MATH 408 Homework 2

Qilin Ye

September 8, 2021

Problem 1

Le $n \geqslant 2$ be an integer. Let $X_1, ..., X_n$ be a random sample of size n (that is, $X_1, ..., X_n$ are i.i.d. random variables). Assume that $\mu := \mathbb{E}X_1 \in \mathbb{R}$ 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 that $\mathrm{var}(\overline{X}) = \sigma^2/n$ and $\mathbb{E}S^2 = \sigma^2$.

Proof. By definition we have $\overline{X} = (X_1 + ... + X_n)/n$ and $S^2 = \sum_{i=1}^{n}(X_i - \overline{X})^2/(n-1)$. Then,

$$\mathrm{var}(\overline{X}) = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{var}(X_i) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}.$$

To show that $\mathbb{E}S^2 = \sigma^2$,

$$\begin{aligned}
\mathbb{E}S^2 &= \frac{1}{n-1}\mathbb{E}\sum_{i=1}^{n}(X_i - \overline{X})^2 \\
&= \frac{1}{n-1}\mathbb{E}\sum_{i=1}^{n}(X_i^2 - 2X_i\overline{X} + \overline{X}^2) \\
&= \frac{1}{n-1}\left[n\mathbb{E}X_1^2 - 2\mathbb{E}\left(\overline{X}\sum_{i=1}^{n}X_i\right) + n\mathbb{E}\overline{X}^2\right] \\
&= \frac{1}{n-1}\left[n\mathbb{E}X_1^2 - 2\mathbb{E}n\overline{X}^2 + n\mathbb{E}\overline{X}^2\right] \\
&= \frac{1}{n-1}\left[n\mathbb{E}X_1^2 - n\mathbb{E}\overline{X}^2\right]. \qquad (1)
\end{aligned}$$

Since

$$\mathrm{var}(X_1) = \mathbb{E}(X - \mathbb{E}X)^2 = \mathbb{E}X^2 - (\mathbb{E}X)^2,$$

we obtain

$$\mathbb{E}X_1^2 = \sigma^2 + \mu^2 \qquad \text{and likewise} \qquad \mathbb{E}\overline{X} = \mathrm{var}(\overline{X}) + \mu^2 = \frac{\sigma^2}{n} + \mu^2.$$

Substituting these values back into (1), we obtain $\mathbb{E}S^2 = (n\sigma^2 + n\mu^2 - \sigma^2 - n\mu^2)/(n-1) = \sigma^2$, as claimed. □

Problem 2

Let $X_1, ..., X_n$ be i.i.d. standard Gaussian random variables. Show that

$$X_1^2 + ... + X_n^2$$

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

Proof. We follow the hint and get

$$\begin{aligned}\int_0^\infty e^{-r^2/2} r^{n-1}\,\mathrm{d}r &= \int_0^\infty e^{-u}(2u)^{(n-1)/2}(2u)^{-1/2}\,\mathrm{d}u \\ &= \int_0^\infty e^{-u}(2u)^{n/2-1}\,\mathrm{d}u = 2^{n/2-1}\Gamma(n/2).\end{aligned}$$

Therefore,

$$\int_{\partial B(0,1)} \mathrm{d}\sigma = \frac{(2\pi)^{n/2}}{2^{n/2-1}\Gamma(n/2)}.$$

Clearly, if $t = 0$, the probability $\mathbb{P}(\sum_{i=1}^n X_i^2 \leqslant 0) = 0$, which coincides with $\int_{-\infty}^0 f_X(x)\,\mathrm{d}x$ where f_X denotes the pdf of chi-squared distribution with n degrees of freedom. Furthermore, differentiating $\mathbb{P}$ with respect to t gives

$$\begin{aligned}&(2\pi)^{-n/2}\int_{\partial B(0,1)} \mathrm{d}\sigma \cdot \frac{\mathrm{d}}{\mathrm{d}t}\left[\int_0^{\sqrt{t}} r^{n-1}e^{-r^2/2}\,\mathrm{d}r\right] \\ =\ &(2\pi)^{-n/2}\cdot\frac{(2\pi)^{n/2}}{2^{n/2-1}\Gamma(n/2)}\cdot\frac{\mathrm{d}}{\mathrm{d}t}\left[\int_0^t u^{(n-1)/2}e^{-u/2}\cdot(2^{-1}u^{-1/2})\,\mathrm{d}u\right] \\ =\ &\frac{t^{n/2-1}e^{-t/2}}{2^{n/2}\Gamma(n/2)},\end{aligned}$$

so indeed $\mathbb{P}(X \leqslant t)$ coincides with $\int_{-\infty}^t f_X(x)\,\mathrm{d}x$, i.e., $X_1^2 + \ldots + X_n^2$ has a chi-squared distribution with n degrees of freedom. □

Problem 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 the **Snedecor's f-distribution** with p and q degrees of freedom

$$f_{(X/p)/(Y/q)}(t) := \frac{t^{p/2-1}(p/q)^{p/2}\Gamma((p+q)/2)}{\Gamma(p/2)\Gamma(q/2)}\left(1+t(p/q)\right)^{-(p+q)/2} \qquad \text{for all } t > 0.$$

Proof. Let X and Y be as stated. By definition, we have the PDFs

$$f_X(x) = \frac{x^{p/2-1}e^{-x/2}}{2^{p/2}\Gamma(p/2)} \qquad \text{and} \qquad f_Y(y) = \frac{y^{q/2-1}e^{-y/2}}{2^{q/2}\Gamma(q/2)}. \tag{1}$$

By independence, we also have the JPDF

$$f_{X,Y}(x,y) = f_X(x)f_Y(y) = \frac{x^{p/2-1}y^{q/2-1}e^{-(x+y)/2}}{2^{(p+q)/2}\Gamma(p/2)\Gamma(q/2)}. \tag{2}$$

Note that $(X/p)/(Y/q) = (X/Y)(q/p)$ and q/p is a constant, so the important part is to compute X/Y. We begin

by computing its CDF. Let $t > 0$. Then

$$\begin{aligned} F_{X/Y}(t) &= P(X/Y \leqslant t) = P(X \leqslant tY) \\ &= \int_0^\infty \int_0^{yt} f_{X,Y}(x,y)\,\mathrm{d}x\,\mathrm{d}y \\ &= \frac{1}{2^{(p+q)/2}\Gamma(p/2)\Gamma(q/2)} \int_0^\infty \left[\int_0^{yt} x^{p/2-1}e^{-x/2}\,\mathrm{d}x\right] y^{q/2-1}e^{-y/2}\,\mathrm{d}y. \end{aligned} \tag{3}$$

We can recover the PDF of X/Y by differentiating (3) with respect to t:

$$\begin{aligned} f_{X/Y}(t) &= \frac{\mathrm{d}}{\mathrm{d}t}(3) = \frac{1}{2^{(p+q)/2}\Gamma(p/2)\Gamma(q/2)} \int_0^\infty \left[(yt)^{p/2-1}e^{-yt/2}\cdot y\right] y^{q/2-1}e^{-y/2}\,\mathrm{d}y \\ &= \frac{t^{p/2-1}}{2^{(p+q)/2}\Gamma(p/2)\Gamma(q/2)} \int_0^\infty y^{(p+q)/2-1}\cdot e^{-y(t+1)/2}\,\mathrm{d}y \\ (\Delta) &= \frac{t^{p/2-1}}{2^{(p+q)/2}\Gamma(p/2)\Gamma(q/2)} \cdot \Gamma(p/2+q/2)\left(\frac{2}{t+1}\right)^{(p+q)/2} \\ &= \frac{\Gamma((p+q)/2)}{\Gamma(p/2)\Gamma(q/2)} \cdot \frac{t^{p/2-1}}{(t+1)^{(p+q)/2}}, \end{aligned} \tag{4}$$

where (Δ) is because

$$g(y) := \frac{y^{(p+q)/2-1}\cdot e^{-y(t+1)/2}}{(2/(t+1))^{(p+q)/2}\cdot\Gamma((p+q)/2}$$

is the PDF of a $\left(\frac{p+q}{2}, \frac{2}{t+1}\right)$-distributed Gamma random variable and thus has integral 1. Finally,

$$\begin{aligned} f_{(X/p)(Y/q)}(t) &= f_{(X/Y)(q/p)}(t) = \frac{p}{q}\cdot f_{X/Y}(t(p/q)) \\ [\text{by (4)}] &= \frac{p}{q}\cdot\frac{\Gamma((p+q)/2)}{\Gamma(p/2)\Gamma(q/2)}\cdot\frac{(t(p/q))^{p/2-1}}{(1+t(p/q))^{(p+q)/2}} \\ &= \frac{t^{p/2-1}(p/q)^{p/2}}{(1+t(p/q))^{(p+q)/2}}\cdot\frac{\Gamma((p+q)/2)}{\Gamma(p/2)\Gamma(q/2)}, \end{aligned}$$

as claimed. □

Problem 4

Let $X : \Omega \to \mathbb{R}$ be a random variable. Let $X_1, ..., X_n$ be a random sample of size n from X. Define $X_{(1)} := \min_{1\leqslant i\leqslant n} X_i$ and for any $2 \leqslant i \leqslant n$, inductively define

$$X_{(i)} := \min\left\{\{X_i, ..., X_n\} - \{X_{(1)}, ..., X_{(i-1)}\}\right\}$$

so that

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

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

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

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

In particular, if X is a continuous random variable with density f_X and CDF F_X, then for $1 \leqslant j \leqslant n$, $F_{X_{(j)}}$ has density

$$f_{X_{(j)}} := \frac{n!}{(j-1)!(n-j)!} f_X(x)(F_X(x))^{j-1}(1-F_X(x))^{n-j} \qquad \text{for all } x \in \mathbb{R}.$$

(2) Let X be a random variable uniformly distributed in $[0,1]$. For any $1 \leqslant j < n$, show that $X_{(j)}$ is a beta distributed random variable with parameters j and $n-j$. Conclude that $\mathbb{E}X_{(j)} = j/(n+1)$.

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

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

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

$$\begin{aligned}\mathbb{P}(X_{(i)} \leqslant a, X_{(j)} \leqslant b) &= \mathbb{P}(U \geqslant i, U+V \geqslant j)\\ &= \sum_{k=i}^{j-1} \sum_{m=j-k}^{n-k} \mathbb{P}(U=k, V=m) + \mathbb{P}(U \geqslant j).\end{aligned}$$

Proof. (1) Fix i. In order that $X_{(j)} \leqslant x_i$, among $X_1, \ldots, X_n$, we need the event $X_k \leqslant x_i$ $(k = 1, 2, \ldots, n)$ to happen at least j times. Considering the event $\{X_i \leqslant x_i\}$ as a success and $\{X_i > x_i\}$ as a failure, we obtain a binomial distribution with parameters n and p_i. Hence

$$\mathbb{P}(X_{(j)} \leqslant x_i) = \sum_{k \geqslant j} \mathbb{P}(k \text{ “successes”}) = \sum_{k=j}^{n} \binom{n}{k} p_i^k (1-p_i)^{n-k}.$$

(2) If X is uniformly distributed in $[0,1]$ we have $f_X(x) \equiv 1$ on $[0,1]$ and $F(x) \equiv x$ on $[0,1]$. Then by (1)'s remark

$$f_{X_{(j)}} = \frac{n!}{(j-1)!(n-j)!} x^{j-1}(1-x)^{n-j}. \tag{1}$$

This is indeed consistent with a $(j, n-j+1)$-distributed beta distribution. (Since the support of (1) is $[0,1]$, the factorial coefficient must be the scaling factor $B(j, n-j+1)$; of course, we could also evaluate the integral $\int_0^1 x^{j-1}(1-x)^{n-j+1}\,\mathrm{d}x$ to verity it.) Therefore,

$$\begin{aligned}\mathbb{E}X_{(j)} &= \int_0^1 x f_{X_{(j)}}\,\mathrm{d}x = \frac{n!}{(j-1)!(n-j)!} \int_0^1 x^j (1-x)^{n-j}\,\mathrm{d}x\\ &= \frac{n!}{(j-1)!(n-j)!} \cdot \frac{j!(n-j)!}{(n+1)!} = \frac{j}{n+1}, \text{ as claimed.}\end{aligned}$$

(3) We can consider the multinomial in which each trial is represented by the value of X_n, with three possible outcomes:

$$(1) X_n \leqslant a, \qquad (2) a < X_n \leqslant b, \qquad \text{and} \qquad (3) X_n > b.$$

It is clear that each trial is independent since $X_1, ..., X_n$ are i.i.d. (so that the outcome of one trial imposes no effect on that of any other trial). Therefore, in plain language,

$$\begin{aligned}&\mathbb{P}(U = u, V = v, n - U - V = n - u - v)\\ &= \mathbb{P}(\text{in } n \text{ trials, get (1) } u \text{ times, (2) } v \text{ times, and (3) } n - u - v \text{ times})\end{aligned}$$

The probability of getting (1) is $\mathbb{P}(X_1 \leqslant a) = F_X(a)$; similarly, that of (2) and (3) are $(F_X(b) - F_X(a))$ and $(1 - F_X(b))$, respectively. The claim then follows by raising each of these terms to the appropriate power and multiplying everything by the corresponding multinomial coefficient. □

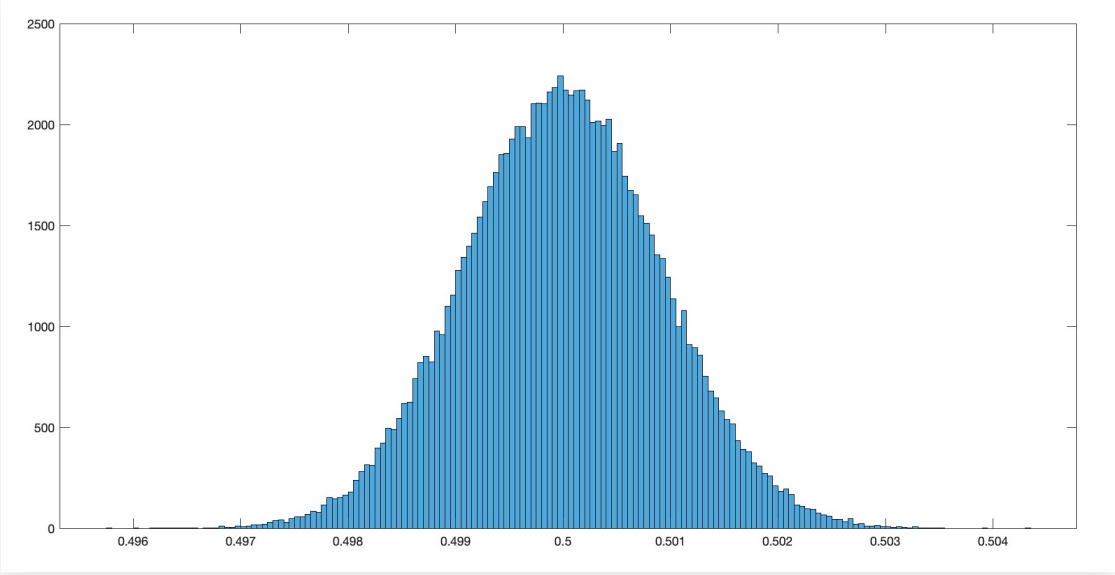
Problem 5

Using Matlab, verify that its random number generator agrees with the LLN. 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 unifom distribution on $[0, 1]$ and check how close the histogram is to a Gaussian.

Solution. I used 10^6 instead as 10^7 takes much longer to run:

```
data = zeros(1,100000);
for i=1:100000
    temp_data = zeros(1,100000);
    for j=1:100000
        temp_data(j) = rand();
    end
    data(i) = mean(temp_data);
end

histogram(data);
```



Problem 6

Here is a file containing the number of sunspots observed in each day, starting from 1800.
Plot the number of sunspots U_t versus time t. Label and sclae the axes appropriate. On the same plot, also plot some moving average of U_t. For example, for a given 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?

Solution. The sample average is 78.7545 and the sample standard deviation is 77.4328. The graphs indeed seem to exhibit an oscillating pattern.

```
x = importdata('SN_d_tot_V2.0.txt');
hold on
time = x(:,4);
sunspot = x(:,5);
plot(time, sunspot);

sunspot_avg = mean(sunspot);
sunspot_stdev = std(sunspot);

sunspot_moving = sunspot;

for i = 1:size(sunspot)
    sunspot_moving(i) = mean(sunspot(max(1,i-19):i));
end
plot(time, sunspot_moving);
```