ClaytonSchneider_HW1

August 28, 2019

[1]:
```
# Problem 1
# Say your research deals with social networks. Your first step is to study the␣
 ↪properties of the Facebook network of college students at CU-Boulder campus.␣
 ↪The next step is to compare your findings to the national college student␣
 ↪Facebook network.
# a) What are the populations you are concerned with?
# b) What is the relationship between these populations?
# c) What are some of the characteristics of the networks you might consider?␣
 ↪Pick three as an example.
# d) If you had infinite time and resources, would you be able to measure these␣
 ↪characteristics for every member of these populations?
# e) Say you don't have infinite time and resources -- how would you go about␣
 ↪estimating those population characteristics?
```

[2]:
```
#1a
"""
Students at CU Boulder who have a Facebook account, and college students in␣
 ↪general who have a Facebook account
"""


#1b
"""
#     ans)
"""


#1c
"""
Average number of friends
Average number of posts
Frequency of posts containing photos
"""


#1d
"""
Yes, though the averages are population-level characteristics
"""
```

```
#1e
"""
I would take smaller, random populations from CU and several other randomly␣
 ↪chosen schools
"""
```

```
[2]: '\nI would take smaller, random populations from CU and several other randomly
chosen schools\n'
```

```
[3]: # Problem 2
# You're working for a US public health surveillance team, keeping an eye on␣
 ↪infectious diseases such as the flu in the US.
# a) If your goal is to estimate the average yearly flu infection rate among␣
 ↪those over 65 years of age in the US, what is the population you are working␣
 ↪with?
# b) Given that surveillance is done only via doctor's offices, what is the␣
 ↪actual population of people whose infection rates you'll be observing?
# c) What kind of estimates will you get? Can they be generalized to the entire␣
 ↪population you'd like to be working with? Under what assumptions the answer␣
 ↪is yes?
```

```
[4]: #2a
"""
US citizens over the age of 65 who have medical access
"""


#2b
"""
The same, US citizens over the age of 65 who have medical access
"""


#2c
"""
#     ans)
"""
```

```
[4]: '\n#     ans) \n'
```

```
[5]: # Problem 3
# a) What is the difference between mean, median, and mode? When would you␣
 ↪prefer to use one and not the others?
# b) What is the difference between standard deviation and range? When would␣
 ↪you report one and not the other to communicate how variable the data are?
```

```
[6]: #3a
"""
mean: the sum of all data points divided by the population size
median: the middle-most data point after sorting a dataset
mode: the most common data point
```

```
You would do well to use a mean when your dataset is relatively balanced, with␣
 ↪a normal distribution. In such a case, the mode or median would yield a␣
 ↪similar result.
If data was unevenly distributed, none of these averages would suffice to␣
 ↪accurately represent the dataset.
"""


#3b
"""
Range is the difference between the two most outlier points in a dataset
#      ans)
"""
```

[6]:

```
'\nRange is the difference between the two most outlier points in a dataset\n#
ans)\n'
```

[7]:

```
# Problem 4
# A friend has given you a 2 pound bag of ordinary M&M's for your birthday.␣
 ↪Incidentally, you've recently had a discussion with the same friend who is␣
 ↪convinced that the blue ones are less frequent than the other 5 colors (red,␣
 ↪yellow, green, orange and brown). You (and the rest of your friends) think␣
 ↪that all colors are equally likely. The 2 pound bag has about 1200 M&M's. So␣
 ↪to put the matter to rest, you actually counted the M&M's and found there␣
 ↪were 1215 in the bag -- and you've found that there are 150 blue ones, 220␣
 ↪red, 230 yellow, 215 orange, 190 green, and 210 brown ones).
# a) Sketch a histogram of the observed relative frequency of colors in that␣
 ↪bag
# b) If your friend is not correct (and you are), what would the true relative␣
 ↪frequency of colors look like (sketch)
# c) How many blue ones would you expect to see if all colors are equally␣
 ↪likely?
# d) Do you think your friend is right, based on the one bag evidence? Give a␣
 ↪heuristic answer here - you don't need to be precise. What are some of the␣
 ↪limitations of this one-bag "evidence" approach? In an ideal world, how␣
 ↪would you design a study to test this more rigorously?
```

[1]:

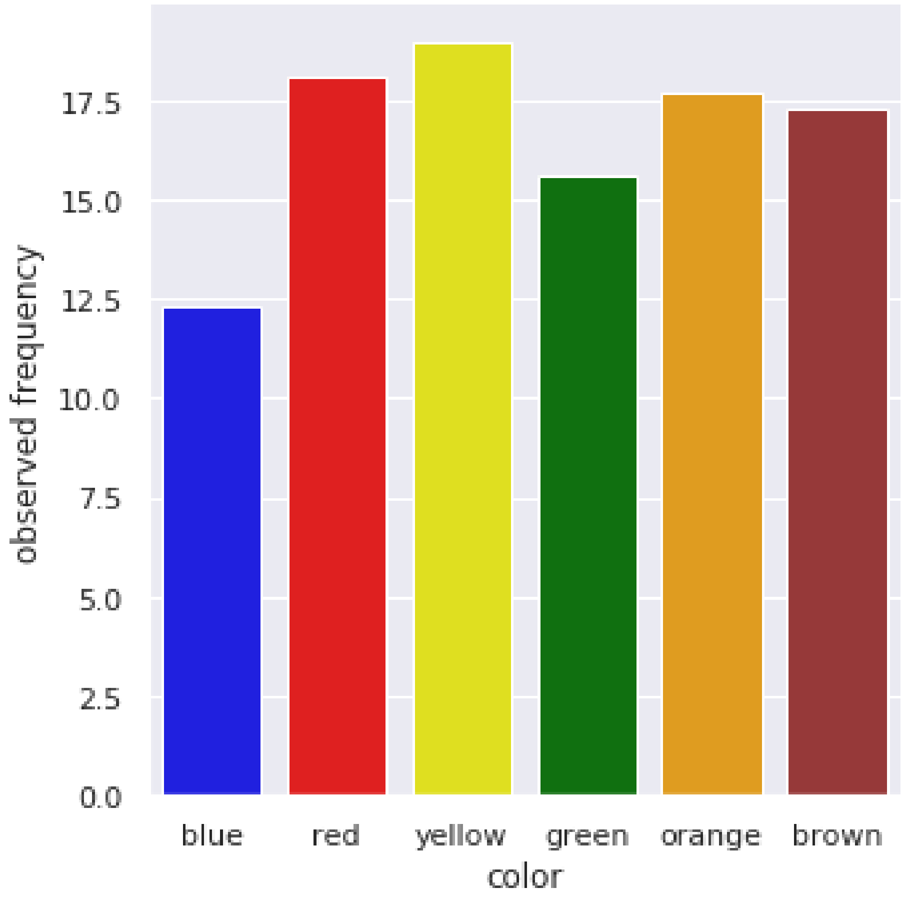
```
import matplotlib.pyplot as plt
import pandas as pd
import numpy as np
import seaborn as sns
sns.set()
```

[2]:

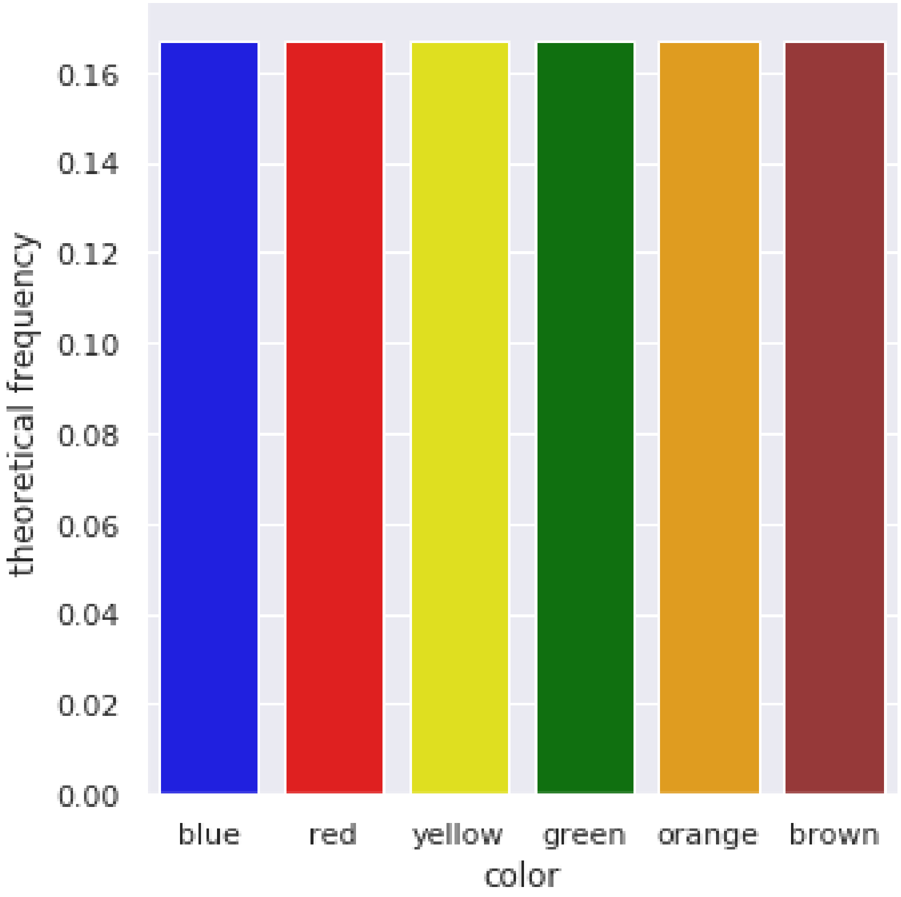
```
colors = ["blue", "red", "yellow", "green", "orange", "brown"]
counts = [150, 220, 230, 190, 215, 210]
total = np.sum(counts)
freqs = [100*count/total for count in counts]

mnm = pd.DataFrame({
    'color': colors,
    'count': counts,
    'observed frequency': freqs,
    'theoretical frequency': [(1/len(colors)) for color in colors]})
```

[3]:
```
#4a
observedData = sns.catplot(x="color", y="observed frequency", kind="bar",␣
 ↪palette=colors, data=mnm)
```



[4]:
```
#4b
theoreticalData = sns.catplot(x="color", y="theoretical frequency", kind="bar",␣
 ↪palette=colors, data=mnm)
```



[5]:

```
#4c
expectBlue = total / len(colors)
print("We would expect", expectBlue, "blue MnMs.")

#4d
"""
It is very well possible that there are fewer blue MnMs produced. It could␣
 ↪perhaps be cheaper to produce blue food coloring. Perhaps bags are made␣
 ↪randomly from an equally distributed manufacturing process. To verify, we'd␣
 ↪want to get a larger sample of data, with bags from various manufacturing␣
 ↪facilities if possible.
"""
```

```
We would expect 202.5 blue MnMs.
```

[5]:

```
"\nIt is very well possible that there are fewer blue MnMs produced. It could
perhaps be cheaper to produce blue food coloring. Perhaps bags are made randomly
from an equally distributed manufacturing process. To verify, we'd want to get a
```

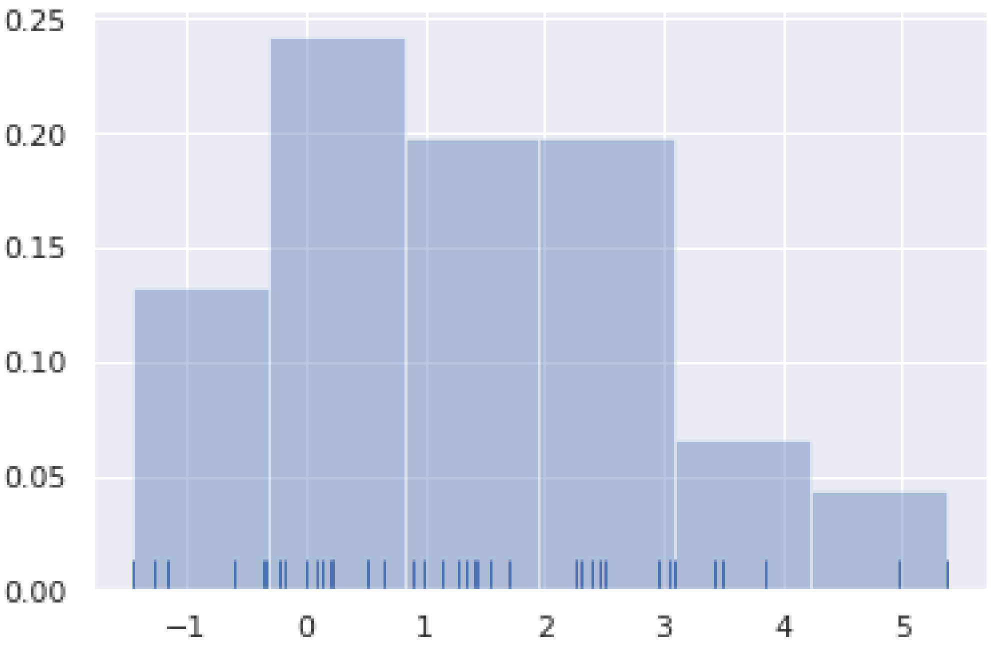
```
larger sample of data, with bags from various manufacturing facilities if
possible. \n"
```

[6]:
```
# Problem 5
# The following dataset (sample size=40) is given to you for further analysis␣
 ↪in a following text file.
# a) Plot a default histogram in your favorite software package/program. How␣
 ↪many bins does it plot by default for this dataset? What is the size of each␣
 ↪bin?
# b) Change the number of bins -- first use 10, then 20, and finally 25. What␣
 ↪differences (if any) can you see between these histograms and your histogram␣
 ↪from part (a)?
# c) Change the starting point to -2, -1.5, and then -1.45 -- and plot a␣
 ↪histogram with 25 bins for each. What differences do you see?
# You can also do this problem by hand if you choose. Some programs might not␣
 ↪let you change all these "inputs" - so if all else fails, sketches of␣
 ↪histograms by hand will be accepted.
```

[7]:
```
for line in open('/home/clayton/school/statisticalmethods/hw1/numbers.txt'):
    floats = [float(x) for x in line.split()]
```

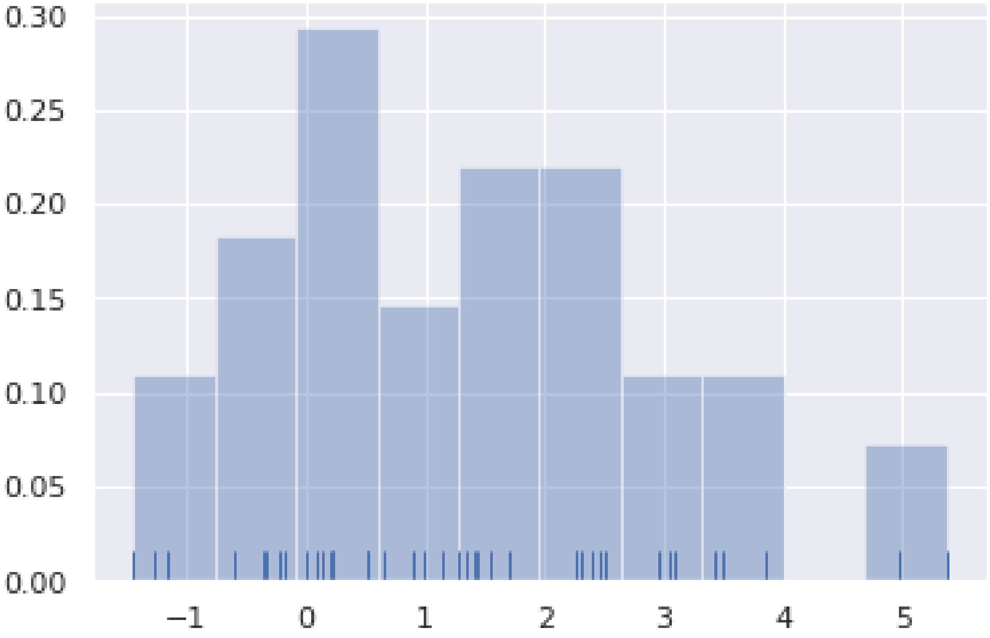
[8]:
```
#5a
sns.distplot(floats, kde=False, rug=True, norm_hist=True)
```

[8]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f1dae574cf8>
```



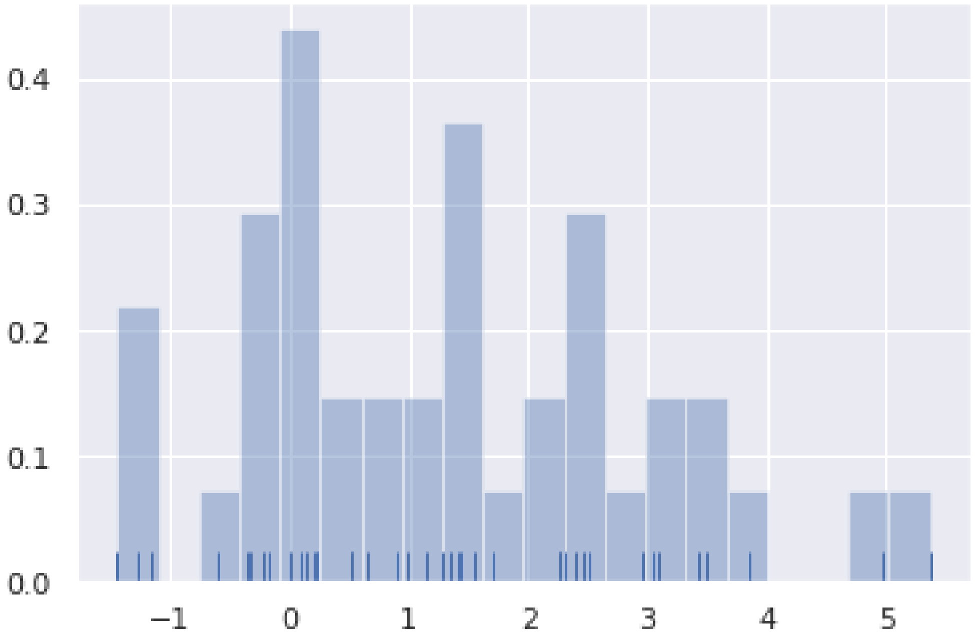
[9]:
```
#5b
sns.distplot(floats, kde=False, rug=True, bins=10, norm_hist=True)
```

[9]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f1dae487630>
```



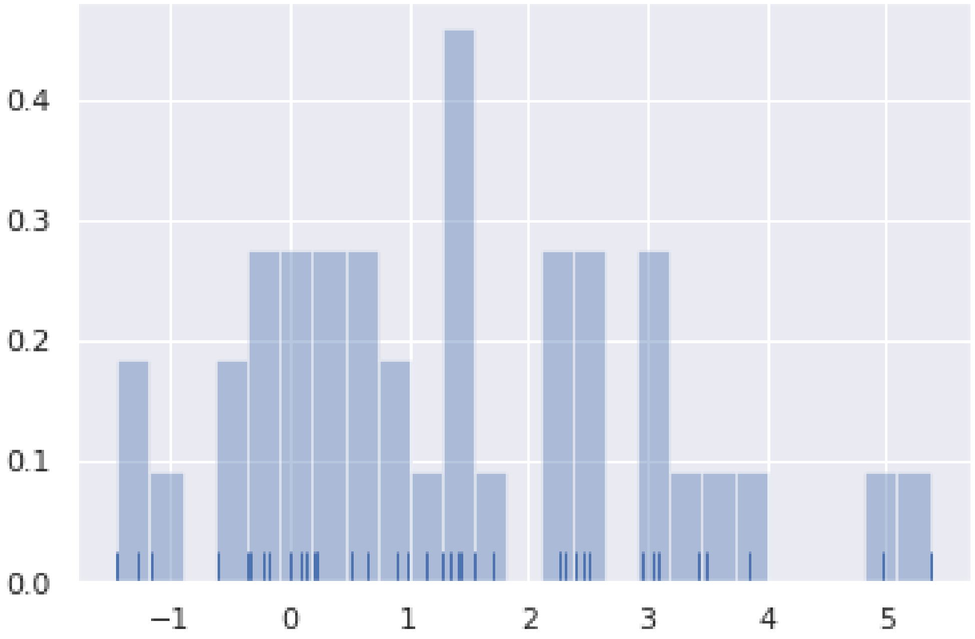
[10]:
```
#5b
sns.distplot(floats, kde=False, rug=True, bins=20, norm_hist=True)
```

[10]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f1dae4f0550>
```



[11]:
```
#5b
sns.distplot(floats, kde=False, rug=True, bins=25, norm_hist=True)
```

[11]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f1dae44dcf8>
```

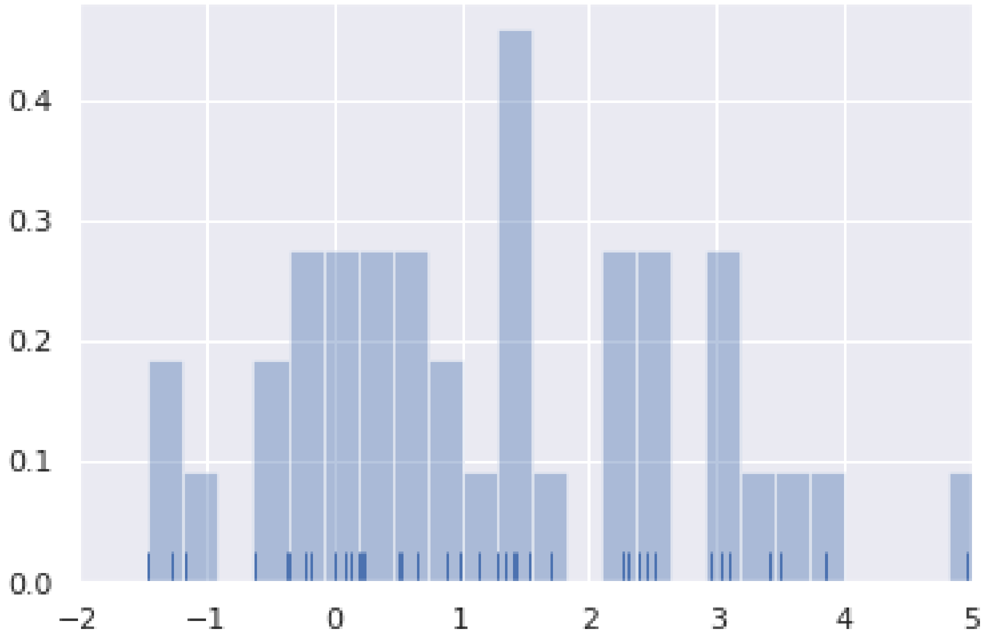


[12]:
```
#5b
"""
Changing the number of bins changes here the range of floats each bucket will␣
 ↪include.
The greater the number of bins, the smaller the range.
"""
```

[12]:
```
'\nChanging the number of bins changes here the range of floats each bucket will
include.\nThe greater the number of bins, the smaller the range.\n'
```

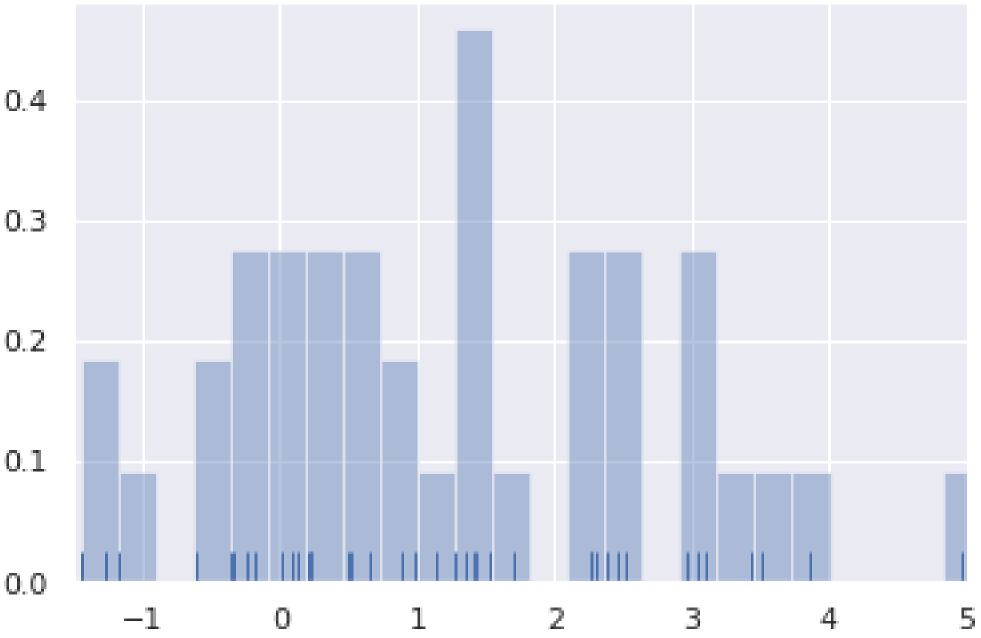
[13]:
```
#5c
hist = sns.distplot(floats, kde=False, rug=True, bins=25, norm_hist=True)
plt.xlim(-2,5)
```

[13]:
```
(-2, 5)
```



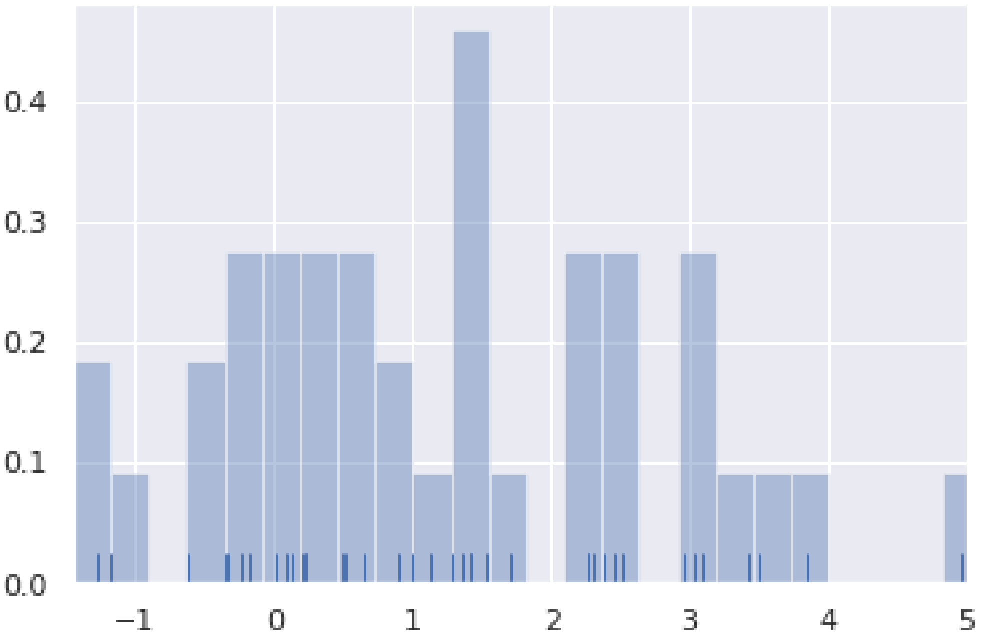
[14]:
```
#5c
hist = sns.distplot(floats, kde=False, rug=True, bins=25, norm_hist=True)
plt.xlim(-1.5,5)
```

[14]:
```
(-1.5, 5)
```



[15]:
```
#5c
hist = sns.distplot(floats, kde=False, rug=True, bins=25, norm_hist=True)
plt.xlim(-1.45,5)
```

[15]:
```
(-1.45, 5)
```



[16]:

```
#5c
"""
It doesn't look like it's actually changed the shape of the data at all, just␣
 ↪the point where the data begins using seaborn.
"""
```

[16]:

```
"\nIt doesn't look like it's actually changed the shape of the data at all, just
the point where the data begins using seaborn.\n"
```