Rakesh Veetekat

Conceptual and theoretical questions

1. For each of parts (a) through (d), indicate whether we would generally expect the performance of a flexible statistical learning method to be better or worse than an inflexible method. Justify your answer.
    a. The sample size n is extremely large, and the number of predictors p is small.

The performance of a flexible statistical learning method in this case would be better than an inflexible method because the flexible method would fit the data better, especially with a larger sample size.

    b. The number of predictors p is extremely large, and the number of observations n is small.

The performance of a flexible statistical learning method in this case would be worse than an inflexible method because the flexible method would overfit since the number of observations is small.

    c. The relationship between the predictors and response is highly non-linear.

The performance of a flexible statistical learning method in this case would be better than an inflexible method because the non-linear relationship means that the flexible model will fit the data better.

    d. The variance of the error terms, i.e. σ2 = Var(), is extremely high.

The performance of a flexible statistical learning method in this case would be worse than an inflexible method because the flexible method would fit to the noise in the data's error terms and increase the variance.

2. Explain whether each scenario is a classification or regression problem, and indicate whether we are most interested in inference or prediction. Finally, provide n and p.
    a. We collect a set of data on the top 500 firms in the US. For each firm we record profit, number of employees, industry and the CEO salary. We are interested in understanding which factors affect CEO salary.

This scenario would be a regression problem and we should be interested in inference to find the relationship between the 3 factors and CEO salary. n = 500 and p = 3.

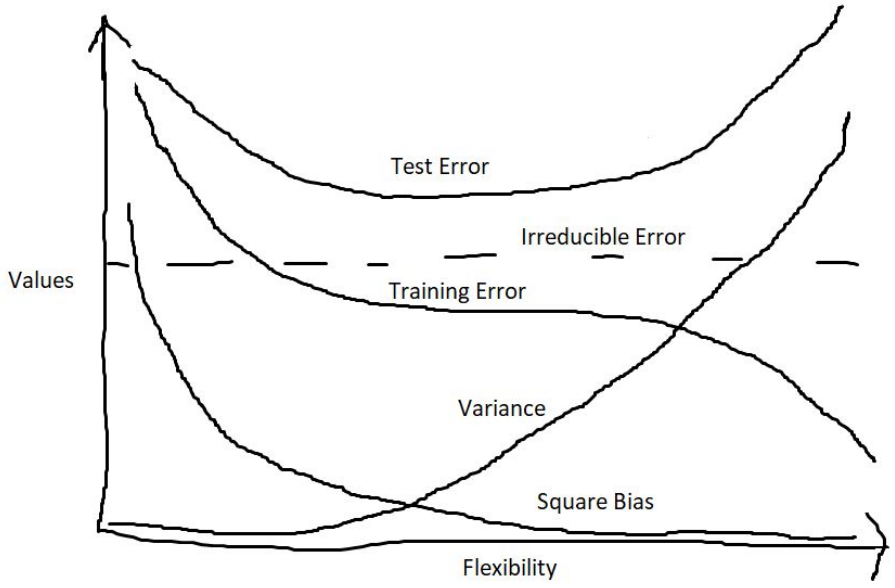
    b. We are considering launching a new product and wish to know whether it will be a success or a failure. We collect data on 20 similar products that were previously launched. For each product we have recorded whether it was a success or failure, price charged for the product, marketing budget, competition price, and ten other variables.

This scenario would be a classification problem and we should be interested in prediction. n = 20 and p = 13.

c. We are interested in predicting the % change in the USD/Euro exchange rate in relation to the weekly changes in the world stock markets. Hence we collect weekly data for all of 2012. For each week we record the % change in the USD/Euro, the % change in the US market, the % change in the British market, and the % change in the German market.

This scenario would be a regression problem and we should be interested in prediction. n = 52 and p = 3.

3. We now revisit the bias-variance decomposition.
   a. Provide a sketch of typical (squared) bias, variance, training error, test error, and Bayes (or irreducible) error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The x-axis should represent the amount of flexibility in the method, and the y-axis should represent the values for each curve. There should be five curves. Make sure to label each one.



   b. Explain why each of the five curves has the shape displayed in part (a).

For “Test Error”, the curve starts out initially by decreasing, then levels out before increasing again. It follows this path because it is underfitting at first, and then overfitting afterwards. For “Irreducible Error”, the error in the model is graphed as the approximation of the noise and it is always constant and greater than 0. For “Training Error”, the curve decreases as flexibility increases and it does this in a monotonic manner. For “Variance”, the curve increases in a monotonic manner as flexibility increases. For “Square Bias”, the curve decreases in a monotonic manner as flexibility increases because less bias is needed to as this happens.

4. You will now think of some real-life applications for statistical learning.

a. Describe three real-life applications in which classification might be useful. Describe the response, as well as the predictors. Is the goal of each application inference or prediction? Explain your answer.

Application 1: Hiring an employee
Response: Hire/Don't hire
Predictors: Resume, Interview results, Fit with company, Potential
Goal: Prediction

Application 2: Product success
Response: Successful/Fails
Predictors: Costs, Revenue, Length of time
Goal: Prediction

Application 3: Contracting coronavirus
Response: Infected/Not infected
Predictors: Location, Age group, Gender, Race, Income
Goal: Prediction

b. Describe three real-life applications in which regression might be useful. Describe the response, as well as the predictors. Is the goal of each application inference or prediction? Explain your answer.

Application 1: Predicting the price of the next generation of iPhones.
Response: Cost of iPhone
Predictors: Last generation sales, Performance of device, Cost to make
Goal: Prediction

Application 2: Determine which factors are related to how many people are going to watch a new movie.
Response: The revenue that the film brings in
Predictors: How many theatres it's being shown in, Genre, Actors, What time of year it comes out?, Length of film
Goal: Inference

Application 3: How many licks does it take to get to the center of a tootsie pop?
Response: Number of licks
Predictors: Radius of lollipop, Ingredients, Temperature
Goal: Prediction

c. Describe three real-life applications in which cluster analysis might be useful.

Application 1: Recommendation system on game consoles to recommend popular titles within different genres based on how the user has been clustered with other users.
Application 2: Cluster users that are listening to music and recommend similar genres to users in the same cluster.

Application 3: Recommend movies and tv shows to users on Netflix by clustering them with other users based on their watch history.

5. What are the advantages and disadvantages of a very flexible (versus a less flexible) approach for regression or classification? Under what circumstances might a more flexible approach be preferred to a less flexible approach? When might a less flexible approach be preferred?

Advantages:

- A flexible approach decreases bias
- May fit non-linear relationships better
- Does not require as many assumptions as less flexible approach

Disadvantages:

- A flexible approach increases variance
- Needs to estimate more parameters than a less flexible approach
- The results are harder to interpret

A flexible approach would be preferred when there is a large amount of data and variables that need to be used. Non-linear relationships also are better modeled using a flexible approach rather than a less flexible one.

A less flexible approach would be preferred when we would like to make an inference that can be interpreted more easily than if the approach was more flexible. If the relationships do not appear to require a flexible approach, then the less flexible approach would be more ideal.

6. Describe the differences between a parametric and a non-parametric statistical learning approach. What are the advantages of a parametric approach to regression or classification (as opposed to a nonparametric approach)? What are its disadvantages?

A parametric statistical learning approach is different from a non-parametric approach because it assumes the shape of f, and so it does not require a very large dataset to estimate f, while the non-parametric approach does not make this assumption and requires a large sample to estimate f.

The advantages of the parametric approach are that it simplifies f so that it can be modeled using few parameters and the computation time is significantly reduced as a result. However, this can make it inaccurate in some cases if the assumptions of f are incorrect or if more flexible models are being used because that can cause overfitting.

7. The table below provides a training data set containing six observations, three predictors, and one qualitative response variable. Suppose we wish to use this data set to make a prediction for Y when X1 = X2 = X3 = 0 using K-nearest neighbors.

Obs.	X_1	X_2	X_3	Y
1	0	3	0	Red
2	2	0	0	Red
3	0	1	3	Red
4	0	1	2	Green
5	−1	0	1	Green
6	1	1	1	Red

a. Compute the Euclidean distance between each observation and the test point, X1 = X2 = X3 = 0.

Observation	Distance
1	3
2	2
3	3.1622777
4	2.236068
5	1.4142136
6	1.7320508

b. What is our prediction with K = 1? Why?

If K = 1, then the nearest neighbor is observation 5. The color associated with observation 5 is Green so the prediction should be Green.

c. What is our prediction with K = 3? Why?

If K = 3, then the nearest neighbors are observations 2, 5, and 6. Observations 2 and 6 are associated with Red, and observation 5 is associated with Green, and since there is a higher chance of Red in the set of neighbors, the prediction should be Red.

d. If the Bayes decision boundary in this problem is highly nonlinear, then would we expect the best value for K to be large or small? Why?

We would expect the best value for K to be small because the boundary becomes linear as K grows.

Applied

8. This exercise involves the Auto data set studied in the lab. Make sure that the missing values have been removed from the data.
    a. Which of the predictors are quantitative, and which are qualitative?

Quantitative predictors:

- mpg
- cylinders
- displacement
- horsepower
- weight
- acceleration
- Year

Qualitative predictors:

- origin
- name

b. What is the range of each quantitative predictor? You can answer this using the range() function.

MAX for each variable:

mpg	46.6
cylinders	8
displacement	455
horsepower	230
weight	5140
acceleration	24.8
year	82
origin	3
name	vw rabbit custom

MIN for each variable:

mpg	9
cylinders	3
displacement	68
horsepower	46
weight	1613
acceleration	8
year	70
origin	1
name	amc ambassador brougham

Range for each variable:

mpg	[9, 46.6]
cylinders	[3, 8]
displacement	[68, 455]
horsepower	[46, 230]
weight	[1613, 5140]
acceleration	[8, 24.8]
year	[70, 82]

origin [1, 3]
name [amc ambassador brougham, vw rabbit custom]

c. What is the mean and standard deviation of each quantitative predictor?

Mean for each quantitative predictor

Standard deviation for each quantitative predictor

mpg 7.805007
cylinders 1.705783
displacement 104.644004
horsepower 38.491160
weight 849.402560
acceleration 2.758864
year 3.683737
origin 0.805518

mpg 23.445918
cylinders 5.471939
displacement 194.411990
horsepower 104.469388
weight 2977.584184
acceleration 15.541327
year 75.979592
origin 1.576531

d. Now remove the 10th through 85th observations. What is the range, mean, and standard deviation of each predictor in the subset of the data that remains?

The range, mean, and standard deviation of each predictor can be found in the image below.

	mpg	cylinders	displacement	horsepower	weight	acceleration	year	origin
count	316.000000	316.000000	316.000000	316.000000	316.000000	316.000000	316.000000	316.000000
mean	24.404430	5.373418	187.240506	100.721519	2935.971519	15.726899	77.145570	1.601266
std	7.867283	1.654179	99.678367	35.708853	811.300208	2.693721	3.106217	0.819910
min	11.000000	3.000000	68.000000	46.000000	1649.000000	8.500000	70.000000	1.000000
25%	18.000000	4.000000	100.250000	75.000000	2213.750000	14.000000	75.000000	1.000000
50%	23.950000	4.000000	145.500000	90.000000	2792.500000	15.500000	77.000000	1.000000
75%	30.550000	6.000000	250.000000	115.000000	3508.000000	17.300000	80.000000	2.000000
max	46.600000	8.000000	455.000000	230.000000	4997.000000	24.800000	82.000000	3.000000

e. Using the full data set, investigate the predictors graphically, using scatterplots or other tools of your choice. Create some plots highlighting the relationships among the predictors. Comment on your findings.

This table shows the correlation value that each predictor has with the other predictors. The main diagonal elements can be dismissed because those values will always be 1.

	mpg	cylinders	displacement	horsepower	weight	acceleration	year	origin
mpg	1.000000	-0.777618	-0.805127	-0.778427	-0.832244	0.423329	0.580541	0.565209
cylinders	-0.777618	1.000000	0.950823	0.842983	0.897527	-0.504683	-0.345647	-0.568932
displacement	-0.805127	0.950823	1.000000	0.897257	0.932994	-0.543800	-0.369855	-0.614535
horsepower	-0.778427	0.842983	0.897257	1.000000	0.864538	-0.689196	-0.416361	-0.455171
weight	-0.832244	0.897527	0.932994	0.864538	1.000000	-0.416839	-0.309120	-0.585005
acceleration	0.423329	-0.504683	-0.543800	-0.689196	-0.416839	1.000000	0.290316	0.212746
year	0.580541	-0.345647	-0.369855	-0.416361	-0.309120	0.290316	1.000000	0.181528
origin	0.565209	-0.568932	-0.614535	-0.455171	-0.585005	0.212746	0.181528	1.000000

The predictors with the highest correlation values in the table are:
Displacement and cylinders: 0.950823
Displacement and weight: 0.932994
Cylinders and weight: 0.897527
Displacement and horsepower: 0.897257
Weight and horsepower: 0.864538

f. Suppose that we wish to predict gas mileage (mpg) on the basis of the other variables. Do your plots suggest that any of the other variables might be useful in predicting mpg? Justify your answer

Weight, displacement, horsepower, and cylinders could be useful in predicting mpg because they all have fairly high correlation values with mpg, although all three are negative. Even though these correlations are negative, they are just as useful as positive values for predicting mpg.

9. This exercise involves the Boston housing data set.
   a. To begin, load in the Boston data set. The Boston data set is part of the MASS library in R.

```
> library(MASS)
```

Now the data set is contained in the object `Boston`.

```
> Boston
```

Read about the data set:

```
> ?Boston
```

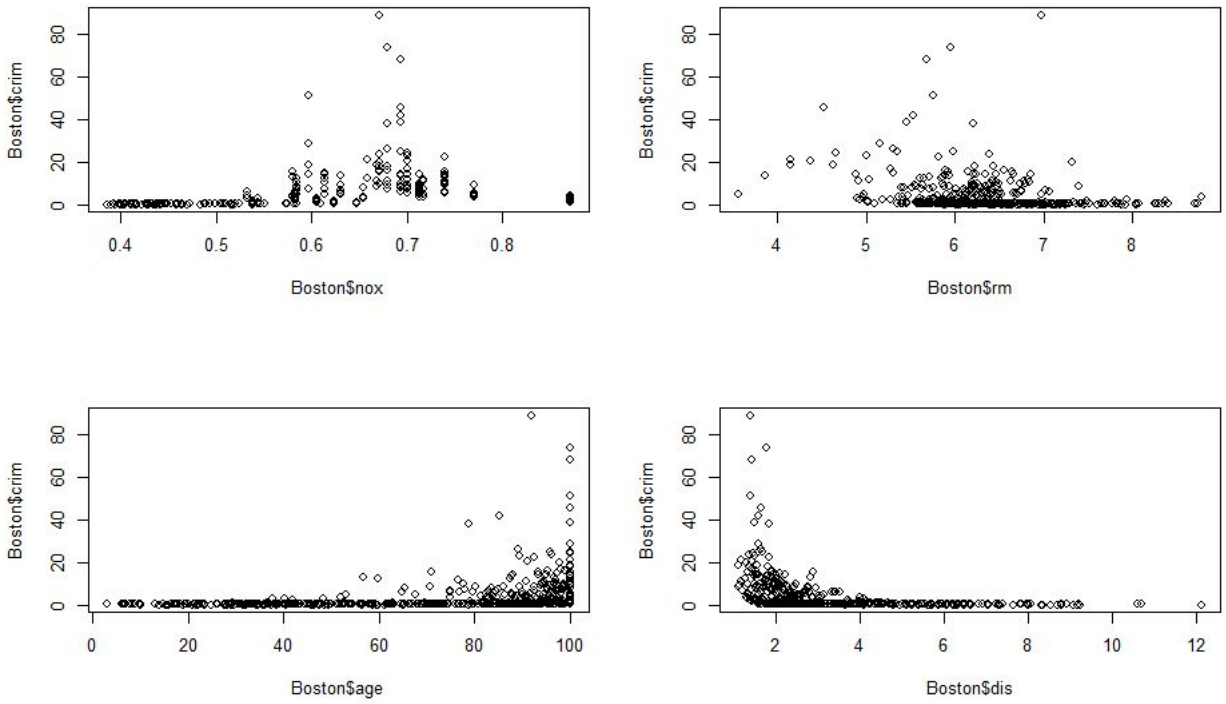
How many rows are in this data set? How many columns? What do the rows and columns represent?

There are 506 rows in the data set. 14 columns.
The rows represent the suburbs of Boston.
The columns represent the following:

- CRIM - per capita crime rate by town
- ZN - proportion of residential land zoned for lots over 25,000 sq.ft.
- INDUS - proportion of non-retail business acres per town.
- CHAS - Charles River dummy variable (1 if tract bounds river; 0 otherwise)
- NOX - nitric oxides concentration (parts per 10 million)

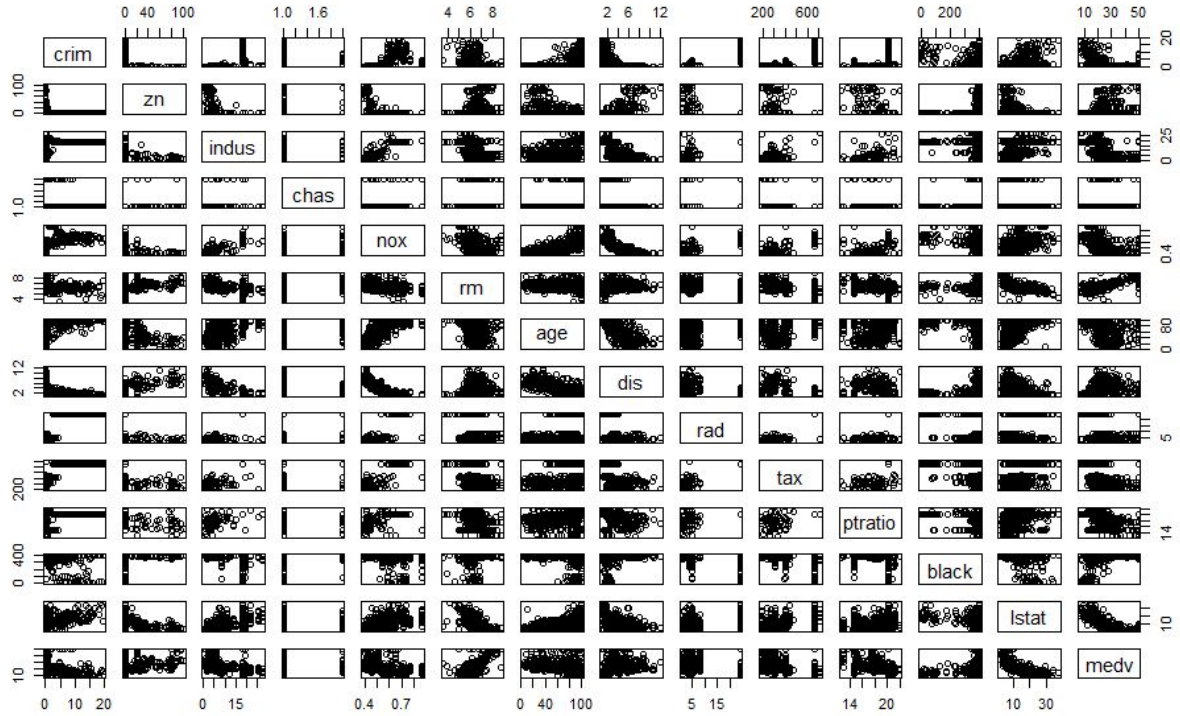
- RM - average number of rooms per dwelling
- AGE - proportion of owner-occupied units built prior to 1940
- DIS - weighted distances to five Boston employment centres
- RAD - index of accessibility to radial highways
- TAX - full-value property-tax rate per $10,000
- PTRATIO - pupil-teacher ratio by town
- B - 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town
- LSTAT - % lower status of the population
- MEDV - Median value of owner-occupied homes in $1000's

b. Make some pairwise scatterplots of the predictors (columns) in this data set. Describe your findings.



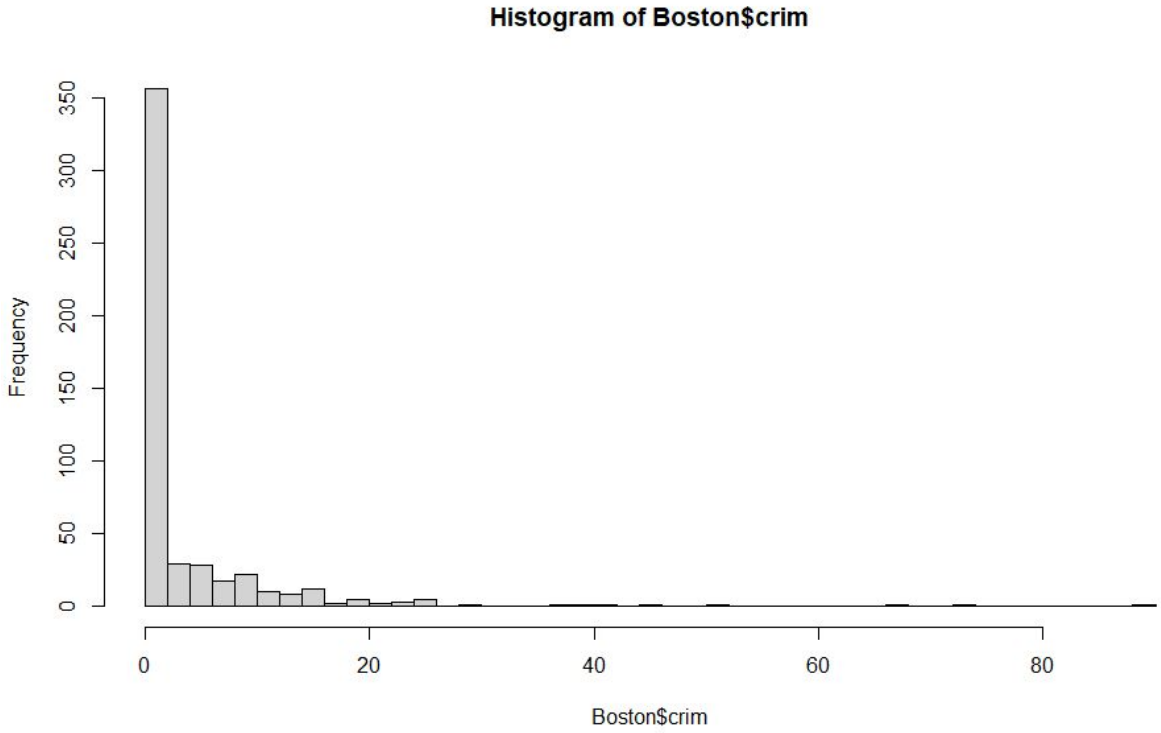
From the scatterplots shown above, we can see that crime appears to spike with these four predictors at different points. When the nitric oxides concentration is between 0.6 and 0.7, the per capita crime rate appears to spike in several towns, and the same happens when the proportion of occupied homes built prior to 1940 is greater than 80.

c. Are any of the predictors associated with per capita crime rate? If so, explain the relationship.

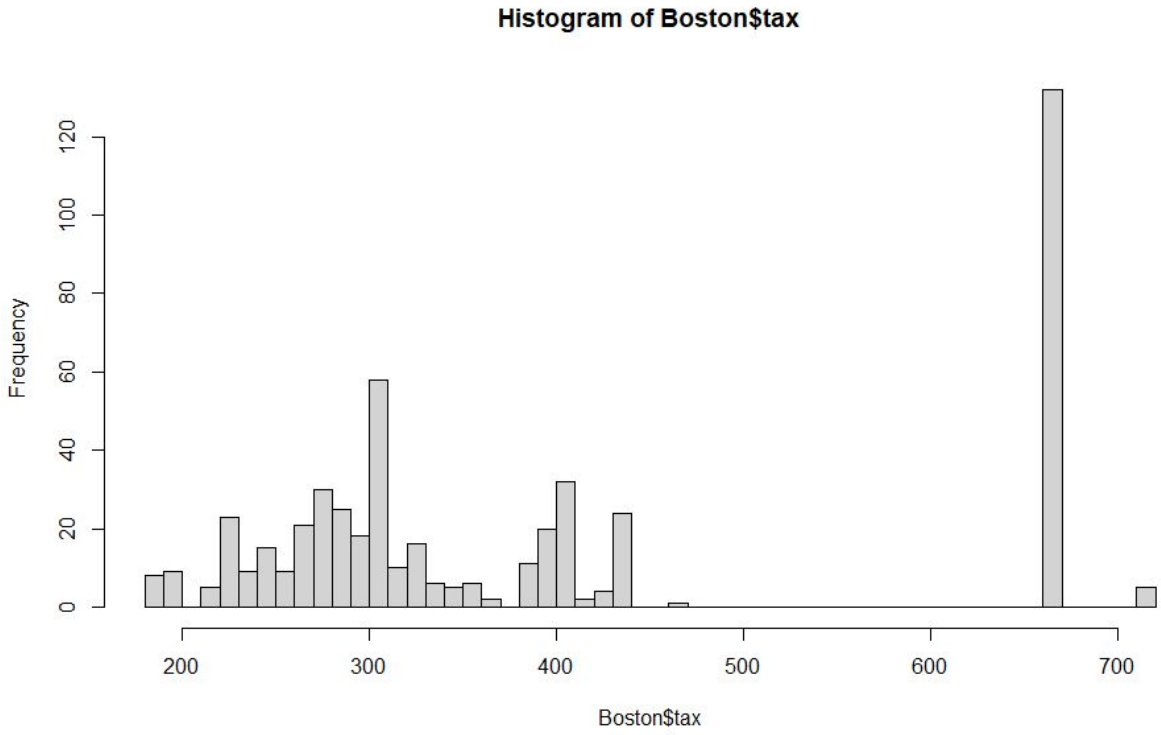


From the first column of graphs that show the correlation between “crim” and the other predictors, “zn”, “nox”, “age”, and “dis” appear to be the predictors that can associate with the “crim” predictor.

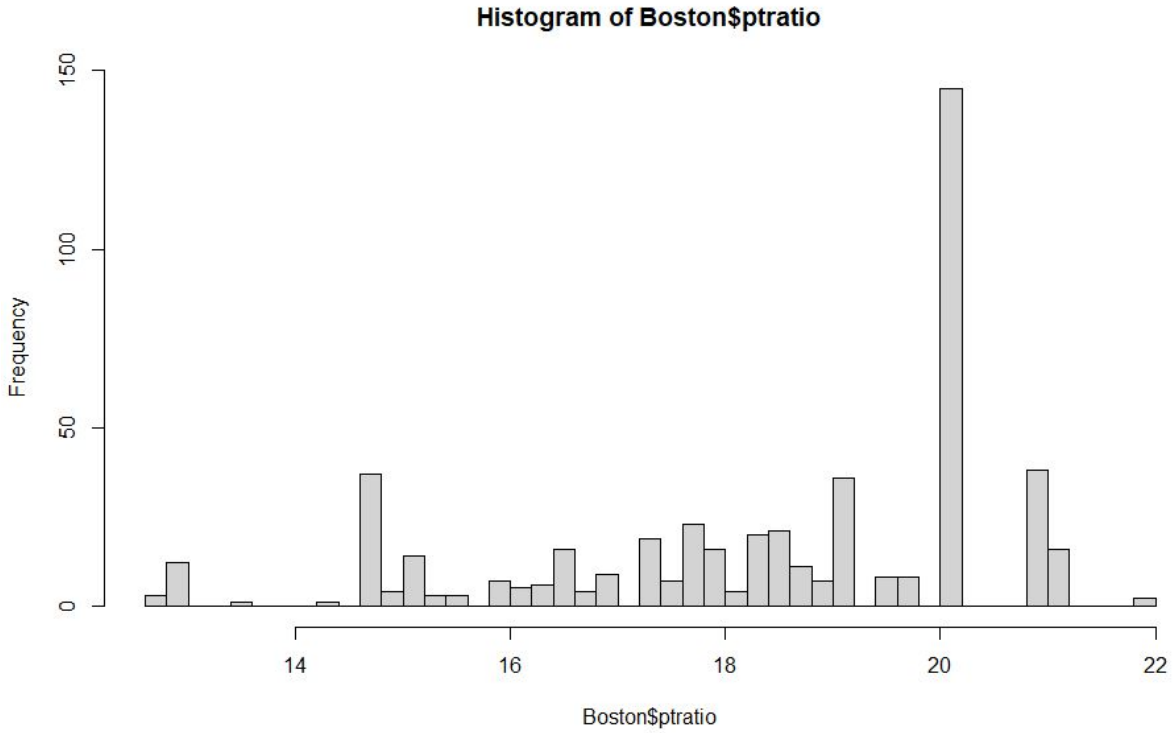
d. Do any of the suburbs of Boston appear to have particularly high crime rates? Tax rates? Pupil-teacher ratios? Comment on the range of each predictor.



The suburbs in Boston do not appear to have high crime rates since the majority of them have per capita crime rates below 10.



There appear to be a sizeable amount of towns that pay a high average tax, while the rest appear to pay between the 200 to 400 range.



There seems to be more towns in this dataset that have a higher ptratio than 16.

e. How many of the suburbs in this data set bound the Charles river?

35.

f. What is the median pupil-teacher ratio among the towns in this data set?

19.05

g. Which suburb of Boston has the lowest median value of owner occupied homes? What are the values of the other predictors for that suburb, and how do those values compare to the overall ranges for those predictors? Comment on your findings.

Suburb 5 appears to have the lowest median value.
The values of the other predictors in this suburb are shown below:

	crim	zn	indus	chas	nox	rm	age	dis	rad	tax	ptratio	black	lstat	medv
5	0.06905	0	2.18	0	0.458	7.147	54.2	6.0622	3	222	18.7	396.9	5.33	36.2

h. In this data set, how many of the suburbs average more than seven rooms per dwelling? More than eight rooms per dwelling? Comment on the suburbs that average more than eight rooms per dwelling.

64 suburbs average more than seven rooms per dwelling and 13 suburbs average more than eight rooms per dwelling.

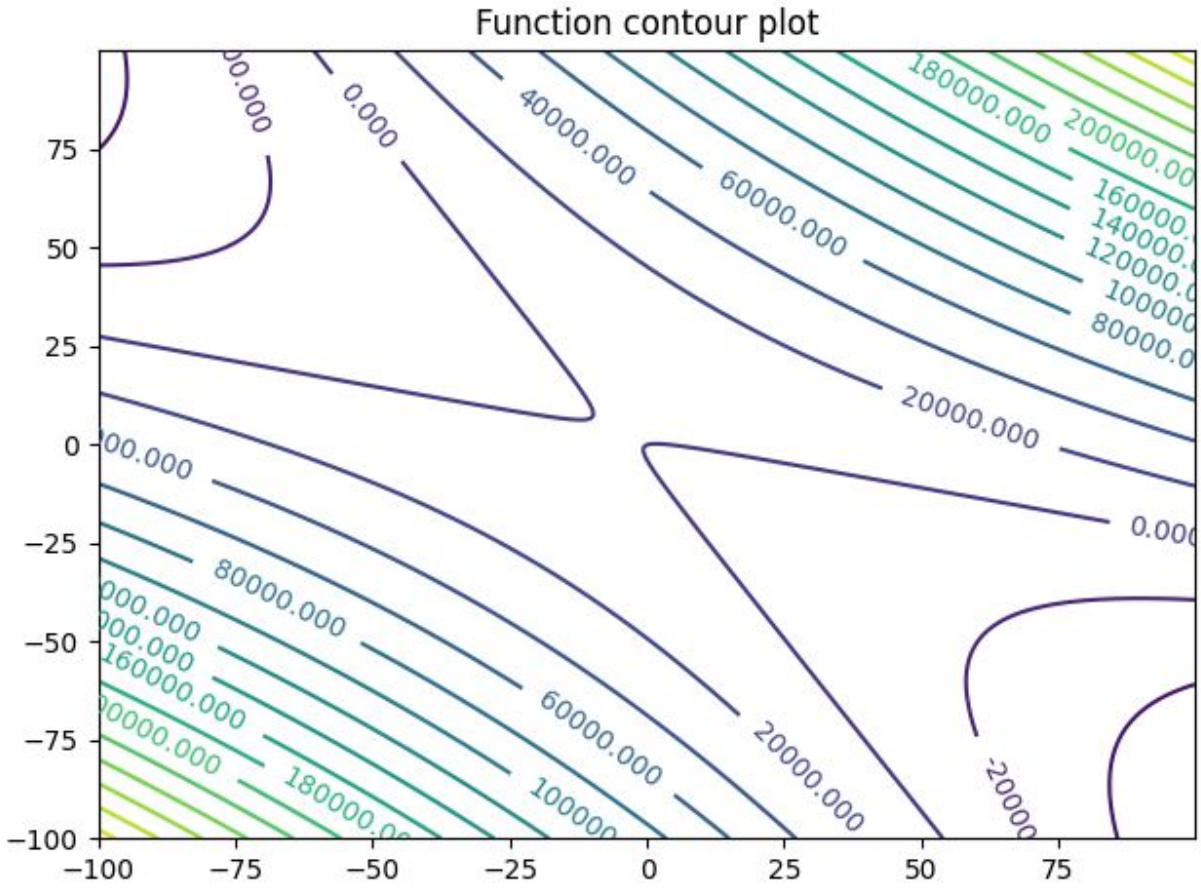
Mathematics and Probability

10. Minimum and Maximum of a function (10)

For the following function

$$f(x, y) = 4x^2 + 9y^2 - 16x + 36y + 18xy + 5$$

a. Show the contour plot



b. Find the partial derivative with respect to x and y

$$f(x, y) = 4\,x^2 + 9\,y^2 - 16\,x + 36\,y + 18\,x\,y + 5$$

df/dx = 8x + 0 - 16 + 0 + 18y
df/dx = 8x + 18y - 16

df/dy = 0 + 18y - 0 + 36 + 18x

df/dy = 18x + 18y + 36

c. Find the minimum point

Set both df/dx and df/dy to 0, and find the intersection between both equations.
8x + 18y - 16 = 18x + 18y + 36
0 = -10x - 52
x = -52/10 = -26/5
y = 16/5

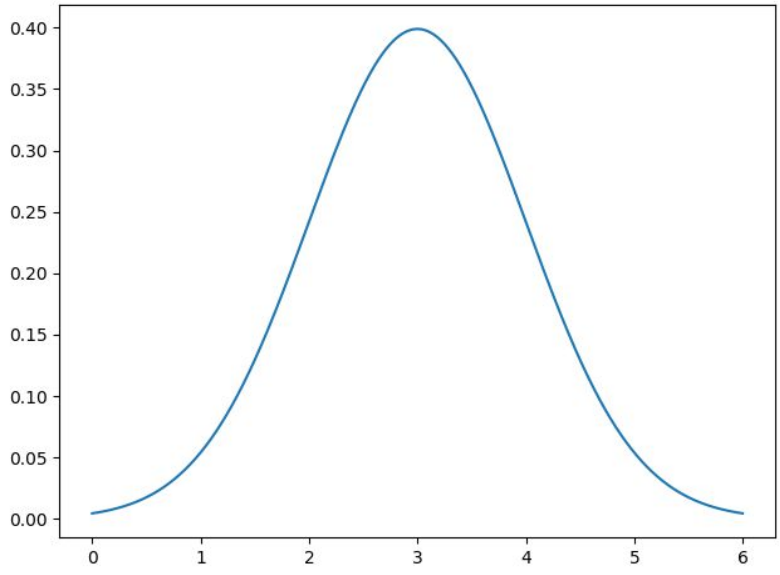
$d^2f/dx^2 = 8$
$d^2f/dy^2 = 18$
Since the second order partial derivatives for x and y are both positive, this one critical point must be a minimum.

So, the minimum point is at (-26/5, 16/5).

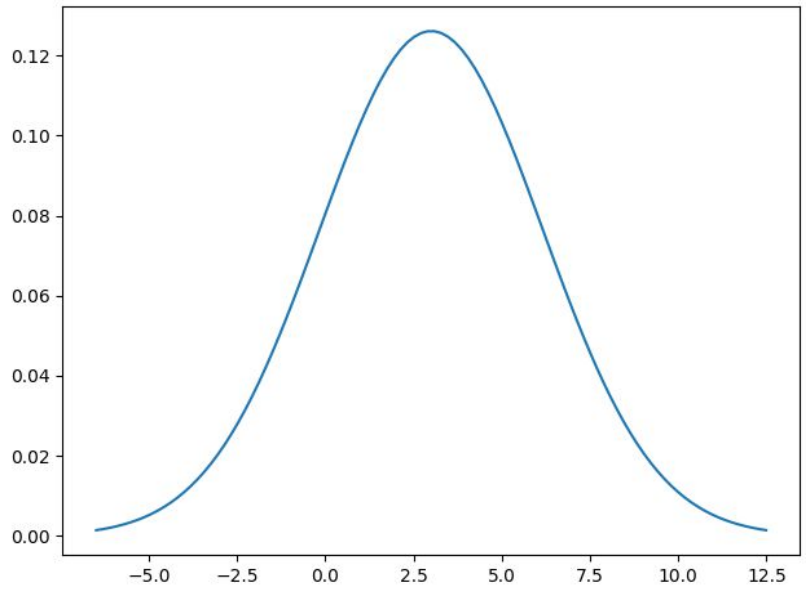
11. For the normal distribution with mean m and variance σ; its pdf is defined by

$$f(x) = \frac{1}{\sqrt{2\pi}\,\sigma}\, e^{-\frac{(x-m)^2}{2\,\sigma^2}}$$

a. Plot the curve for m = 3 and σ = 1



b. Plot the curve for m = 3 and σ = √10



c. Let's assume, we have N samples - {x1, x2, … . , xN}. The likelihood function is defined by

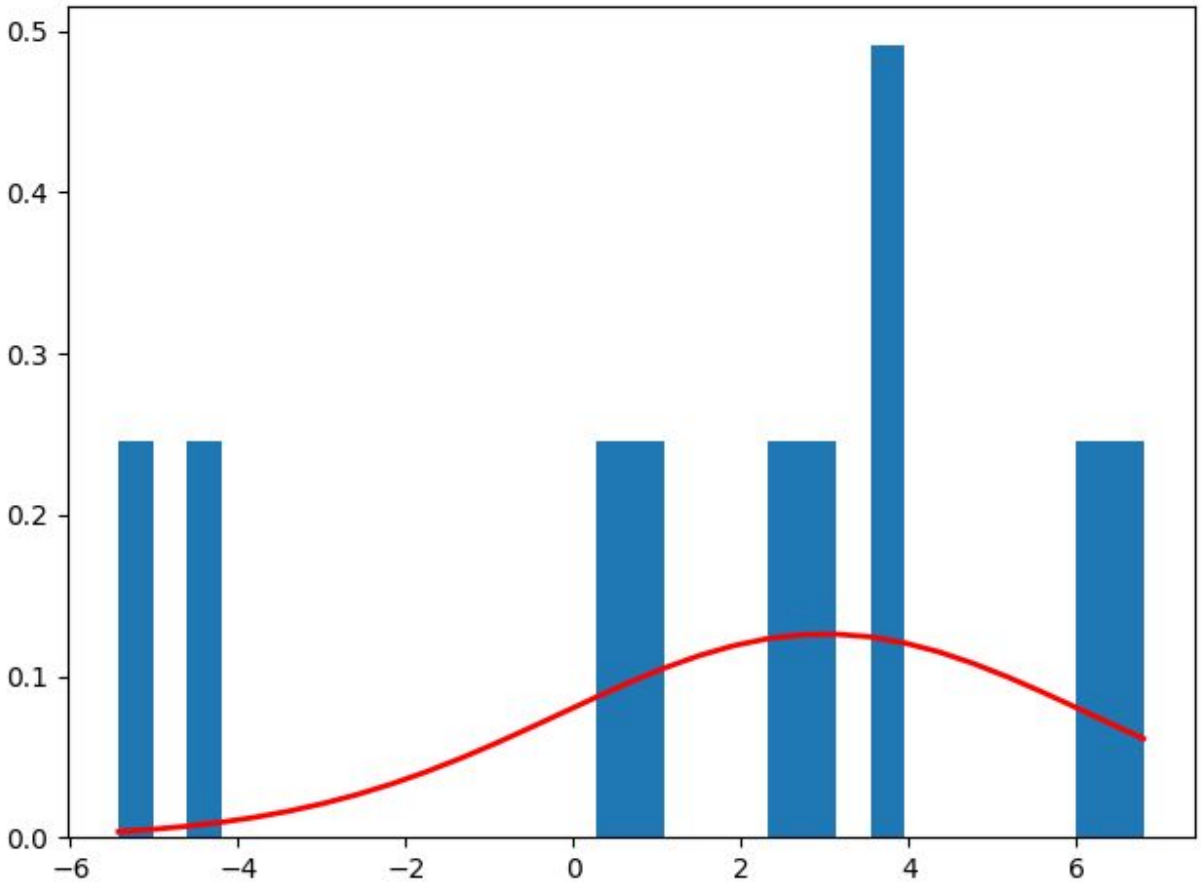
$$L(x_1, x_2, \dots, x_N; m, \sigma) = \prod_{i=1}^{N} f(x_i) = f(x_1) \times f(x_2) \times \cdots \times f(x_N)$$

find the MLE for m and σ.

$$m = \frac{1}{n}\sum_{i=1}^{n} X_i$$

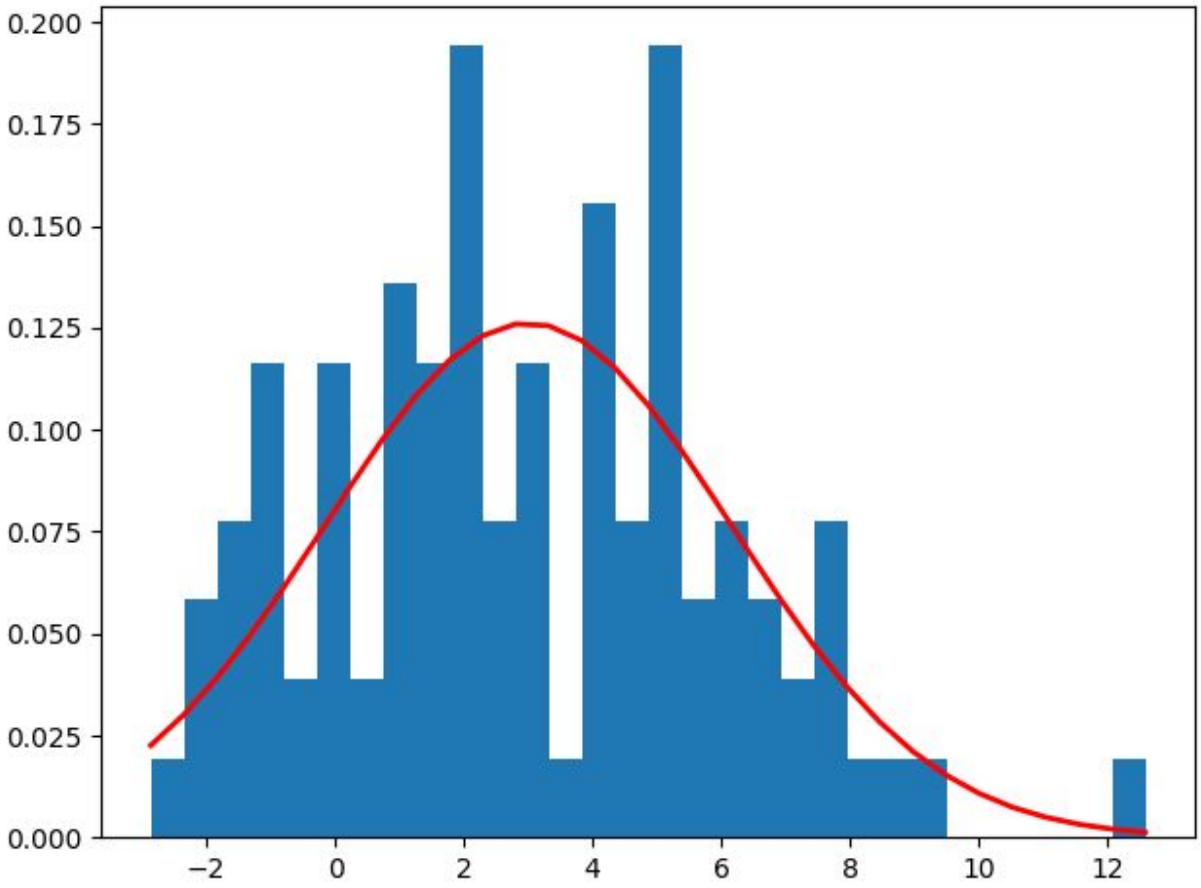
$$\sigma = \frac{1}{n}\sum_{i=1}^{n} (x_i - m)^2$$

d. Draw 10 samples from a normal distribution with m = 3 and σ = √10. Show its histogram, and calculate it's mean and variance



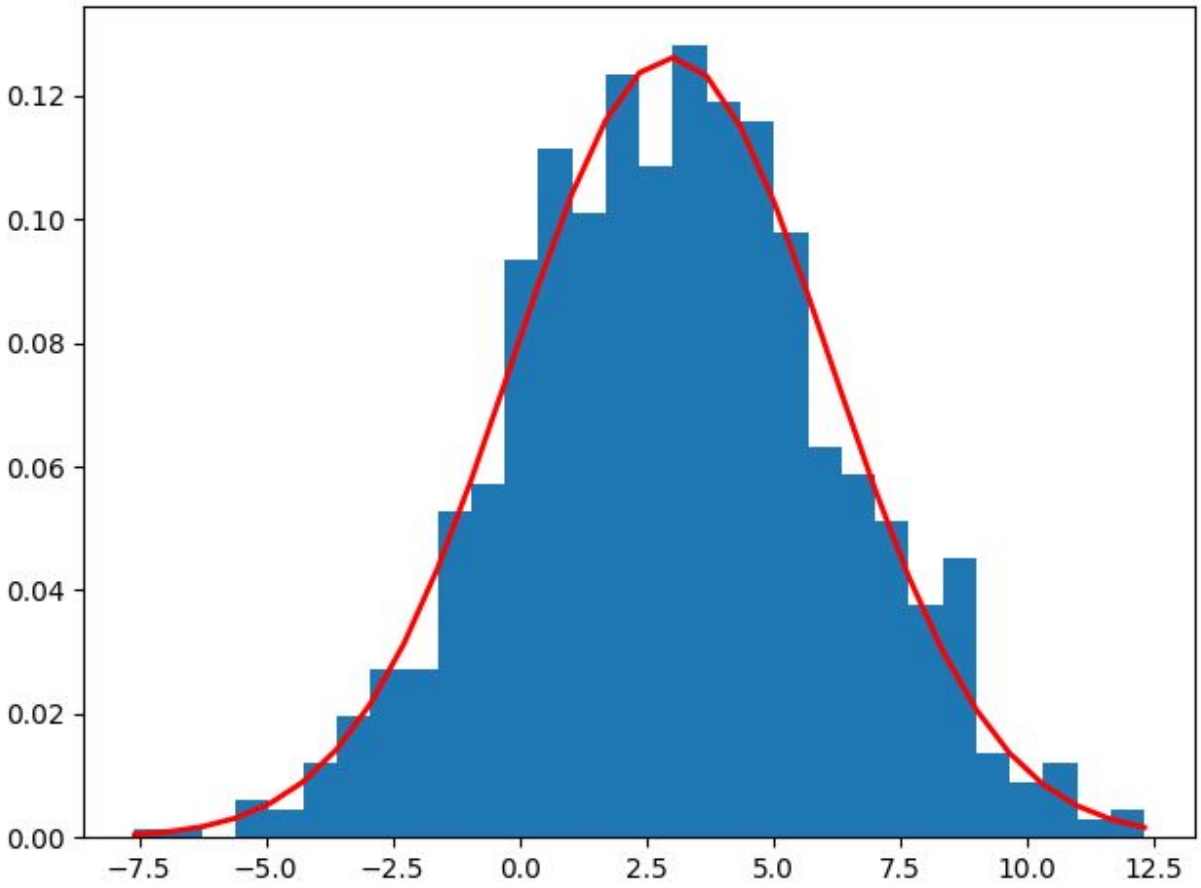
Mean of histogram: 1.8355038143829834
Variance of histogram: 16.711230436088684

e. Draw 100 samples from a normal distribution with m = 3 and σ = √10. Show its histogram, and calculate it's mean and variance



Mean of histogram: 2.9504055310335446
Variance of histogram: 9.330108018708456

f. Draw 1000 samples from a normal distribution with m = 3 and σ = √10. Show its histogram, and calculate it's mean and variance



Mean of histogram: 3.061011890916824
Variance of histogram: 10.214792186545374

12. For a company, we have collected the following information for their hiring process over the last 10 years.

	Education	Ph.D.	Engineering	Ph.D. in Engineering
Accepted		10	25	45
Rejected		90	125	55

a. What is the probability of an applicant to have PhD in Engineering?

P(PhD in Engineering) = (45 + 55) / (10 + 25 + 45 + 90 + 125 + 55)
P(PhD in Engineering) = 100 / 350

b. What is probability of being accepted if you have an Engineering background?

P(accepted | Engineering) = P(Engineering | accepted) * P(accepted) / P(Engineering)
P(accepted | Engineering) = (70/80) * (80/350) / (250/350)
P(accepted | Engineering) = 70 / 250

c. What is probability of being accepted?

P(accepted) = (Total accepted) / (Total applicants)
P(accepted) = (10 + 25 + 45) / 350
P(accepted) = 80 / 350

d. What is probability of having Ph.D. if the candidate being accepted?

P(PhD | accepted) = P(accepted | PhD) * P(PhD) / P(accepted)
P(PhD | accepted) = (55/200) * (200/350) / (80/350)
P(PhD | accepted) = 55 / 80