Please provide complete and well-written solutions to the following exercises.

Due September 7, 12PM noon PST, to be uploaded as a single PDF document to Gradescope.

Homework 2

Exercise 1. Let $n \geq 2$ be an integer. Let $X_1, \ldots, X_n$ be a random sample of size n (that is, $X_1, \ldots, X_n$ are i.i.d. random variables). Assume that $\mu := \mathbf{E}X_1 \in \mathbf{R}$ and $\sigma := \sqrt{\mathrm{var}(X_1)} < \infty$. Let $\overline{X}$ be the sample mean and let S be the sample standard deviation of the random sample. Show the following

- $\mathrm{Var}(\overline{X}) = \sigma^2/n$.
- $\mathbf{E}S^2 = \sigma^2$.

Exercise 2 (Optional)**.** Let $X_1, \ldots, X_n$ be i.i.d. standard Gaussian random variables (i.e. Gaussian random variables with mean zero and variance one). Show that

$$X_1^2 + \cdots + X_n^2$$

has a chi-squared distribution with n degrees of freedom.

(Hint: Let $B(0,r) := \{(x_1, \ldots, x_n) \in \mathbf{R}^n \colon x_1^2 + \cdots + x_n^2 \leq r^2\}$. Using hyperspherical coordinates, write

$$\begin{aligned}\mathbf{P}(X_1^2 + \cdots + X_n^2 \leq t) &= (2\pi)^{-n/2} \int_{B(0,\sqrt{t})} e^{-(x_1^2+\cdots+x_n^2)/2} dx_1 \cdots dx_n \\ &= (2\pi)^{-n/2} \int_{r=0}^{r=\sqrt{t}} \int_{\partial B(0,1)} r^{n-1} e^{-r^2/2} d\sigma dr,\end{aligned}$$

where $d\sigma$ denotes integration on the boundary of the unit ball $\partial B(0,1)$. To find the latter quantity, let $t = \infty$ to note that

$$1 = (2\pi)^{-n/2} \int_{r=0}^{r=\infty} r^{n-1} e^{-r^2/2} dr \cdot \int_{\partial B(0,1)} d\sigma,$$

so that

$$\int_{\partial B(0,1)} d\sigma = \frac{(2\pi)^{n/2}}{\int_{r=0}^{r=\infty} r^{n-1} e^{-r^2/2} dr},$$

and then change variables to obtain the Gamma function on the right side denominator.)

Exercise 3. Let X be a chi squared random variables with p degrees of freedom. Let Y be a chi squared random variable with q degrees of freedom. Assume that X and Y are independent. Show that $(X/p)/(Y/q)$ has the following density, known as **Snedecor's f-distribution** with p and q degrees of freedom

$$f_{(X/p)/(Y/q)}(t) := \frac{t^{(p/2)-1}(p/q)^{p/2}\Gamma((p+q)/2)}{\Gamma(p/2)\Gamma(q/2)} \Big(1 + t(p/q)\Big)^{-(p+q)/2}, \qquad \forall\, t > 0.$$

Exercise 4 (Order Statistics)**.** Let $X\colon \Omega \to \mathbf{R}$ be a random variable. Let $X_1, \ldots, X_n$ be a random sample of size n from X. Define $X_{(1)} := \min_{1\le i\le n} X_i$, and for any $2 \le i \le n$, inductively define

$$X_{(i)} := \min\Big\{\{X_1, \ldots, X_n\} \smallsetminus \{X_{(1)}, \ldots, X_{(i-1)}\}\Big\},$$

so that

$$X_{(1)} \le X_{(2)} \le \cdots \le X_{(n)} = \max_{1\le i\le n} X_i.$$

The random variables $X_{(1)}, \ldots, X_{(n)}$ are called the **order statistics** of $X_1, \ldots, X_n$.

- Suppose X is a discrete random variable and we can order the values that X takes as $x_1 < x_2 < \cdots$. For any $i \ge 1$, define $p_i := \mathbf{P}(X \le x_i)$. Show that, for any $1 \le i, j \le n$,

  $$\mathbf{P}(X_{(j)} \le x_i) = \sum_{k=j}^{n} \binom{n}{k} p_i^k (1-p_i)^{n-k}.$$

  (Hint: Let Y be the number of indices $1 \le j \le n$ such that $X_j \le x_i$. Then Y is a binomial random variable with parameters n and p_i.)

  You don't have to show it, but if X is a continuous random variable with density f_X and cumulative distribution function F_X, then for any $1 \le j \le n$, $F_{X_{(j)}}$ has density

  $$f_{X_{(j)}}(x) := \frac{n!}{(j-1)!(n-j)!} f_X(x)(F_X(x))^{j-1}(1-F_X(x))^{n-j}, \qquad \forall\, x \in \mathbf{R}.$$

  (This follows by differentiating the above identity for the cumulative distribution function, i.e. by differentiating $\mathbf{P}(X_{(j)} \le x) = \sum_{k=j}^{n} \binom{n}{k} F_X(x)^k (1-F_X(x))^{n-k}$, where $F_X(x) := \mathbf{P}(X \le x)$ for any $x \in \mathbf{R}$.)
- Let X be a random variable uniformly distributed in $[0, 1]$. For any $1 \le j \le n$, show that $X_{(j)}$ is a beta distributed random variable with parameters j and $n - j + 1$. Conclude that (as you might anticipate)

  $$\mathbf{E}X_{(j)} = \frac{j}{n+1}.$$

- Let $a, b \in \mathbf{R}$ with $a < b$. Let U be the number of indices $1 \le j \le n$ such that $X_j \le a$. Let V be the number of indices $1 \le j \le n$ such that $a < X_j \le b$. Show that the vector $(U, V, n - U - V)$ is a multinomial random variable, so that for any nonnegative integers u, v with $u + v \le n$, we have

  $$\begin{aligned}&\mathbf{P}(U = u, V = v, n - U - V = n - u - v)\\ &\quad = \frac{n!}{u!v!(n-u-v)!} F_X(a)^u (F_X(b) - F_X(a))^v (1 - F_X(b))^{n-u-v}.\end{aligned}$$

  Consequently, for any $1 \le i, j \le n$,

  $$\mathbf{P}(X_{(i)} \le a, X_{(j)} \le b) = \mathbf{P}(U \ge i, U + V \ge j) = \sum_{k=i}^{j-1} \sum_{m=j-k}^{n-k} \mathbf{P}(U = k, V = m) + \mathbf{P}(U \ge j).$$

  So, it is possible to write an explicit formula for the joint distribution of $X_{(i)}$ and $X_{(j)}$ (but you don't have to write it yourself).

Remark 1. We might occasionally do some computer-based exercises. You can use whatever program you want to do these exercises. Here are some links for downloading such software:

Matlab software download

R software download

Exercise 5. Using Matlab (or any other mathematical system on a computer), verify that its random number generator agrees with the law of large numbers and central limit theorem. For example, average 10^7 samples from the uniform distribution on $[0, 1]$ and check how close the sample average is to $1/2$. Then, make a histogram of 10^7 samples from the uniform distribution on $[0, 1]$ and check how close the histogram is to a Gaussian.

Exercise 6 (Sunspot Data)**.** This exercise deals with sunspot data from the following files (the same data appears in different formats)

txt file csv (excel) file

These files are taken from http://www.sidc.be/silso/datafiles#total

To work with this data, e.g. in Matlab you can use the command

```
x=importdata('SN_d_tot_V2.0.txt')
```

to import the .txt file.

The format of the data is as follows.

- Columns 1-3: Gregorian calendar date (Year, Month, then Day)
- Column 4: Date in fraction of year
- Column 5: Daily total number of sunspots observed on the sun. A value of -1 indicates that no number is available for that day (missing value).
- Column 6: Daily standard deviation of the input sunspot numbers from individual stations.
- Column 7: Number of observations used to compute the daily value.
- Column 8: Definitive/provisional indicator. A blank indicates that the value is definitive. A '*' symbol indicates that the value is still provisional and is subject to a possible revision (Usually the last 3 to 6 months)

For this data set, do the following:

• Plot the number of sunspots U_t versus time t. Label and scale the axes appropriately. On this same plot, also plot some moving averages of U_t. For example, for a given time t, plot the average of the twenty previous days' sunspot counts, versus time t.

• Find the sample average and sample standard deviation of U_t, averaging over all t given in the data.

• Do you notice any periodic behavior in U_t versus t?