ProblemSet2

October 16, 2023

```
[ ]: import numpy as np
     import matplotlib.pyplot as plt
```

```
[ ]: %matplotlib inline
```

1 Problem Set 2

1.0.1 PHYS 441

Adam A Miller
02 October 2023
version 0.2

Problem sets for PHYS 441/DATA SCI 421 are due 1 week after they are assigned at 11:59 pm.

Every student is responsible for submitting their own *individual* solutions. Solutions must be either an executable jupyter notebook or Adobe PDF file submitted via Canvas. You must **show all your work** (if you submit a pdf, be sure the pdf does not cut off text or lines of code). *Submissions that only include answers will have points deducted.*

If the problem set calls for an integral, please calculate the integral by hand (in general any problem with integrals will not require the use of mathematica or similar tools).

If you wish to "write mathematics" in a Jupyter notebook, this can be done using LaTeX formatting.

LaTeX is great at typesetting mathematics. Let $X_1, X_2, \ldots, X_n$ be a sequence of independent and identically distributed random variables with $\mathrm{E}[X_i] = \mu$ and $\mathrm{Var}[X_i] = \sigma^2 < \infty$, and let

$$S_n = \frac{X_1 + X_2 + \cdots + X_n}{n} = \frac{1}{n} \sum_{i}^{n} X_i$$

denote their mean. Then as n approaches infinity, the random variables $\sqrt{n}(S_n - \mu)$ converge in distribution to a normal $\mathcal{N}(0, \sigma^2)$.

You can find a summary of all the LaTeX math symbols from Overleaf.

1.1 Problem 1) 12 points

Suppose the probability function for a random variable X is described by

$$f(x) = \begin{cases} cx^3 & 0 \leq \text{x} \leq 4 \\ 0 & \text{otherwise} \end{cases}$$

Problem 1a What is the value of c?

The p.d.f. $f(x)$ is normalized such that the total probability is one, i.e.$\int_{-\infty}^{\infty} f(x)dx = 1$. Hence

$$\begin{aligned} \int_{-\infty}^{\infty} f(x)dx &= && (1) \\ \int_0^4 cx^3 dx &= && (2) \\ c\int_0^4 x^3 dx &= && (3) \\ c \cdot \frac{1}{4} x^4 \Big|_0^4 &= 1 && (4) \end{aligned}$$

$\Rightarrow$

$$\begin{aligned} c \cdot \frac{1}{4}(4^4 - 0^4) &= && (5) \\ c \cdot \frac{1}{4} \cdot 256 &= && (6) \\ c \cdot 64 &= 1 && (7) \end{aligned}$$

$\Rightarrow$

$$c = \frac{1}{64} \qquad (8)$$

Problem 1b

What is the cumulative distribution function $F(x)$?

$$F(x) = \int_{-\infty}^{x} f(t)dt \qquad (9)$$

Hence,

$$F(x) = \begin{cases} 1 & \text{x} > 1 \\ \frac{1}{256}x^4 & 0 \leq \text{x} \leq 4 \\ 0 & x < 0 \end{cases} \tag{10}$$

Problem 1c

Calculate the mean for the random variable X.

$$\begin{align} E[X] &= \int_{-\infty}^{\infty} x \cdot f(x)dx \tag{11} \\ &= \int_{0}^{4} x \cdot \frac{1}{64}x^3 dx \tag{12} \\ &= \frac{1}{64}\int_{0}^{4} x^4 dx \tag{13} \\ &= \frac{1}{64} \cdot \frac{1}{5}x^5\Big|_0^4 \tag{14} \\ &= \frac{1}{64 \cdot 5} \cdot (4^5 - 0^5) \tag{15} \\ &= \frac{1024 - 0}{64 \cdot 5} \tag{16} \\ &= 3.2 \tag{17} \end{align}$$

Problem 1d

Calculate the variance for X.

$$\begin{align} \text{Var}(X) &= \int_{-\infty}^{\infty} (x - \mu)^2 \cdot f(x)dx \tag{18} \\ &= \int_{0}^{4} (x - \frac{16}{5})^2 \cdot \frac{1}{64}x^3 dx \tag{19} \\ &= \frac{1}{64}\int_{0}^{4} (x - \frac{16}{5})^2 \cdot x^3 dx \tag{20} \\ &= \frac{1}{64}\int_{0}^{4} (x^2 - \frac{32}{5}x + \frac{256}{25}) \cdot x^3 dx \tag{21} \\ &= \frac{1}{64} \cdot (\frac{1}{6}x^6\Big|_0^4 + \frac{32}{25}x^5\Big|_0^4 + \frac{256}{100}x^4\Big|_0^4) \tag{22} \\ &= \frac{1}{64} \cdot (\frac{4096}{6} + \frac{32 \cdot 1024}{25} + \frac{256 \cdot 256}{100}) \approx 41.38 \tag{23} \end{align}$$

1.2 Problem 2) 10 points

Let X and Y be independent variables with variances σ_x^2 and σ_y^2.

Let $Z = X + Y$.

Problem 2a

What is the variance σ_z^2?

$$\begin{aligned}
\text{Var}[Z] &= \text{Var}[X+Y] && (24)\\
&= E[(X+Y)^2] - (E[X+Y])^2 && (25)\\
&= (E[X^2] + E[Y^2] + 2E[XY]) - (E[X] + E[Y])^2 && (26)\\
&= (E[X^2] + E[Y^2] + 2E[XY]) - [(E[X])^2 + (E[Y])^2 + 2E[X]E[Y]] && (27)\\
&= [E[X^2] - (E[X])^2] + [E[Y^2] - (E[Y])^2] + (2E[XY] - 2E[X]E[Y]) && (28)
\end{aligned}$$

Since X and Y are independent, $E[XY] = E[X]E[Y]$, so

$$\begin{aligned}
2E[XY] - 2E[X]E[Y] &= 2E[X]E[Y] - 2E[X]E[Y] && (29)\\
&= 0 && (30)
\end{aligned}$$

Hence,

$$\begin{aligned}
\text{Var}[Z] &= [E[X^2] - (E[X])^2] + [E[Y^2] - (E[Y])^2] + 0 && (31)\\
&= [E[X^2] - (E[X])^2] + [E[Y^2] - (E[Y])^2] && (32)\\
&= \text{Var}[X] + \text{Var}[Y] && (33)\\
&= \sigma_x^2 + \sigma_y^2 && (34)
\end{aligned}$$

Problem 2b

What is the covariance σ_{zx}^2?

$$\begin{aligned}
\text{Cov}(X,Z) &= \text{Cov}(X, X+Y) && (35)\\
&= E[(X - E[X])((X+Y) - E[X+Y])] && (36)\\
&= E[(X - E[X])(X + Y - E[X] - E[Y])] && (37)\\
&= E[(X - E[X])(X - E[X] + Y - E[Y])] && (38)\\
&= E[(X - E[X])(X - E[X]) + (X - E[X])(Y - E[Y])] && (39)\\
&= E[(X - E[X])^2] + E[(X - E[X])(Y - E[Y])] && (40)\\
&= \text{Var}[X] + \text{Cov}(X,Y) && (41)
\end{aligned}$$

Since X and Y are independent, $\text{Cov}(X,Y) = 0$. Hence,

$$\begin{aligned}
\text{Cov}(X,Z) &= \text{Var}[X] + \text{Cov}(X,Y) && (42)\\
&= \text{Var}[X] + 0 && (43)\\
&= \sigma_x^2 && (44)
\end{aligned}$$

1.3 Problem 3) 14 points

Consider the following array of data:

[12.849, 9.185, 6.973, 0.727, 11.376, 0.881, 18.355, 13.544, 16.946, 16.907, 9.959, 4.328, 5.435, 33.384, 4.128, 5.882, 3.58, 4.405, 5.438, 9.55, 5.508, 5.183, 3.988, 8.082, 1.927]

(you may conveniently execute the cell below to load these data into an array called `dat`

```
[ ]: dat = np.array([12.849, 9.185, 6.973, 0.727, 11.376, 0.881, 18.355, 13.544, 16.
      ↪946, 16.907, 9.959, 4.328, 5.435, 33.384, 4.128, 5.882, 3.58,  4.405, 5.438,␣
      ↪9.55, 5.508, 5.183, 3.988, 8.082, 1.927])
     print(len(dat))
```

```
25
```

Problem 3a

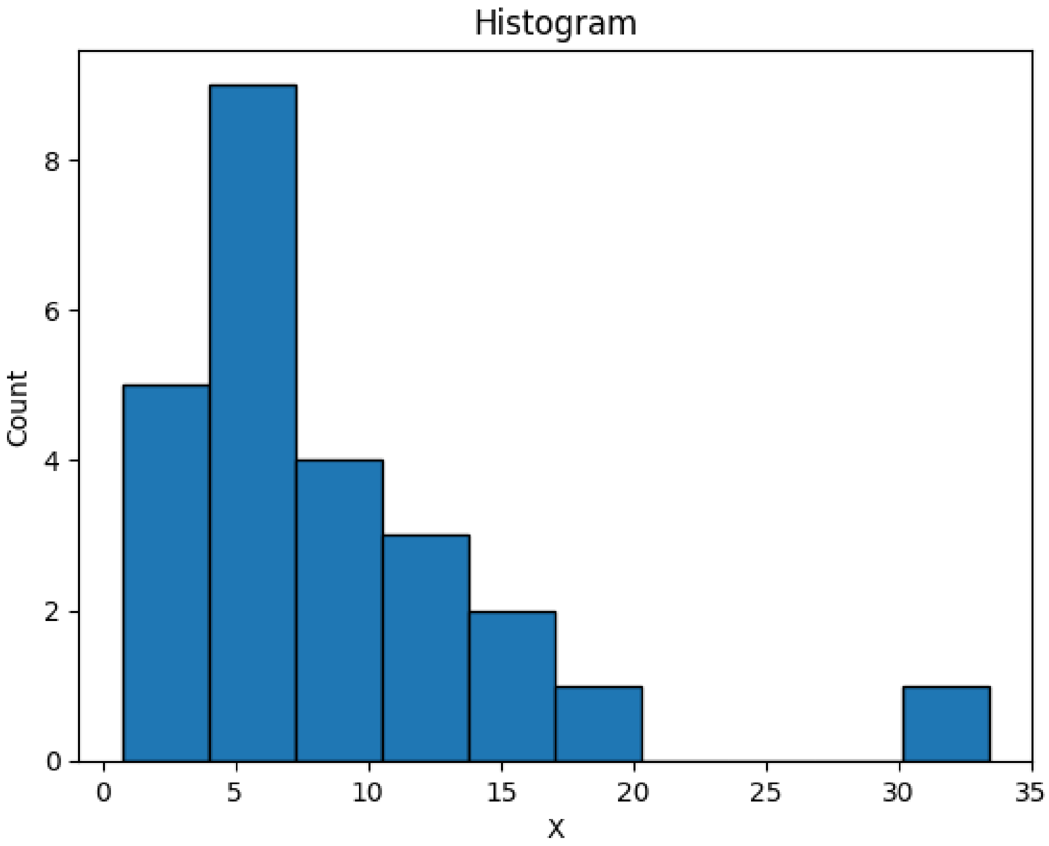
Estimate the p.d.f. for these data by displaying them as a histogram.

Hint – use the default settings for whatever plotting program you are using

```
[ ]: plt.hist(dat, edgecolor='black')  # You can adjust the number of bins as needed

     plt.xlabel('X')
     plt.ylabel('Count')
     plt.title('Histogram')

     plt.show()
```



Problem 3b

Using the same data plot two new histograms, but adjust the plotting parameters – in the first set the total number of bins equal to 6, in the second center the bars on the right edge of the bins.

Hint – the latter can be done by setting `align = right`.

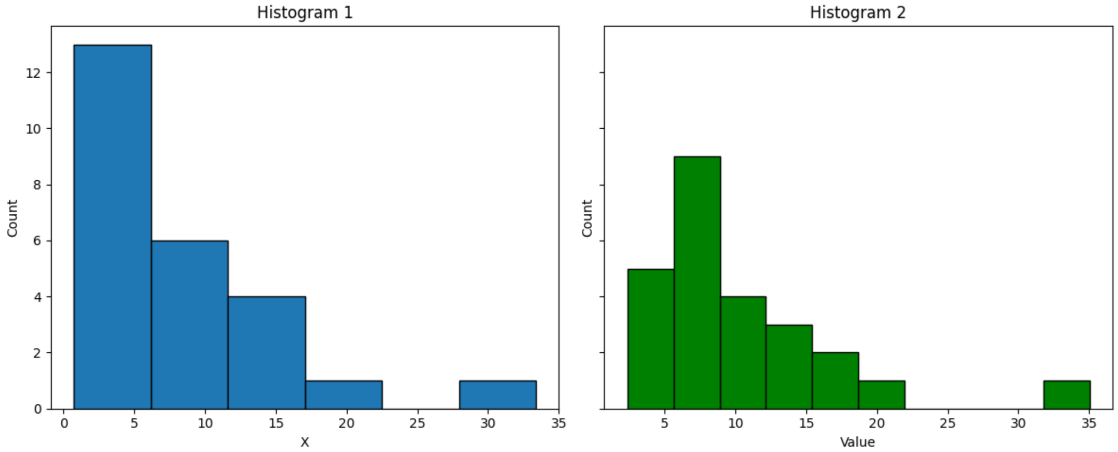
[ ]:

```
fig, (ax1, ax2) = plt.subplots(1, 2, figsize=(12, 5), sharey=True)

ax1.hist(dat, bins=6, edgecolor='black')
ax1.set_xlabel('X')
ax1.set_ylabel('Count')
ax1.set_title('Histogram 1')

ax2.hist(dat, edgecolor='black', color='green', align='right')
ax2.set_xlabel('Value')
ax2.set_ylabel('Count')
ax2.set_title('Histogram 2')

plt.tight_layout()
plt.show()
```



Problem 3c

You now have 3 different estimates for $f(x)$, but they are all very different. Does this make sense? Which of the three estimates is best?

Not really, since the $f(x)$ is being estimated by the frequency of numbers in a given range. The accuracy is dependent on the number of bins used to plot the histogram. The one with the most bins is the best estimate.

Here we have demonstrated one of the major challenges with using histograms to try and estimate p.d.f.s: production of the histogram requires user specified input which affects the output. Given that reasonable people may disagree on what parameters are reasonable, perhaps histograms are not the best tool for analysis (i.e., do not fit models to histograms).

For example, do you think it makes sense that $P(X = 25) = 0$? Each of the above histograms shows this is the case, but the true probability is likely greater than 0.

Problem 3d

Replot the histogram with a variable bin width such that each bin contains exactly 5 sources. Note – it is important that the histogram be normalized for this problem if you previously were not doing that.

Is this representation better?

[ ]:

```
import math

bin_edges = []
s_dat = sorted(dat)
bin_edges.append(s_dat[0])
for i in range(0, len(dat)):
    if i%5 == 0:
        bin_edges.append(s_dat[i])
bin_edges.append(s_dat[-1])
print(s_dat)
```

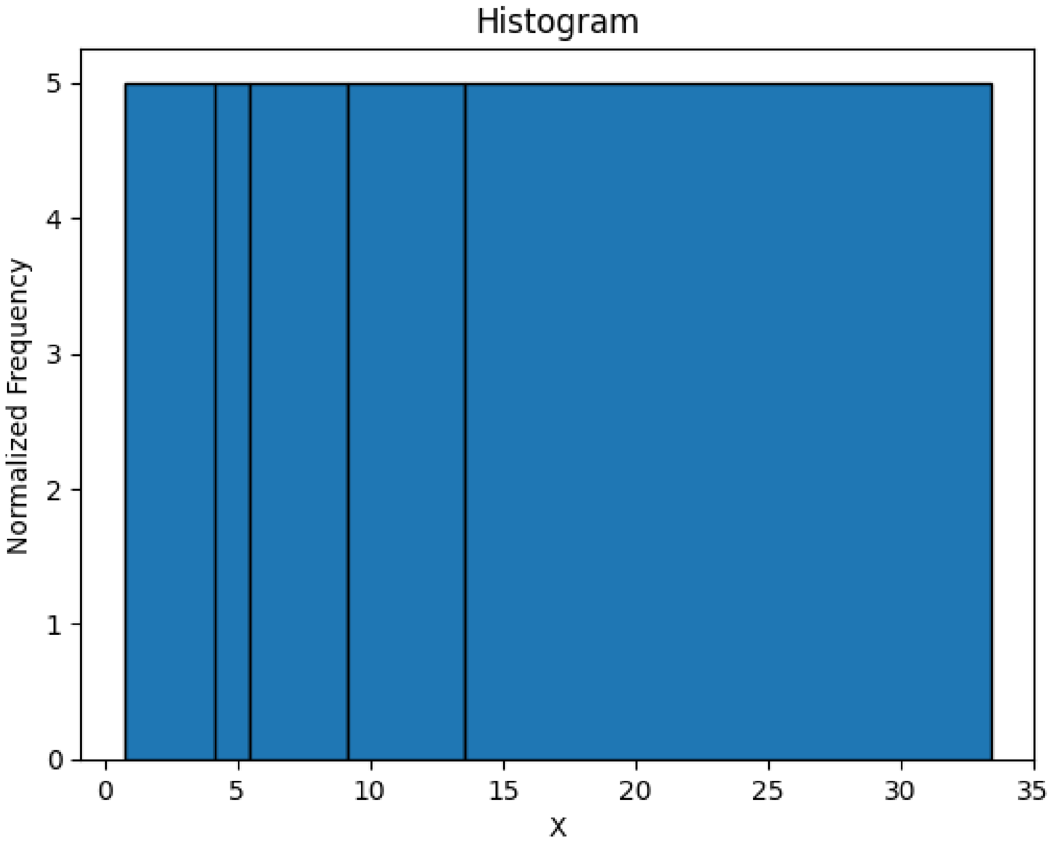
```python
print(bin_edges)

plt.hist(dat, bins=bin_edges, edgecolor='black')

plt.xlabel('X')
plt.ylabel('Normalized Frequency')
plt.title('Histogram')

plt.show()
```

```
[0.727, 0.881, 1.927, 3.58, 3.988, 4.128, 4.328, 4.405, 5.183, 5.435, 5.438,
5.508, 5.882, 6.973, 8.082, 9.185, 9.55, 9.959, 11.376, 12.849, 13.544, 16.907,
16.946, 18.355, 33.384]
[0.727, 0.727, 4.128, 5.438, 9.185, 13.544, 33.384]
```



I think this plot is a better representation of the data distribution. The high density around 5 corresponds and low density from 14-33 especially clearly corresponds to the characteristics of the original data distribution.

1.4 Problem 4) 18 points

Execute the cell below to load samples from the joint p.d.f. $f(x, y)$ into an array called `z`.

```
[ ]:
```

```
z = np.array([[2.4630e+00, 4.1940e+00], [-1.2260e+00, 3.2910e+00], [-8.
 ↪1600e-01, 4.3470e+00], [3.1040e+00, 4.9180e+00], [3.3880e+00, 4.7000e+00],␣
 ↪[1.3470e+00, 3.7620e+00], [3.9970e+00, 2.0420e+00], [1.5140e+00, 1.
 ↪7440e+00], [1.0730e+00, 1.4790e+00], [2.2650e+00, 2.2700e+00], [1.4230e+00,␣
 ↪2.4090e+00], [1.9440e+00, 1.6090e+00], [1.6800e+00, 5.0000e-03], [1.
 ↪3210e+00, 6.0590e+00], [5.8850e+00, 2.3880e+00], [-5.3300e-01, 1.0820e+00],␣
 ↪[4.3760e+00, 1.7210e+00], [9.1200e-01, 7.7560e+00], [2.1730e+00, 4.
 ↪2700e+00], [-8.6400e-01, 2.0420e+00], [1.1080e+00, 5.4400e+00], [2.2010e+00,␣
 ↪3.6520e+00], [4.1730e+00, 6.9550e+00], [1.2990e+00, 4.7300e-01], [4.
 ↪5800e-01, 5.5530e+00], [2.2000e+00, 2.0840e+00], [-1.8100e-01, 3.7700e-01],␣
 ↪[1.8990e+00, 8.5000e-01], [3.1210e+00, 1.5800e+00], [2.8250e+00, 3.
 ↪1500e+00], [3.0160e+00, 2.6770e+00], [6.1100e-01, 2.9480e+00], [3.7020e+00,␣
 ↪5.4580e+00], [2.4650e+00, 8.0780e+00], [-1.7200e-01, 1.9700e+00], [-9.
 ↪3800e-01, 5.3640e+00], [9.1700e-01, 4.2460e+00], [2.0840e+00, 2.8930e+00],␣
 ↪[1.3280e+00, 1.7470e+00], [1.6880e+00, 2.0630e+00], [-2.6200e-01, 4.
 ↪3720e+00], [1.2300e+00, 2.5860e+00], [8.9900e-01, 5.7210e+00], [3.3200e+00,␣
 ↪6.5390e+00], [1.5000e+00, 6.7490e+00], [2.7380e+00, 4.7300e+00], [1.
 ↪2020e+00, 2.9100e-01], [5.3100e-01, 2.1660e+00], [-2.1700e-01, 1.5490e+00],␣
 ↪[2.3260e+00, 1.8980e+00], [-3.9000e-01, 1.5480e+00], [1.7920e+00, 5.
 ↪3970e+00], [1.6840e+00, 3.1040e+00], [2.6750e+00, 1.6870e+00], [1.8850e+00,␣
 ↪1.8140e+00], [-6.9000e-01, 7.7000e-02], [4.5300e+00, 1.7850e+00], [4.
 ↪7290e+00, 3.0530e+00], [2.7600e+00, 3.9260e+00], [1.7990e+00, 1.7040e+00],␣
 ↪[4.1630e+00, 6.7770e+00], [1.1670e+00, 1.6900e+00], [1.6780e+00, 6.
 ↪0740e+00], [1.0100e+00, 5.2150e+00], [4.2950e+00, 4.1230e+00], [2.5810e+00,␣
 ↪4.8170e+00], [1.9240e+00, 3.3280e+00], [3.6240e+00, 2.7350e+00], [2.
 ↪9010e+00, 5.4980e+00], [9.1200e-01, 2.8800e-01], [2.1580e+00, 3.6530e+00],␣
 ↪[2.5240e+00, 1.3600e-01], [1.0620e+00, 9.0140e+00], [4.2600e+00, 2.
 ↪8480e+00], [1.3140e+00, 2.1440e+00], [2.8340e+00, 3.2750e+00], [3.2360e+00,␣
 ↪1.9060e+00], [2.3510e+00, 1.3650e+00], [1.8880e+00, 3.8510e+00], [2.
 ↪1790e+00, 3.8550e+00], [1.7500e+00, 6.0470e+00], [6.8000e-02, 3.1590e+00],␣
 ↪[1.3380e+00, 5.3700e-01], [3.7300e-01, 3.1400e-01], [9.3000e-01, 1.
 ↪3400e-01], [3.3400e+00, 3.0910e+00], [1.1560e+00, 1.2890e+00], [4.1630e+00,␣
 ↪4.6790e+00], [4.2600e-01, 6.3940e+00], [2.6310e+00, 1.7540e+00], [2.
 ↪6640e+00, 5.0410e+00], [2.4110e+00, 3.5480e+00], [1.9990e+00, 1.2610e+00],␣
 ↪[1.7680e+00, 4.8720e+00], [-1.5900e+00, 6.1290e+00], [4.2570e+00, 1.
 ↪9640e+00], [4.2530e+00, 4.6850e+00], [4.0700e-01, 3.5400e+00], [3.9570e+00,␣
 ↪1.5860e+00], [1.4400e-01, 9.2000e-01], [7.6100e-01, 7.2300e-01], [2.
 ↪6790e+00, 1.9560e+00], [4.1300e+00, 8.5300e-01], [1.5380e+00, 4.5290e+00],␣
 ↪[-3.6200e-01, 1.0630e+00], [4.7810e+00, 2.9150e+00], [5.3300e-01, 2.
 ↪8280e+00], [1.2730e+00, 2.9690e+00], [-8.8300e-01, 1.3340e+00], [4.0450e+00,␣
 ↪2.6110e+00], [3.7260e+00, 5.5660e+00], [2.7230e+00, 3.3920e+00], [2.
 ↪2310e+00, 5.2760e+00], [1.5430e+00, 3.7160e+00], [1.0630e+00, 5.7090e+00],␣
 ↪[3.3990e+00, 4.3810e+00], [3.2450e+00, 4.7550e+00], [1.1010e+00, 2.
 ↪0880e+00], [1.3790e+00, 2.1710e+00], [1.0730e+00, 5.9400e-01], [5.2100e-01,␣
 ↪1.7760e+00], [1.0800e+00, 3.9720e+00], [4.1340e+00, 3.3150e+00], [2.
 ↪5440e+00, 6.7890e+00], [2.6720e+00, 5.2650e+00], [5.0090e+00, 8.0710e+00],␣
 ↪[2.1410e+00, 5.5590e+00], [2.7160e+00, 2.3790e+00], [1.6790e+00, 3.
 ↪0100e+00], [5.0200e-01, 4.8190e+00], [3.1000e-02, 1.2520e+00], [1.5720e+00,␣
 ↪4.6100e-01], [1.4960e+00, 1.1180e+00], [3.4720e+00, 1.7670e+00], [9.
 ↪3600e-01, 4.0700e+00], [3.8240e+00, 3.5080e+00], [2.1420e+00, 4.9250e+00],␣
 ↪[3.4310e+00, 4.2400e+00], [-3.3800e-01, 5.8000e-01], [3.2140e+00, 1.
 ↪5130e+00], [2.6390e+00, 6.3080e+00], [3.8520e+00, 6.2160e+00], [9.6200e-01,␣
 ↪3.0010e+00], [1.9440e+00, 1.5680e+00], [3.6020e+00, 6.0220e+00], [-7.
```

```
250
```

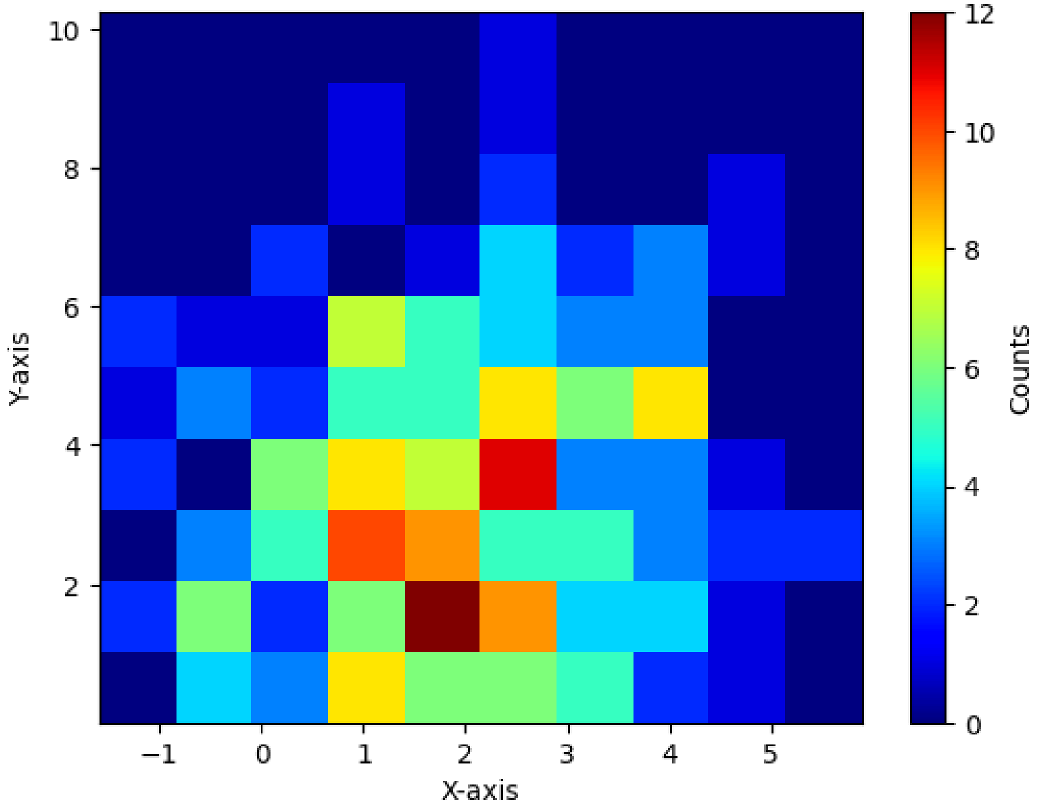
Problem 4a

Plot a 2d histogram to get a sense of the joint p.d.f.

[ ]:
```
plt.hist2d(z[:, 0], z[:, 1], cmap=plt.cm.jet)

# Add labels and a colorbar
plt.xlabel('X-axis')
plt.ylabel('Y-axis')
plt.colorbar(label='Counts')

# Show the plot
plt.show()
```

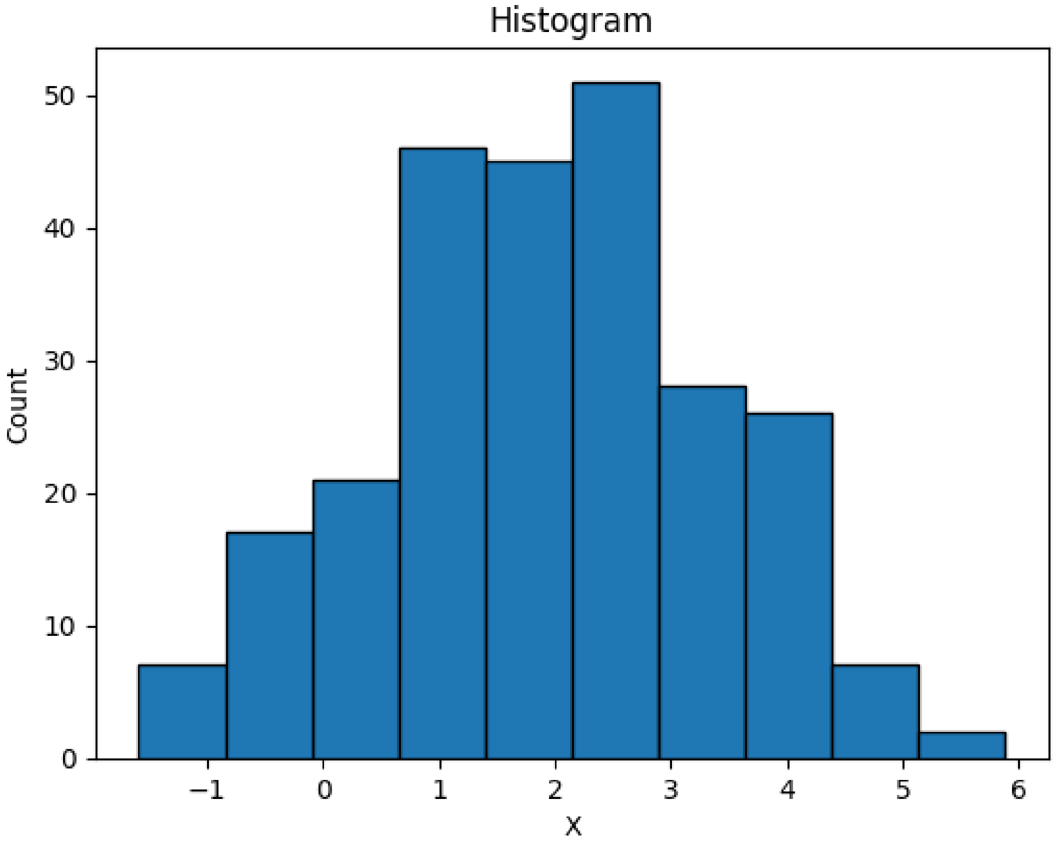


Problem 4b

Plot the marginal distribution along the first axis (the one that is often generically called “x”), $f_x(x)$.

[ ]:
```
plt.hist(z[:, 0], edgecolor='black')
plt.xlabel('X')
plt.ylabel('Count')
plt.title('Histogram')

plt.show()
```



Problem 4c

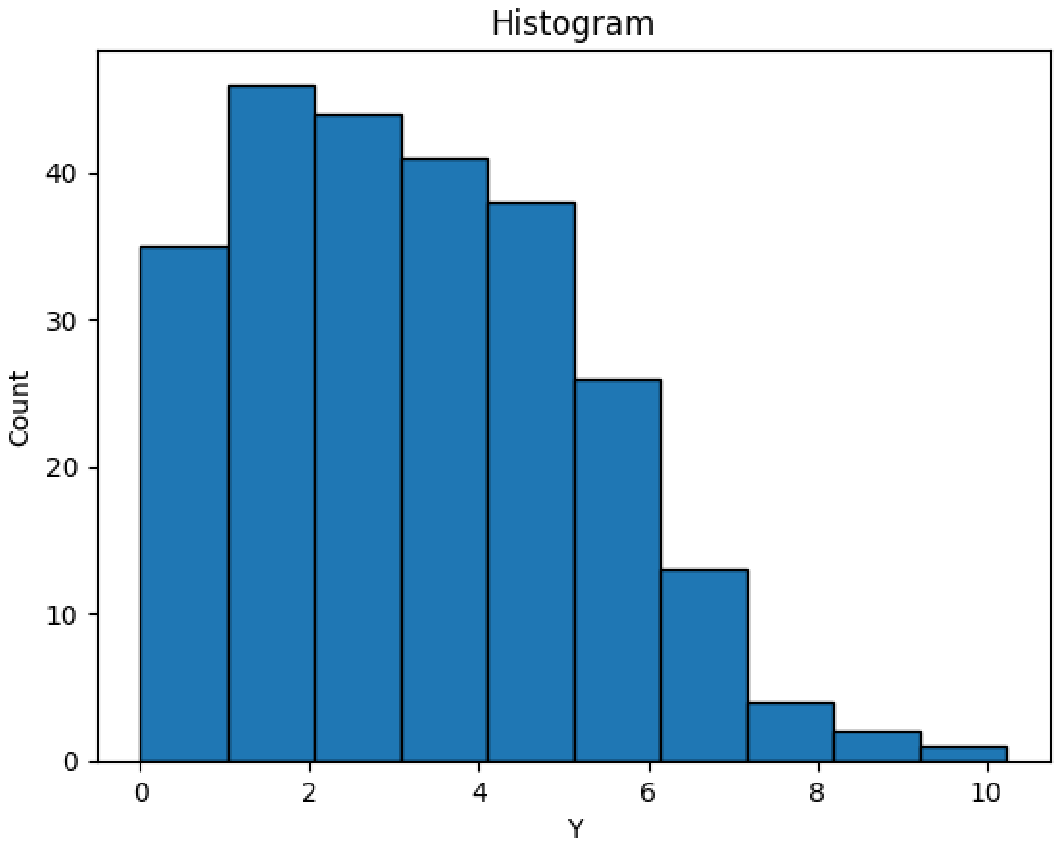
Plot the marginal distribution along the second axis (the one that is often generically called “y”), $f_y(y)$.

[ ]:

```
plt.hist(z[:, 1], edgecolor='black')

plt.xlabel('Y')
plt.ylabel('Count')
plt.title('Histogram')

plt.show()
```



Problem 4d

Which axis has more variance? Could you tell this from your plot of the joint distribution?

The X axis seems to have (slightly) more variance, although personally I cannot tell this from the joint histogram.

Problem 4e

Plot the conditional $h(y|x)$, for x given to be between 1 and 1.75.

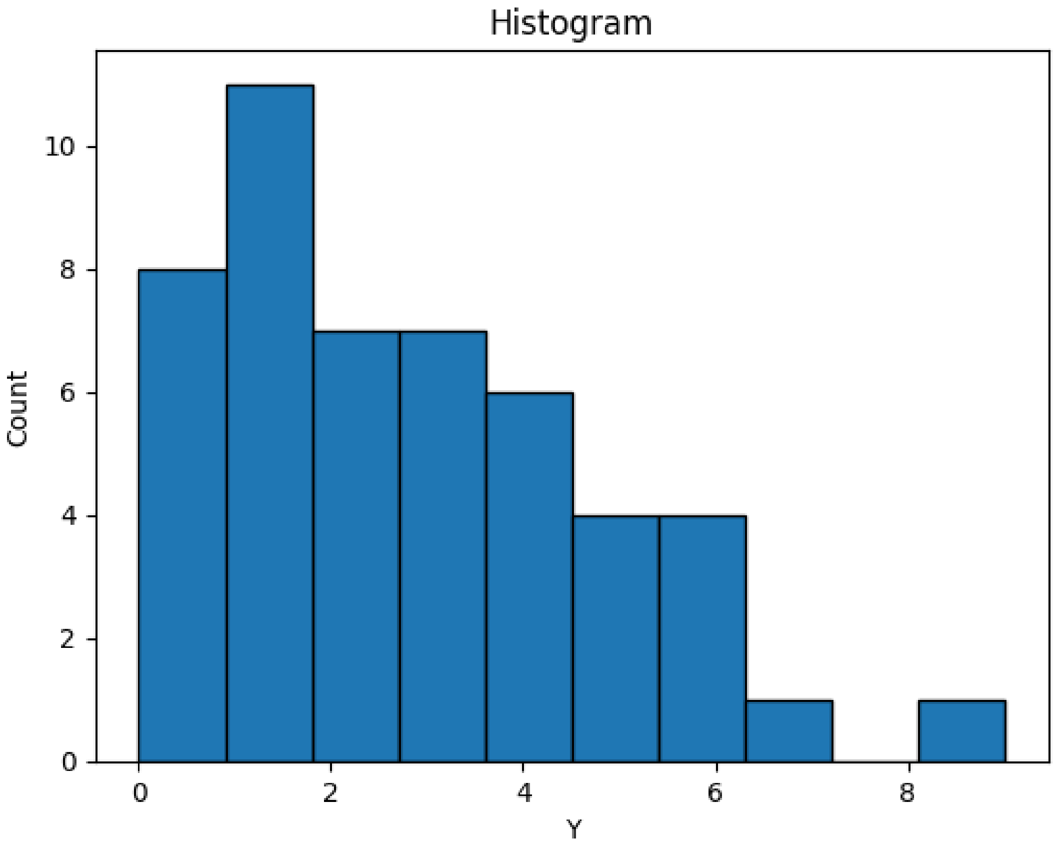
Does this distribution look different from $f_y(y)$? Which of these two has a higher median? Are x and y positively or negatively correlated?

[ ]:

```
h = z[(1 < z[:, 0]) & (z[:, 0] < 1.75)]

plt.hist(h[:, 1], edgecolor='black')
plt.xlabel('Y')
plt.ylabel('Count')
plt.title('Histogram')

plt.show()
```



```
[ ]: median_y = int(len(z[:, 1])/2)
     median_h = int(len(h[:, 1])/2)

     print(z[:, 1][median_y])
     print(h[:, 1][median_h])
```

```
8.071
5.709
```

The distribution looks slightly different than $f_y(y)$. In this case $f_y(y)$ has a higher median of 8.071. x and y are positively correlated.

1.5 Problem 5) 16 points

X and Y have the joint density function

$$f(x,y) = \begin{cases} 8xy & 0 \le x \le 1 \text{ and } 0 \le y \le \text{x} \\ 0 & \text{otherwise} \end{cases}$$

Note the range for y.

Problem 5a

Find the marginal density of X: $f_x(x)$

$$\begin{aligned}
f_x(x) &= \int_{-\infty}^{\infty} f_{XY}(x,y)dy && (45)\\
&= \int_{0}^{x} 8xydy && (46)\\
&= 8x\int_{0}^{x} ydy && (47)\\
&= 8x\cdot \frac{y^2}{2}\Big|_0^x && (48)\\
&= 8x\cdot (\frac{x^2}{2}-0) && (49)\\
&= 4x^3 && (50)
\end{aligned}$$

Hence,

$$f_x(x) = \begin{cases} 4x^3 & 0 \le x \le 1 \\ 0 & \text{otherwise} \end{cases}$$

Problem 5b

Find the marginal density of Y: $f_y(y)$

$$\begin{aligned}
f_y(y) &= \int_{-\infty}^{\infty} f_{XY}(x,y)dx && (51)\\
&= \int_{0}^{1} 8xydx && (52)\\
&= 8y\int_{0}^{1} xdx && (53)\\
&= 8y\cdot \frac{x^2}{2}\Big|_0^1 && (54)\\
&= 8y\cdot (\frac{1}{2}-0) && (55)\\
&= 4y && (56)
\end{aligned}$$

Hence,

$$f_y(y) = \begin{cases} 4y & 0 \le y \le 1 \\ 0 & \text{otherwise} \end{cases}$$

Problem 5c

Find the conditional density of X: $g(x|y)$.

$$
\begin{align}
g(x|y) &= \frac{P(x,y)}{P(y)} && \text{Probability of joint dist / Marginal dist of Y} \tag{57}\\
&= \frac{8xy}{4y} \tag{58}\\
&= 2x \tag{59}
\end{align}
$$

Hence,

$$g(x|y) = \begin{cases} 2x & 0 \le x \le 1, 0 \le y \le 1 \\ 0 & \text{otherwise} \end{cases}$$

Problem 5d

Find the condition density of Y: $h(y|x)$.

$$
\begin{align}
h(x|y) &= \frac{P(x,y)}{P(h)} && \text{Probability of joint dist / Marginal dist of X} \tag{60}\\
&= \frac{8xy}{4x^3} \tag{61}\\
&= \frac{8y}{4x^2} \tag{62}\\
&= \frac{2y}{x^2} \tag{63}
\end{align}
$$

Hence,

$$g(x|y) = \begin{cases} \frac{2y}{x^2} & 0 \le x \le 1, 0 \le y \le 1 \\ 0 & \text{otherwise} \end{cases}$$

Problem 5e

Are the random variables X and Y independent?

If X and Y are independent, $P(X \cap Y) = P(X) \cdot P(Y)$.

In this case, $P(X \cap Y) = 8xy \neq 4x^3 \cdot 4y$, hence X and Y are not independent.