Homework_08

October 30, 2019

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.

```
In [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 $p_0, p_1, \ldots, p_N$" means that $P(W = i) = p_i$ for $0 \leq i \leq N$ 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 $X_1, X_2, \ldots, X_8$ 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 $Y_1, Y_2, \ldots, Y_{12}$ 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.

```
In [3]: #1a
        dist_X = Table().values(make_array(0,1,2,3,4,5))\
            .probabilities(make_array(0.45, 0.25, 0, 0.2, 0.05, 0.05))
        probs_X = dist_X.column(1) # Extract array of probabilities for X
        coeffs_X = np.flipud(probs_X) # put the coefficients in the appropriate order
        pgf_X = np.poly1d(coeffs_X) # pgf of X
        print(pgf_X)
```

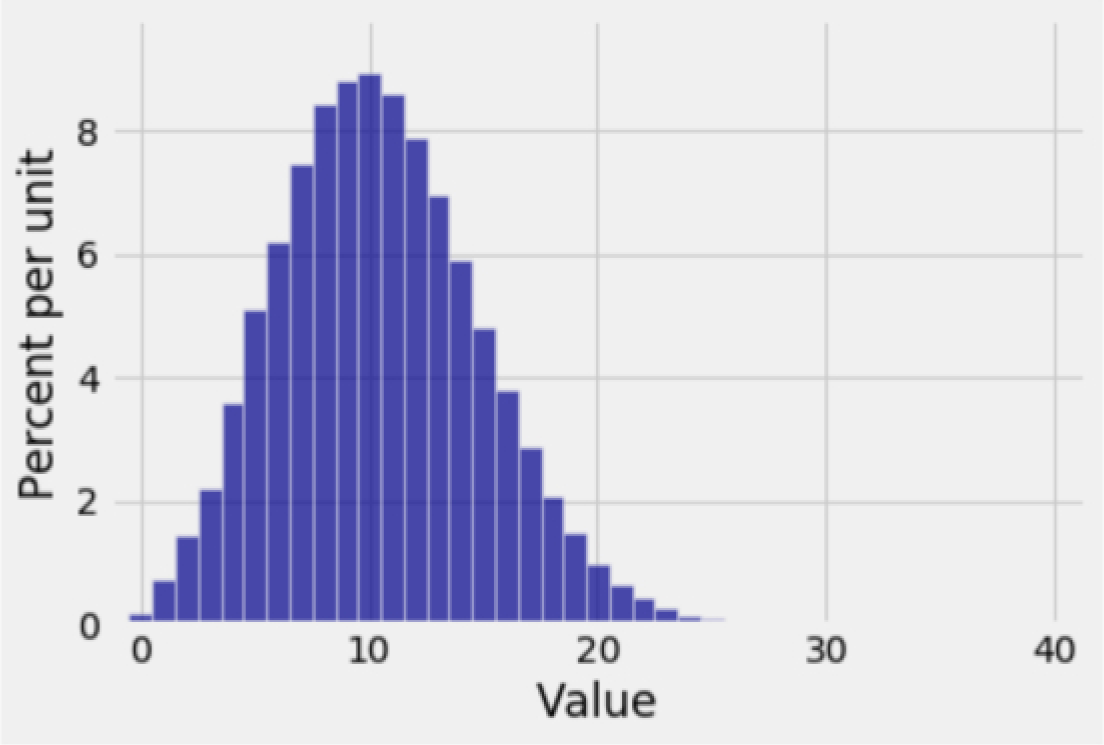
```
      5        4       3
0.05 x + 0.05 x + 0.2 x + 0.25 x + 0.45
```

```
In [13]: #1b


         pgf_SX = pgf_X**8      # PGF of S_X
         coeffs_SX = pgf_SX.c   # Coefficients of PGF


         dist_SX = Table().values(np.arange(len(coeffs_SX)))\
             .probabilities(np.flipud(coeffs_SX)) # Distribution object for S_x

         Plot(dist_SX)
```



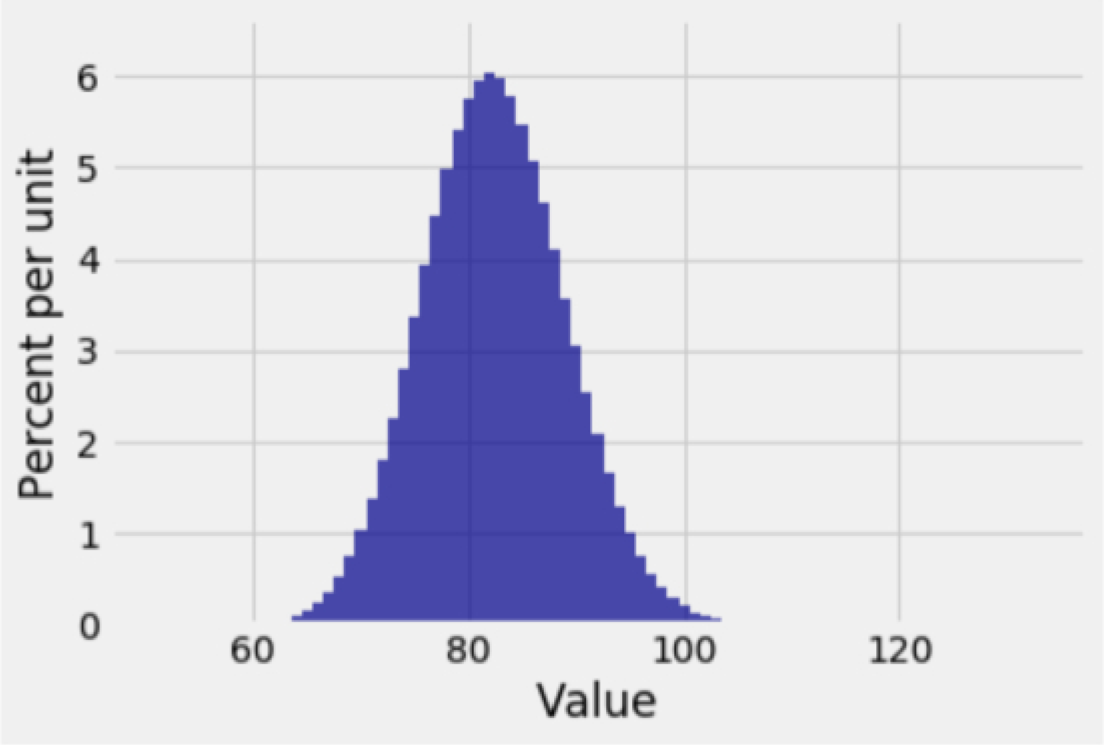
```
In [7]: #1c
        dist_SX.column(1).item(13)
```

```
Out[7]: 0.06964776875000002
```

```
In [14]: #1d
         probs_Y = np.append(0*np.ones(4), 0.2*np.ones(5))
         dist_Y = Table().values(np.arange(len(probs_Y)))\
             .probabilities(probs_Y) # distribution object for Y
         coeffs_Y = np.flipud(probs_Y)
         pgf_Y = np.poly1d(coeffs_Y) # 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
         coeffs_W = pgf_W.c
         probs_W = np.flipud(coeffs_W)
         dist_W = Table().values(np.arange(len(probs_W)))\
             .probabilities(probs_W) # distribution object for W

         Plot(dist_W)
```



```
In [11]: #Answer to 1e

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

         # Use dist_X and dist_Y here

         print("E(W) =", 8*dist_X.ev() + 12*dist_Y.ev())
         print("SD(W) =", np.sqrt(8*dist_X.sd()**2 + 12*dist_Y.sd()**2))
```

```
E(W) = 82.40000000000005
SD(W) = 6.578753681359412
E(W) = 82.4
SD(W) = 6.578753681359411
```

#newpage

1.0.3 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)`:

```
In [3]: z = 2
        stats.norm.cdf(2, 0, 1)
```

```
Out[3]: 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)`:

```
In [4]: stats.norm.ppf(0.9772498680518208, 0, 1)
```

```
Out[4]: 2.0000000000000004
```

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.

1.0.4 [Solution] What's Normal?

a) The Central Limit Theorem only states that large random sample sums and means are normally distributed - not the sample itself. In addition, it does not make sense for the shape of the distribution to be normal because this implies that some faculty members have published a negative number of papers. The shape of this distribution is likely to be skewed right with a hard lower bound of 0.

b) $1.1 \pm 1.645 \times \frac{1.8}{\sqrt{1000}}$. Even though the underlying distribution of the number of papers is not normal, the CLT implies that the probability distribution of the average number of papers in a large random sample will be roughly normal. The large sample size also allows us to estimate $\sigma/\sqrt{n}$ based on the corresponding quantity in the sample.

#newpage

1.0.5 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.

1.0.6 [Solution] Widths of Confidence Intervals

a) The observed value of $\bar{X}_n$ is \$68,000.

The confidence interval is $\bar{X}_n \pm z\sigma/\sqrt{n}$, where z depends on the level of confidence.

z for a 95% confidence interval is 1.96 (`stats.norm.ppf(0.975)`)

$1.96\sigma/\sqrt{n} = (70000 - 66000)/2 = 2000$

$\sigma/\sqrt{n} = 1020.4$

z for a 99% confidence interval is 2.576.

99% confidence interval for the population mean: $68000 \pm 2.576 \times 1020.4$

b) If the population proportion of voters who favor the proposition is p, then the random sample proportion has expectation p and SD $\sqrt{pq/n}$.

The largest possible value of $\sqrt{pq}$ is 0.5, achieved at $p = 0.5$.

We want $2.576\sqrt{0.25/n} = 0.06/2 = 0.03$

$n = 1843.27$, round up to $n = 1844$

#newpage

1.0.7 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.

1.0.8 [Solution] A Mixture

a) P(transistor is from first machine) = $\frac{4}{12} = \frac{1}{3}$

The lifetime of a transistor from Machine 1 has exponential $(1/100)$ distribution.

The survival function of a random variable Y that has the exponential (λ) distribution is $P(Y > y) = e^{-\lambda y}$ for $y > 0$.

By conditioning on the type of machine, $P(X > 200) = \frac{1}{3}e^{\frac{-200}{100}} + \frac{2}{3}e^{\frac{-200}{200}}$

b) By iteration, $E(X) = \frac{1}{3} \times 100 + \frac{2}{3} \times 200 = 166.667$

c) For Y with exponential (λ) distribution, $f_Y(y)dy \sim P(Y \in dy) \sim \lambda e^{-\lambda y}dy$ for $y > 0$.

By conditioning,

$P(X \in dx) \sim \frac{1}{3} \cdot \frac{1}{100}e^{-x/100}dx + \frac{2}{3} \cdot \frac{1}{200}e^{-x/200}dx$
Therefore, $f_X(x) = \frac{1}{3} \cdot \frac{1}{100}e^{-x/100} + \frac{2}{3} \cdot \frac{1}{200}e^{-x/200}$
#newpage

1.0.9 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.

1.0.10 [Solution] Relations Between Three Well Known Distributions

a) Possible values of X: 0 to ∞. For $x > 0$,

$F_X(x) = P(X \le x) = P(-\log(U) \le x) = P(U \ge e^{-x}) = 1 - e^{-x}$ because for any $u \in (0,1)$, $P(U \ge u) = 1 - u$.

This is the CDF of the exponential (1) distribution.

b) $\log(Y_n)$ is roughly normally distributed when n is large because $\log(Y_n) = \frac{1}{n}(\log(U_1) + \log(U_2) + ... + \log(U_n))$, which is the average of n iid random variables.

The parameters of the normal are $E(\log(U_1))$ and $\frac{SD(\log(U_1))}{\sqrt{n}}$.

In Part **a** we showed that $-\log(U_1)$ is exponential (1). So $E(\log(U_1)) = -1$ and $SD(\log(U_1)) = 1$.

So for large n, the distribution of $\log(Y_n)$ is approximately normal $(-1, \frac{1}{\sqrt{n}})$.

c) For $w > 0$, $F_W(w) = P(W < w) = P(e^Z < w) = P(Z < \log(w)) = \Phi(\log(w))$.

By differentiation, $f_W(w) = \frac{1}{w}\phi(\log(w))$ for $w > 0$.
#newpage

1.1 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.

1.1.1 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.

1.1.2 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.

1.1.3 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.**

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