Probability Theory for EOR 2021/2022

—

Assignment 2

Deadline: Friday December 24th, 19h00.
Hand-in (handwritten okay, but prefer the typed) PDF report via Nestor.

Instructions

- **This assignment is made available much earlier, so you can hand in a bit early rather than till the night before Christmas**.
- **After week 4, Ex1 doable, after week 5, Ex2 (a)-(c) doable.**
- Explain your answers. If you just answer $P(A) = 0.0030$ you get 0 points, even if this is the correct answer.
- It is better to define and use clear notation. E.g., prefer not to write:

  $$P(\text{we have two pairs in the first four cards}) = \ldots$$

  Instead, define the event A: we have two pairs in the first four cards, and then write $P(A) = \ldots$
- Make sure that it is clear what your final answer is.
- Use *first* a formula and *then* fill in the numbers. So suppose you have defined N_A and N_T and have calculated that $N_A = 10$ and $N_T = 20$. The answer could look something like: "Since the outcomes are equally likely, by the naive definition of probability, $P(A) = N_A/N_T = 10/20 = 1/2$." Note that the **underlined part is a crucial piece** of information. **Because it it important to motivate your steps.**
- During the exam, it is okay if you want to use **definitions/theorems/propositions** from the textbook and you do not need to invoke the exact numbers of the **definitions/theorems/propositions** but only need to rephrase the contents via your own words, same rule applies here.

1. X, Y are two independent discrete random varaibles.

   (a) Prove $\text{Var}(X+Y) = \text{Var}(X) + \text{Var}(Y)$.

Chapter 3 Independent r.v.'s and Chapter 4: Discrete r.v.'s and Expectation

Denote $X_0 = X - E(X), Y_0 = Y - E(Y)$, and $\text{Var}(X+Y) = \text{Var}(X_0+Y_0)$ since we shift by a constant the varaince would remain the same.

$$\text{Var}(X_0+Y_0) = E(X_0+Y_0)^2 - (E(X_0+Y_0))^2 = E(X_0+Y_0)^2.$$

By LOTUS, we know

$$\begin{aligned}
E(X_0+Y_0)^2 &= \sum_{x\in\text{Supp}(X_0),y\in\text{Supp}(Y_0)} (x+y)^2 P(X_0=x, Y_0=y) \\
&= \sum_{x\in\text{Supp}(X_0),y\in\text{Supp}(Y_0)} (x^2+y^2+2xy) P(X_0=x, Y_0=y) \\
&= \sum_{x\in\text{Supp}(X_0),y\in\text{Supp}(Y_0)} (x^2) P(X_0=x, Y_0=y) \\
&+ \sum_{x\in\text{Supp}(X_0),y\in\text{Supp}(Y_0)} (y^2) P(X_0=x, Y_0=y) \\
&+ \sum_{x\in\text{Supp}(X_0),y\in\text{Supp}(Y_0)} (2xy) P(X_0=x, Y_0=y) \qquad (1)
\end{aligned}$$

Note that by axioms of probability or LOTP and LOTUS:

$$\begin{aligned}
&\sum_{x\in\text{Supp}(X_0),y\in\text{Supp}(Y_0)} (x^2) P(X_0=x, Y_0=y) \\
=&\sum_{x\in\text{Supp}(X_0)} x^2 \sum_{y\in\text{Supp}(Y_0)} P(X_0=x, Y_0=y) \\
=&\sum_{x\in\text{Supp}(X_0)} x^2 P(X_0=x) \\
=&E(X_0)^2
\end{aligned}$$

So we know

$$\sum_{x\in\text{Supp}(X_0),y\in\text{Supp}(Y_0)} (x^2) P(X_0=x, Y_0=y) = \text{Var}(X_0) \qquad (2)$$

since $\text{Var}(X_0) = E(X_0)^2 - (E(X_0))^2$ for mean-zero X_0 ($E(X_0) = 0$). Similarly, we know

$$\sum_{x\in\text{Supp}(X_0),y\in\text{Supp}(Y_0)} (y^2) P(X_0=x, Y_0=y) = \text{Var}(Y_0) \qquad (3)$$

By independence,

$$\begin{aligned}
&\sum_{x\in\text{Supp}(X_0),y\in\text{Supp}(Y_0)} (2xy) P(X_0=x, Y_0=y) = \sum_{x\in\text{Supp}(X_0)} \sum_{y\in\text{Supp}(Y_0)} (2xy) P(X_0=x) P(Y_0=y) \\
=&2 \sum_{x\in\text{Supp}(X_0)} x P(X_0=x) \sum_{y\in\text{Supp}(Y_0)} y P(Y_0=y) = 2E(X_0)E(Y_0) = 0 \qquad (4)
\end{aligned}$$

Equations (1)-(4) togenter imply that $\text{Var}(X_0+Y_0) = \text{Var}(X_0) + \text{Var}(Y_0)$.

   (b) Provide a counter-example such that Z_1, Z_2 are two random varaibles and $\text{Var}(Z_1 + Z_2) > \text{Var}(Z_1) + \text{Var}(Z_2)$. Compare with the result you proved in the subquestion

1.(a), provide one intuitive explaination for the difference.

Chapter 4: Expectation

Let $Z_1 = Z_2 = X$ and assume that $\mathrm{Var}(X) > 0$. Then

$$\mathrm{Var}(Z_1 + Z_2) = \mathrm{Var}(2X) = 4\mathrm{Var}(X) > 2\mathrm{Var}(X) = \mathrm{Var}(Z_1) + \mathrm{Var}(Z_2).$$

The intuiation is when we sum up two independent r.v.'s, since they are independent, when one is large the other one could be small, and thus after the summation they may not be far away from the mean. If we add two highly dependent r.v.'s then they can be simutaneously large and small and thus increase the spread.

2. A group of $n \geq 4$ persons independently draw numbers from $\mathrm{Unif}(0,1)$. The person who has the highest number wins our textbook free of charge. Let X_i be the number drawn by the ith individual, so that $X_i \sim_{i.i.d.} \mathrm{Unif}(0,1)$.

   (a) What is the probability of having strictly more than one winner?

   Chapter 5: continuous random variables.

   The probability of having strictly more than one winner is zero (**use words to motivate is also okay, here I only show solutions motivated by formulas** Methods are not unique, using indicator functions is also okay.):

   $$\mathbb{P}\left(\left\{\bigcup_{i=1}^{n}\left\{\max_{l\neq i}\{X_l\} = \max_{1\leq j\leq n}\{X_j\}\right\}\right\}^c\right) \leq \mathbb{P}\left(\bigcup_{i\neq j}\{X_i = X_j\}\right) \leq \sum_{i\neq j}\mathbb{P}\left(X_i = X_j\right) = 0$$

   where the last equation is due to the fact that $X_i - X_j$ is one continuous r.v. and thus $\mathbb{P}\left(\{X_i = X_j\}\right) = \mathbb{P}\left(\{X_i - X_j = 0\}\right) = 0$ (the probability of one continuous r.v. equal to one fixed constant is zero).

   (b) What is the probability of the jth person having the largest number among the first j individuals?

Chapter 5: continuous random variables

By symmetry of continuous random variables we know

$$\mathbb{P}\left(X_{a_1} < X_{a_2} < \cdots < X_{a_j}\right) = 1/j!$$

for arbitrary permutation $(a_1, \cdots, a_j)$ of $(1, \cdots, j)$. Now among all permutations, there are $(j-1)!$ permutations of the format $(a_1, \cdots, a_{j-1}, j)$: therefore,

$$\mathbb{P}\left(\bigcup_{(a_1,\cdots,a_{j-1})} \left\{X_{a_1} < X_{a_2} < \cdots < X_{a_{j-1}} < X_j\right\}\right) = 1/j$$

(c) Calculate the expectations: $\mathbb{E}\left(\log\left(\frac{X_1+X_2}{2-(X_{n-1}+X_n)}\right)\right)$.

Chapter 4: Expectation + Chapter 5: Continuous random variables.

Note that $X_1 + X_2$ follows the same distribution as $1 - X_{n-1} + 1 - X_n$ (they have the same PDF). Therefore, by the LOTUS,

$$\mathbb{E}\left(\log\left(\frac{X_1+X_2}{2-(X_{n-1}+X_n)}\right)\right) = \mathbb{E}\left(\log\left(X_1+X_2\right)\right) - \mathbb{E}\left(\log\left(2-(X_{n-1}+X_n)\right)\right)$$
$$= \int_{-\infty}^{+\infty} \log(x)g(x)dx - \int_{-\infty}^{+\infty} \log(x)g(x)dx = 0$$

(d) Derive the moment generating function (MGF) of $Y = a + bX_1$.

Chapter 6: MGF.

We first derive the MGF of X_1 (for $t \neq 0$):

$$\mathbb{E}[e^{tX_1}] = \int_0^1 e^{tx}dx = (e^t - 1)/t$$

for $t = 0$, we know $\mathbb{E}[e^{tX_1}] = \mathbb{E}[1] = 1$ Therefore the MGF is well defined. Next, for $t \neq 0$

$$M_{a+bX_1}(t) = e^{at}M_{X_1}(bt) = e^{at}\left(e^{bt} - 1\right)/t$$

for $t = 0$, $M_{a+bX_1}(t) = 1$.

(e) Denote $X^* = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(X_i - 1/2)$, what is the variance and the MGF of X^*? If n is very large, what distribution would serve as a proper approximation for X^*? For the second question, make use of the MGF to motivate your choice.

Hint: When n is large, in your calculations you can approximate $\sqrt{n}\left(e^{t/(2\sqrt{n})} - e^{-t/(2\sqrt{n})}\right)$ with $t + t^3/(24n)$ and use the fact that $e^x = \lim_{n\to\infty}(1 + x/n)^n$.

Chapter 4: Variance and Chapter 6: MGF.

The variance (by independence) is $Var(X^*) = \sum_{i=1}^n Var(X_i)/n = 1/12$.

For $t \neq 0$,

$$
\begin{aligned}
M_{X^*}(t) = M_{\left(\sum_{i=1}^n X_i\right)-\sqrt{n}/2}(t/\sqrt{n}) &= e^{-\sqrt{n}t/2}\prod_{i=1}^n M_{X_i}(t/\sqrt{n}) \\
=& e^{-\sqrt{n}t/2}\left(\sqrt{n}\left(e^{t/\sqrt{n}} - 1\right)/t\right)^n = \left(e^{-t/(2\sqrt{n})}\sqrt{n}\left(e^{t/\sqrt{n}} - 1\right)/t\right)^n \\
=& \left(\sqrt{n}\left(e^{t/(2\sqrt{n})} - e^{-t/(2\sqrt{n})}\right)/t\right)^n =_{large\ n} \left((t + t^3/24/n)/t\right)^n \\
\to_{n\to\infty}& e^{t^2/24}
\end{aligned}
$$

When $t = 0$, $M_{X^*}(t) = 1$. The limiting distribution should be $N(0, 1/12)$ by the format of our derived limiting MGF, since the normal distribution has the MGF $e^{\mu t+\sigma^2 t^2/2}$.

For nearly any problem you'll encounter in business, consultancy, you'll need the computer to do the computations and simulations. The earlier you become comfortable with using the computer, the better.

Here we provide a short discussion in R. Later on in other courses, e.g., *Probability distribution* you may need to learn to read these as well. Both R and Python are used widely! We also provide Python code (open with colab https://colab.research.google.com/) for this assignment.

No need to hand in any codes, they are not graded and only here to help you solve the aforementioned exercises.

R help!

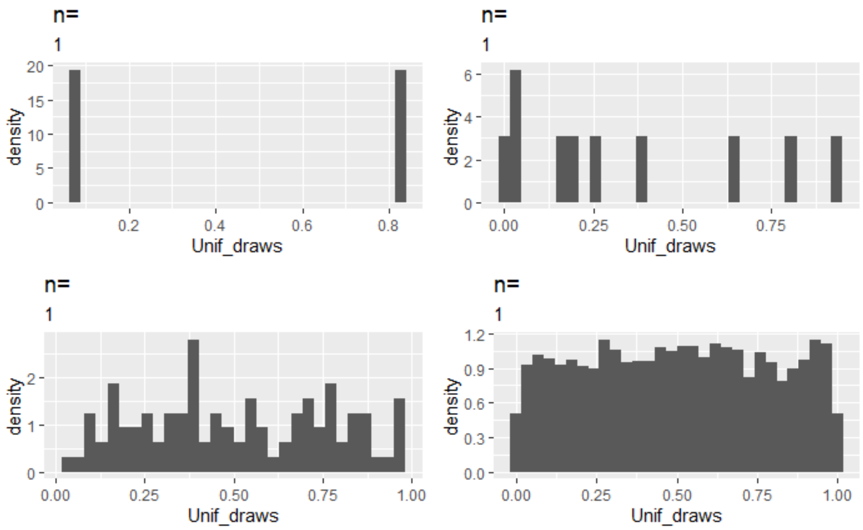
For continuouse random varaible, we can redraw many of its realised values, and then draw a histogram. Histogram can be regarded as a sample-version of the PDF (actually, it is indeed one estimator for the PDF). You can see the more data you use to draw the histogram, the histogram is smoother and closer to its PDF.

We may use the following codes to draw $S = 2, 10, 100, 5000$ realised values from Unif(0,1) and then draw histogram:

```
set.seed(12)
n=1
sigma2=1/12
par(mfrow=c(2,2))
require(ggplot2)
loop.vector <- c(2,10,100,5000)
plot_list = list()
j=0
for(i in loop.vector){
  S=i
  j=j+1
  Unif_draws= runif(n*S, min = 0, max = 1)

  plot <- ggplot(data.frame(Unif_draws), aes(Unif_draws)) +
  geom_histogram(aes(y=..density..)) +
  ggtitle("n=",n)
  plot_list[[j]] = plot
}
#library(gridExtra)
require(gridExtra)
grid.arrange(plot_list[[1]], plot_list[[2]],plot_list[[3]],
     plot_list[[4]], nrow = 2)
```

The output is the following figure



and indeed the larger the S the closer the histogram to the PDF of Unif(0,1).

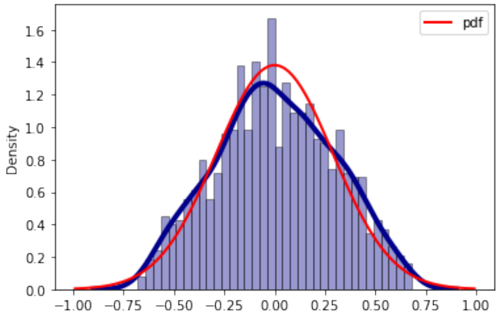


Figure 1: n=**2**,S=1000

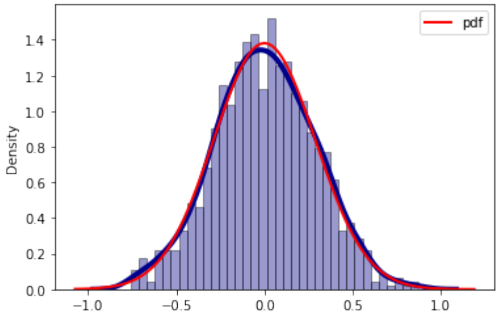


Figure 2: n=**20**,S=1000

From the comparison between the histograms of X_n^* (blue) and normal density (red), we can see it indeed gets closer and closer as n increases.

Here we fix a large S to make sure the histograms are good approximations for the density of X_n^*.

The Rnotebook file also on Nestor. Open it with Rstudio, and the rests come easily as shown by our introduction video on Nestor. Exam of our course this time **would not** involve any codes, but you will need to learn some basic coding ideas for the following course, e.g., Probability Distribution; and when you get stuck somewhere, simulations by codes are also a nice way to explore possible directions.

Each line of code can be executed using ctrl + Enter in Rstudio.

This draws 4 values from Unif(0,1)

```
runif(4, min = 0, max = 1)
```

Here are the codes that you may find how many are the largest number

```
rand.a=runif(4, min = 0, max = 1)
sum(rand.a==max(rand.a))
```

We may repeat this game many many times and check how many more than 1 winner in S repeated games via the following codes:

```
n=4;
S=100;
rand.Srepeats=matrix(runif(n*S, min = 0, max = 1),ncol=S)
colMax<- apply(rand.Srepeats, 2, max)
# this gives you the a True and false matrix, and only true if
   the max in that column
result=colMax==t(matrix(rep((colMax),n),nrow=S))
# how many more than 1 winner in S repeated games
sum(colSums(result)>1)
```

Here we comparing the histogram of X* (we generate S values drawn from the same distribution as the one of X* and then draw histogram) with the density function of a normal distribution.

```
set.seed(12)
require(ggplot2)
n=4
S=100
sigma2=1/12
# we generate S realised values of X* (simu_Sbar) by drawing
   from Unif distribution
simu_Sbar= colSums(matrix(runif(n*S, min = 0, max = 1)-1/2,ncol
   =S))/sqrt(n)

ggplot(data.frame(simu_Sbar), aes(simu_Sbar)) +
geom_histogram(aes(y=..density..)) +
stat_function(fun=function(x)1/sqrt(2*pi*sigma2)*exp(-(x)^2/
    sigma2/2),
color=rgb(0.6, 0.2, 0.2, 0.35), size=2)
```

Here are also Python codes

```
import numpy as np
np.random.seed(10)

# now again this draws 4 values from Unif(0,1)
np.random.uniform(0,1,4)

n=4;
S=100;
rand_Srepeats= np.random.uniform(0,1,n*S).reshape(-1,S)
# how many more than 1 winner in S repeated games
np.sum(np.sum(rand_Srepeats==rand_Srepeats.max(axis=0), axis
    =0)>1)


# X* distribution analysis
n=4
S=100
sigma2=1/12

rand_Srepeats= np.random.uniform(0,1,n*S).reshape(-1,S) -1/2
simu_Sbar= np.sum(rand_Srepeats.reshape(-1,S),axis=0)/np.sqrt
    (n)
#  Histogram
import seaborn as sns
sns.distplot(simu_Sbar, hist=True, kde=True,
bins=int(180/5), color = 'darkblue',
hist_kws={'edgecolor':'black'},
kde_kws={'linewidth': 4})

# we add some normal density
import math
ax=sns.distplot(simu_Sbar, hist=True, kde=True,
bins=int(180/5), color = 'darkblue',
hist_kws={'edgecolor':'black'},
kde_kws={'linewidth': 4})
# calculate the pdf
sigma2=1/12
x0, x1 = ax.get_xlim()  # extract the endpoints for the x-
```

```
    axis
x_pdf = np.linspace(x0, x1, 100)

y_pdf = 1/(np.sqrt(2*math.pi*sigma2))*np.exp(-np.power(x_pdf
    ,2)/(2*sigma2))

ax.plot(x_pdf, y_pdf, 'r', lw=2, label='pdf')
ax.legend()
```