Homework_08

October 29, 2019

Probability for Data Science

UC Berkeley, Fall 2019

Ani Adhikari and Jim Pitman

1 Homework 8

1.0.1 Instructions

Your homeworks have two components: a written portion and a portion that also involves code. Written work should be completed on paper, and coding questions should be done in the notebook. You are welcome to LaTeX your answers to the written portions, but staff will not be able to assist you with LaTeX related issues. It is your responsibility to ensure that both components of the homework are submitted completely and properly to Gradescope. Refer to the bottom of the notebook for submission instructions.

[1]:

```
# Run this cell to set up your notebook

import numpy as np
from scipy import stats
from datascience import *
from prob140 import *

# These lines do some fancy plotting magic
import matplotlib
%matplotlib inline
import matplotlib.pyplot as plt
plt.style.use('fivethirtyeight')

# These lines make warnings look nicer
import warnings
warnings.simplefilter('ignore', FutureWarning)
```

1.0.2 1. The Exact Distribution of a Sum

In this exercise we will use the same shorthand as in the textbook: ”A random variable W has distribution given by the probabilities $p_0, p_1, \dots, p_N$” means that $P(W = i) = p_i$ for $0 \le i \le N$ and $\sum_{i=0}^{N} p_i = 1$.

Before you start this exercise, carefully go through the code in Section 14.2 of the textbook. As always, feel free to create more code cells as needed.

a) [**CODE**] Let X have the distribution given by $p_0 = 0.45$, $p_1 = 0.25$, $p_3 = 0.2$, $p_4 = 0.05$, $p_5 = 0.05$. Construct the pgf of X.

b) [**CODE**] Let $X_1, X_2, \dots, X_8$ be i.i.d. with the same distribution as X in (a). Let $S_X = X_1 + X_2 + \cdots + X_8$. Use `Plot` to plot the probability histogram of S_X.

c) [**CODE**] Find $P(S_X = 13)$.

d) [**CODE**] Let Y have the uniform distribution on the integers 4 through 8. Let $Y_1, Y_2, \dots, Y_{12}$ be i.i.d. with the same distribution as Y, and let $S_Y = Y_1 + Y_2 + \cdots + Y_{12}$. Use `Plot` to plot the histogram of the distribution of $W = S_X + S_Y$.

e) [**CODE**] For a `prob140` distribution object `dist`, the expression `dist.ev()` evaluates to the expectation and `dist.sd()` evaluates to the SD. At this point you should already have a distribution object representing W, so use these methods to find $E(W)$ and $SD(W)$. To check that you found the right distribution of W, use `.ev()` and `.sd()` to find the expectations and SDs of X and Y, and then use rules of expectation and variance to find $E(W)$ and $SD(W)$. Confirm that these are the same as what you got from directly using the distribution of W.

[2]:

```
#Answer to 1a

# Construct the distribution of X
dist_X = Table().with_column("X",np.array([0,1,2,3,4,5]))
dist_X = dist_X.with_column("probability",np.array([0.45,0.25,0,0.2,0.05,0.05]))
                              #"probability",np.array([0.45,0.25,0.2,0.05,0.
 ↪05]))

# Extract the array of probabilities
probs_X = dist_X['probability']

# Get the coefficients of the pgf in the appropriate order
coeffs_X = np.flipud(probs_X)

# Construct the pgf
pgf_X = np.poly1d(coeffs_X)

# Display the pgf
print(pgf_X)
```

```
      5          4         3
0.05 x + 0.05 x + 0.2 x + 0.25 x + 0.45
```

[3]:

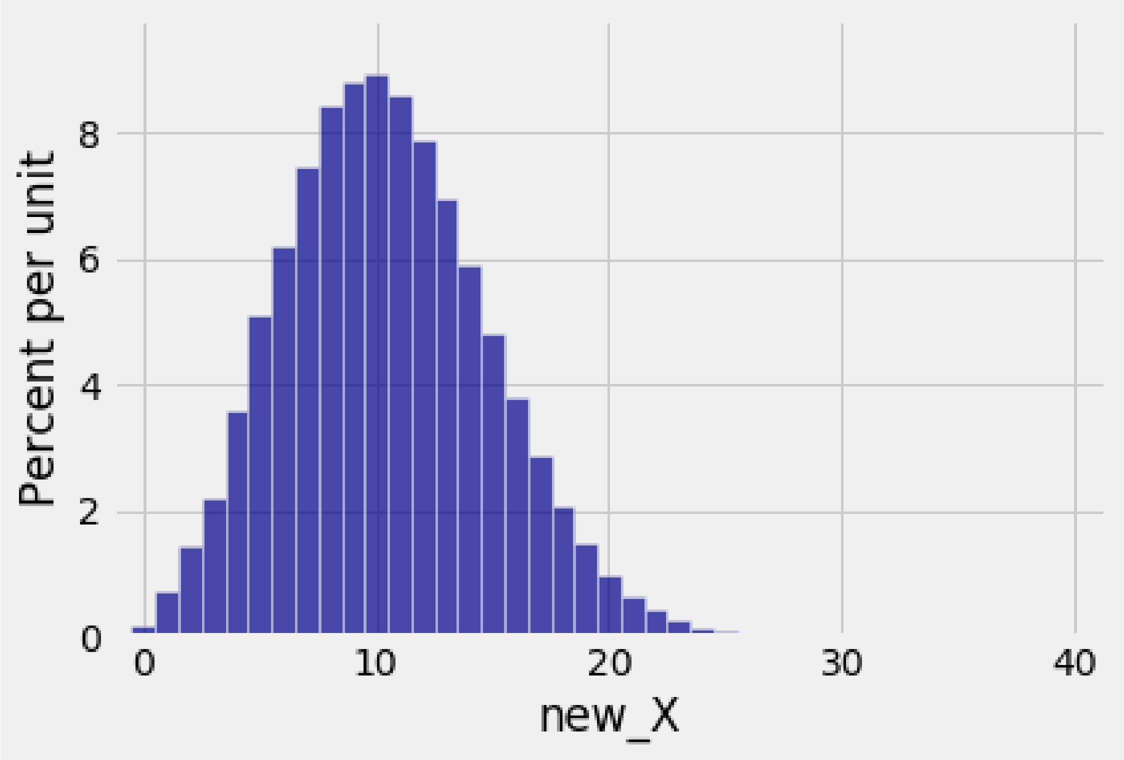
```
#Answer to 1b

pgf_SX = pgf_X**8 # pgf of S_X
coeffs_SX =np.flipud(pgf_SX.c) # coefficients of pgf of S_X

# Distribution object for S_X
# Careful ...
# Think how you will extract the possible values and corresponding chances.
# Use extra lines if you need them.

dist_SX = Table().with_column("new_X",np.arange(len(coeffs_SX)))
dist_SX = dist_SX.with_column("probability",coeffs_SX)

Plot(dist_SX)
```



[4]:

```
#Answer to 1c
dist_SX.where('new_X',13)['probability']
```

[4]:

```
array([0.06964777])
```

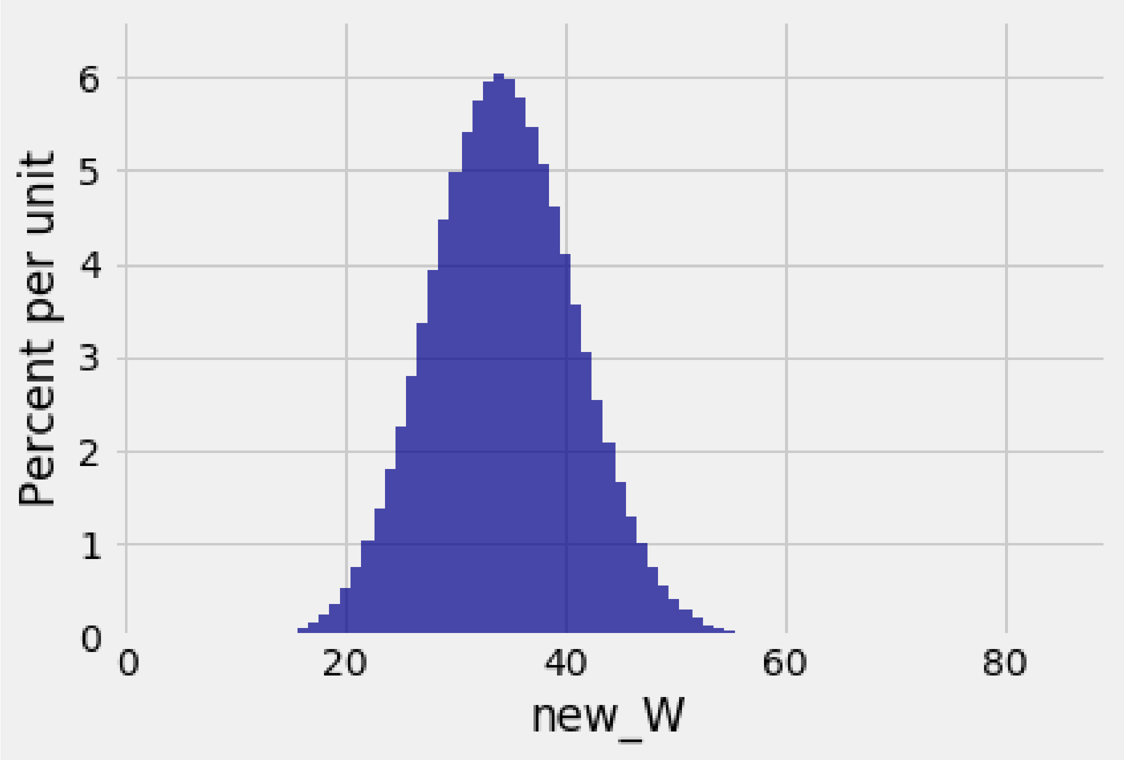
[5]:

```
#Answer to 1d

# The following is provided as a very brief skeleton
# The solution uses more lines than are provided in the skeleton
# Be sure to refer to your setup for (1a) and (1b)

#dist_Y = ... # distribution object for Y
dist_Y = Table().with_column("Y",np.arange(4,9))
dist_Y = dist_Y.with_column("probability",np.array([0.2,0.2,0.2,0.2,0.2]))
pgf_Y = np.poly1d(dist_Y['probability']) # pgf of Y
pgf_SY = pgf_Y**12 # pgf of S_Y
pgf_W = pgf_SX * pgf_SY # pgf of W = S_X + S_Y
#dist_W = ... # distribution object for W
dist_W = Table().with_column("new_W",np.arange(len(pgf_W.c)))
dist_W = dist_W.with_column("probability",np.flipud(pgf_W.c))

Plot(dist_W)
```



[6]:

```
#Answer to 1e
dist_NY = Table().with_column("new_y",np.arange(len(pgf_SY.c)))
dist_NY = dist_NY.with_column("probability",np.flipud(pgf_SY.c))

# Use dist_W here
print("E(W) =", dist_W.ev())
print("SD(W) =", dist_W.sd())
```

```python
# Use dist_X and dist_Y here

print("E(W) =", dist_SX.ev() + dist_NY.ev())
print("SD(W) =", (dist_SX.sd()**2+dist_NY.sd()**2)**(1/2))
```

```
E(W) = 34.40000000000003
SD(W) = 6.578753681359413
E(W) = 34.400000000000006
SD(W) = 6.578753681359412
```

2 newpage

2.0.1 2. What's Normal?

Before you answer this question, please read all of Section 14.5 of the textbook. We did almost all of it in lecture on Thursday 10/17 but stopped a bit before the end. The bit we didn't do is a review of Data 8.

As a preliminary (which is also in the textbook section), let Φ be the standard normal cdf, that is, $\Phi(z) = P(Z \leq z)$ where Z is a standard normal random variable. Then you know that for a specified z you can find $\Phi(z)$ by using `stats.norm.cdf(z, mean, sd)`:

[7]:

```python
z = 2
stats.norm.cdf(2, 0, 1)
```

[7]:

```
0.9772498680518208
```

The function Φ^{-1} returns the z for a specified value of Φ. That is, $\Phi^{-1}(p)$ is the value of z such that $\Phi(z) = p$.

In the `stats` module, Φ^{-1} is called the "percent point function" and the call is `stats.norm.ppf(p, mean, SD)`:

[8]:

```python
stats.norm.ppf(0.995, 0, 1)
```

[8]:

```
2.5758293035489004
```

In any part of this question that involves a sample size, you can assume the sample size is big enough for the Central Limit Theorem approximation to be good. But pay attention to what is being approximated by the CLT.

a) In a simple random sample of 1000 faculty taken among all universities in a country, the number of papers published by the sampled faculty in the past year had a mean of 1.1 and an SD of 1.8. Does the Central Limit Theorem say that the distribution of the number of papers published by the sampled faculty in the past year is roughly normal? If not, what do you think is the shape of that distribution? Explain based on the information given in the problem.

b) Continuing Part **a**, construct an approximate 90% confidence interval for the mean number of papers published by faculty at all universities in the country in the past year. Justify your answer. If it is not possible to construct the interval, explain why not.

3 newpage

3.0.1 3. Widths of Confidence Intervals

In any part of this question that involves a sample size, you can assume the sample size is big enough for the Central Limit Theorem approximation to be good.

a) A survey organization has used the methods of our class to construct an approximate 95% confidence interval for the mean annual income of households in a county. The interval runs from $\$66,000$ to $\$70,000$. If possible, find an approximate 99% confidence interval for the mean annual income of households in the county. If this is not possible, explain why not.

b) A survey organization is going to take a simple random sample of n voters from among all the voters in a state, to construct a 99% confidence interval for the proportion of voters who favor a proposition. Find an n such that the total width of the confidence interval (left end to right end) will be no more than 0.06. Remember that you can bound the variance of an indicator.

4 newpage

4.0.1 4. A Mixture

This is adapted from a problem from Pitman's text.

Transistors produced by one machine have a lifetime that is exponentially distributed with mean 100 hours. Those produced by a second machine have an exponentially distributed lifetime with mean 200 hours. A package of 12 transistors contains 4 produced by the first machine and 8 produced by the second. Let X be the lifetime of a transistor picked at random from the package.

We say that the distribution of X is a *mixture* of the two exponential distributions. Conditioning is the most natural way to study mixtures. **Answer each part below by conditioning.**

a) Find the numerical value of $P(X > 200)$. You don't have to turn in the code; just show your math, then create a cell in any of your notebooks to calculate the value, and report the value at the end of your math calculation. For a number c, the expression `np.exp(c)` evaluates to e^c.

b) Find the numerical value of $E(X)$.

c) For $x > 0$, find $P(X \in dx)$ and hence find the density of X.

```
[9]: p_larger200 = (4/12)*(np.exp(-2))+(4/12)*(np.exp(-1))
     p_larger200
```

```
[9]: 0.16773824146935168
```

And #newpage

4.0.2 5. Relations Between Three Well Known Distributions

a) Let U be uniform on $(0, 1)$ and let $X = -\log(U)$. Find the possible values of X and the cdf of X. Recognize that X has a well known distribution and provide its name and parameters.

b) Products of uniform $(0, 1)$ random samples arise when the data are "fractions of fractions of fractions of ..." some quantity. Let $U_1, U_2, \ldots, U_n$ be an i.i.d. uniform $(0, 1)$ sample and let $Y_n = (U_1U_2 \cdots U_n)^{\frac{1}{n}}$ be the *geometric mean* of the sample. Show that when n is large the distribution of $\log(Y_n)$ is close to one of the famous ones, and provide its name and parameters.

c) Let Z be standard normal and let $W = e^Z$. Then $\log(W) = Z$, that is, the log of W has a normal distribution. That is why the distribution of W is called *lognormal.* Find the cdf of W in terms of the standard normal cdf Φ, and hence find the density of W in terms of the standard normal density ϕ. State the possible values of W.

5 newpage

5.1 Submission Instructions

Many assignments throughout the course will have a written portion and a code portion. Please follow the directions below to properly submit both portions.

5.1.1 Written Portion

- Scan all the pages into a PDF. You can use any scanner or a phone using an application. Please **DO NOT** simply take pictures using your phone.
- Please start a new page for each question. If you have already written multiple questions on the same page, you can crop the image or fold your page over (the old-fashioned way). This helps expedite grading.
- It is your responsibility to check that all the work on all the scanned pages is legible.

5.1.2 Code Portion

- Save your notebook using File > Save and Checkpoint.
- Generate a PDF file using File > Download as > PDF via LaTeX. This might take a few seconds and will automatically download a PDF version of this notebook.
  - If you have issues, please make a follow-up post on the general HW 8 Piazza thread.

5.1.3 Submitting

- Combine the PDFs from the written and code portions into one PDF. Here is a useful tool for doing so.
- Submit the assignment to Homework 8 on Gradescope.
- **Make sure to assign each page of your pdf to the correct question.**

- **It is your responsibility to verify that all of your work shows up in your final PDF submission.**

5.1.4 We will not grade assignments which do not have pages selected for each question.

[ ]:

hw8

2.

a.

⇒ The central limit theorm does say that when n is large, the sum and the mean is normal. but it doesn't say anything abt the sample itself.

⇒ thus, the CTL don't say anything about the sample itself, so the dist of the sample is not roughly normal, and the distribution have some random shape.

b.

⇒ $\phi^{-1}(0.95)$ = stats.norm.pdf(0.95)

$z = 1.64485$

⇒ $0.9 \approx P(\mu \in (\overline{X}_n - z_\lambda \frac{\sigma}{\sqrt{n}}, \overline{X}_n - z_\lambda \frac{\sigma}{\sqrt{n}})$

⇒ since $z_\lambda = 1.64485$

⇒ $\mu \in (1.1 - 1.64485 * (\frac{1.8}{\sqrt{1000}}), 1.1 + 1.64485 * (\frac{1.8}{\sqrt{1000}}))$

3.

a.

$\Rightarrow$ mean $= 66{,}000 + \frac{70{,}000 - 66{,}000}{2} = 68{,}000$

$\Rightarrow$ since it 95% → the z is 1.96.

$\Rightarrow \quad 68000 - 1.96 * \frac{\sigma}{\sqrt{n}} = 66{,}000$

$68000 - 1.96 * \frac{\sigma}{\sqrt{n}} = 70{,}000.$

$\Rightarrow \frac{\sigma}{\sqrt{n}} = \frac{2000}{1.96}$

$\Rightarrow$ stats. norm. ppf (0.995) = 2.5758

$\Rightarrow$ new interval

$68000 - 2.5758 * \frac{2000}{1.96} \approx 65372$

$68000 + 2.5758 * \frac{2000}{1.96} \approx 70628$

$\Rightarrow$ thus, the new interval is

$\mu \in (65372, 70628)$

b.

$\Rightarrow$ for 99% → $z =$ stats. norm. ppf (0.995) = 2.5758

$\Rightarrow$ total width of the confidence interval ≤ 0.06

$\Rightarrow$ since it can only favor / not favor a population

⇒ the population is a binominal (n, p)

→ where p is the chance that a voter will voter favor → chance who vote not favor $= q = 1-p$.

⇒ $E(\text{porportion}) = \frac{n*p}{n} = p$

⇒ $\text{variance} = \frac{n*p*q}{n^2} = \frac{p*q}{n}$

⇒ $SD(\text{porportion}) = \frac{\sqrt{pq}}{\sqrt{n}}$

⇒ $2.5758 * \left(\frac{\sqrt{pq}}{\sqrt{n}}\right) / \sqrt{n} \leq 0.03$

$2.5758 * \frac{\sqrt{pq}}{n} \leq 0.03$

⇒ $\sqrt{pq} = \sqrt{0.5*0.5} = 0.5$

⇒ $n \geq \frac{2.5758*0.5}{0.03}$

4.

a.

⇒ $\text{mean} = E(T) = \frac{1}{\lambda}$

⇒ $\lambda = \frac{1}{\text{mean}}$

⇒ $\lambda_{x_1} = \frac{1}{100}$ & $\lambda_{x_2} = \frac{1}{200}$

⇒ density machine 1 = $\frac{1}{100} * e^{-\frac{1}{100}t}$ & density machine 2 = $\frac{1}{200} * e^{-\frac{1}{200}t}$

$S_{machin1}(t) = e^{-\frac{1}{100}t}$ & $S_{machine\ 2} = e^{-\frac{1}{200}t}$

⇒ $P(X>200) = P(\text{machine } 1) * P(\text{machine } 1 > 200)$
$+$
$P(\text{machin } 2) * P(\text{machine } 2 > 200$

⇒ $P(x>200) = \frac{4}{12} * e^{-\frac{1}{100}*200} + \frac{8}{12} * e^{-\frac{1}{200}*200}$

$= \frac{4}{12} * e^{-2} + \frac{8}{12} * e^{-1}$

$= 0.167738$

⇒ thus. $P(X>200)$ is 0.167738

b.

⇒ $E(X) = P_{x_1} * E(X_1) + P_{x_2} * E(X_2)$

$= \frac{4}{12} * 100 + \frac{8}{12} * 200$

$= \frac{2000}{12} = \frac{500}{3}$

⇒ Thus, the numerical value for $E(x)$ is $\frac{500}{3}$ year.

c.

⇒ $P(X \in dx) = \int_{x-a}^{x+a} f(x)\,dx = F(x)$

→ let $\frac{1}{2}\Delta dx = a$ & $\lim_{a \to 0}$.

$\Rightarrow P(X \in dx) = \frac{4}{12} * \left(1 - e^{-\frac{1}{100}t}\right)\Big|_{x-a}^{x+a}$.

$\frac{8}{12} + \left(1 \overset{+}{-} e^{-\frac{1}{200}t}\right)\Big|_{x-a}^{x+a}$

$\Rightarrow$ for $\lim_{a \to 0} \Rightarrow P(X \in dx)$

$= \frac{4}{12} * e^{\frac{x}{-100}} + \frac{8}{12} * e^{\frac{x}{-200}}$

$\Rightarrow$ density of $X = \frac{d}{dx} F(x)$

$= \frac{4}{12} * \left(-\frac{x}{100}\right) * e^{\frac{-x}{100}} + \frac{8}{12} * \left(-\frac{x}{200}\right) * e^{\frac{-x}{200}}$

$\Rightarrow$ thus, $P(X \in dx) = \frac{4}{12} * e^{\frac{-x}{100}} + \frac{8}{12} * e^{\frac{-x}{200}}$, and

density of $X = \frac{4}{12} * \left(-\frac{x}{100}\right) * e^{\frac{-x}{100}} + \frac{8}{12} * \left(-\frac{x}{200}\right) * e^{\frac{-x}{200}}$.

5.

a.

$\Rightarrow X = -\log(u)$ & $0 < u < 1$

$\Rightarrow 0 < X < \infty$

$\Rightarrow$ thus the possible value of X is $0 < X < \infty$.

$\Rightarrow$ for cdf of X

$\Rightarrow F(x) = P(X < x)$

$= P(-\log(u) < x)$

$= P(u^{-1} < e^{x})$

$= p(u > e^{-x})$

$\Rightarrow$ since u is uniformly distribute

$\Rightarrow P(u > e^{-x}) = 1 - e^{-x}$

$\Rightarrow$ thus cdf of $x = P(X < x) = 1 - e^{-x}$

$\Rightarrow$ since $F(x) = 1 - e^{-x}$

$\Rightarrow f(x) = \frac{d}{dx} F(x) = e^{-x}$

$\Rightarrow$ thus, X is exponentional distribution with parameter $\lambda = 1$

b.

$\Rightarrow$ since $0 < u < 1$

$\Rightarrow 0 < Y_n < 1$

$\Rightarrow P(Y_n < x) = P(\log((U_1 U_2 \cdots\cdots U_n)^{1/n}) < x)$

$$\Rightarrow Y_n = \log((U_1 U_2 U_3 \cdots\cdots U_n)^{1/n})$$

$$= \log(U_1^{\frac{1}{n}} * U_2^{\frac{1}{n}} \cdots\cdots * U_n^{\frac{1}{n}})$$

$$= \log(U_1^{\frac{1}{n}}) + \log(U_2^{\frac{1}{n}}) + \cdots\cdots + \log(U_n^{\frac{1}{n}})$$

$$= \frac{1}{n}\log(U_1) + \frac{1}{n}\log(U_2) + \cdots\cdots + \frac{1}{n}\log(U_n)$$

$$= \frac{1}{n}\sum_{i=1}^{n}\log(U_i) \leftarrow \text{mean of the sample}$$

$\Rightarrow$ CTL say that the mean of the sample is normal

⇒ so Y_n have a normal distribution

⇒ mean of $Y_n = E(Y_n) = E(\frac{1}{n} \sum_{i=1}^{n} \log(U_i))$

$$= \frac{1}{n} * \sum_{i=1}^{n} E(\log(U_i))$$

$$= \frac{1}{n} * n * E(\log(U_i)) \leftarrow E(\log(U_i)) = \int_0^1 (\log(U_i))$$

$$= -1$$

$$= \frac{1}{n} * n * (-1)$$

$$= -1$$

⇒ SD of $Y_n = \sqrt{\sum_{i=1}^{n} \frac{Var(\log(U_i))}{n^2}}$

⇒ $Var((\log(U_i)) = E((\log(U_i))^2) + [E(\log(U_i))]^2$

$$= 2 - (-1)^2 = 2 - 1$$

$$= 1$$

⇒ SD of $Y_n = \sqrt{\frac{n * 1}{n^2}} = \sqrt{\frac{1}{n}} = \frac{1}{\sqrt{n}}$

⇒ thus Y_n is normal distribution with parameter $(-1, \frac{1}{\sqrt{n}})$.

c.

⇒ $P(W < x) = P(e^z < x)$

$$= P(z < \log(x))$$

$$= \Phi(\log(x))$$

$\Rightarrow$ density of $W = \frac{d}{dx} \Phi(\log(x))$

$$= \frac{1}{x} \Phi(\log(x))$$