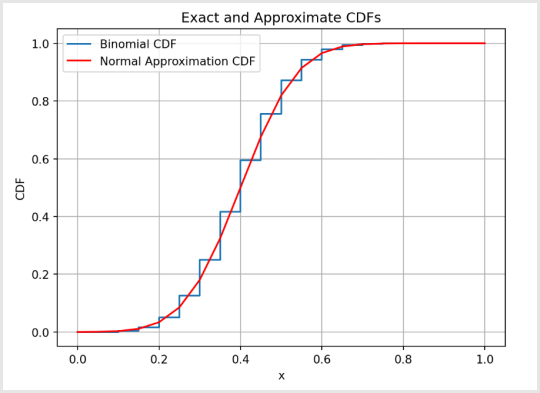
(a) Let $W = \sum X_i$. Then $W \sim \text{Bin}(n, p)$ and $\mathbb{P}(W = w) = \binom{n}{w} p^w (1-p)^{n-w}$.

Therefore, $\mathbb{P}(\bar{X} = x) = \mathbb{P}(W = nx) = \binom{n}{nx} p^{nx} (1-p)^{n-nx}$.

Thus, $\bar{X}$ is the sample proportion from n Bernoulli trials.

(b) The mean and variance of each X_i is $\mathsf{E}\, X_i = p$ and $\mathsf{Var}\, X_i = p(1-p)$. By the central limit theorem $\bar{X} \sim N\left(p, \frac{p(1-p)}{n}\right)$ approximately.

(c) The two CDFs can be seen in the figure below



with the red line denoting the approximate normal CDF and the blue line the exact binomial proportion. The exact CDF is a step function because the random variable is discrete.

Below is some Python code to produce the plot.

```
import numpy as np
import matplotlib.pyplot as plt
from scipy.stats import binom, norm
n = 20
p = 0.4
x = np.arange(0, n+1)
binom_cdf = binom.cdf(x, n, p)
norm_cdf = norm.cdf(x, loc=n*p, scale=np.sqrt(n*p*(1-p)))
plt.step(x, binom_cdf, label='Binomial CDF', where='post
')
plt.plot(x, norm_cdf, label='Normal Approximation CDF', color='red
')
plt.xlabel('x')
plt.ylabel('CDF')
plt.title('Exact and Approximate CDFs')
plt.legend()
plt.grid()
plt.show()
```