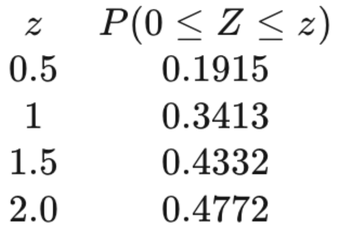
Data analysis and Visualization

1. There are a total of 400 students in a secondary school. Their heights obey a normal distribution with mean 169cm and standard deviation 4 cm. What is the approximate number of students whose heights are below 163 cm or above 171 cm?

   Note: Use the standard normal distribution table below, where Z has mean μ = 0 and standard deviation σ = 1 **(#MEDIUM #MCQ)**



z	$P(0 \leq Z \leq z)$
0.5	0.1915
1	0.3413
1.5	0.4332
2.0	0.4772

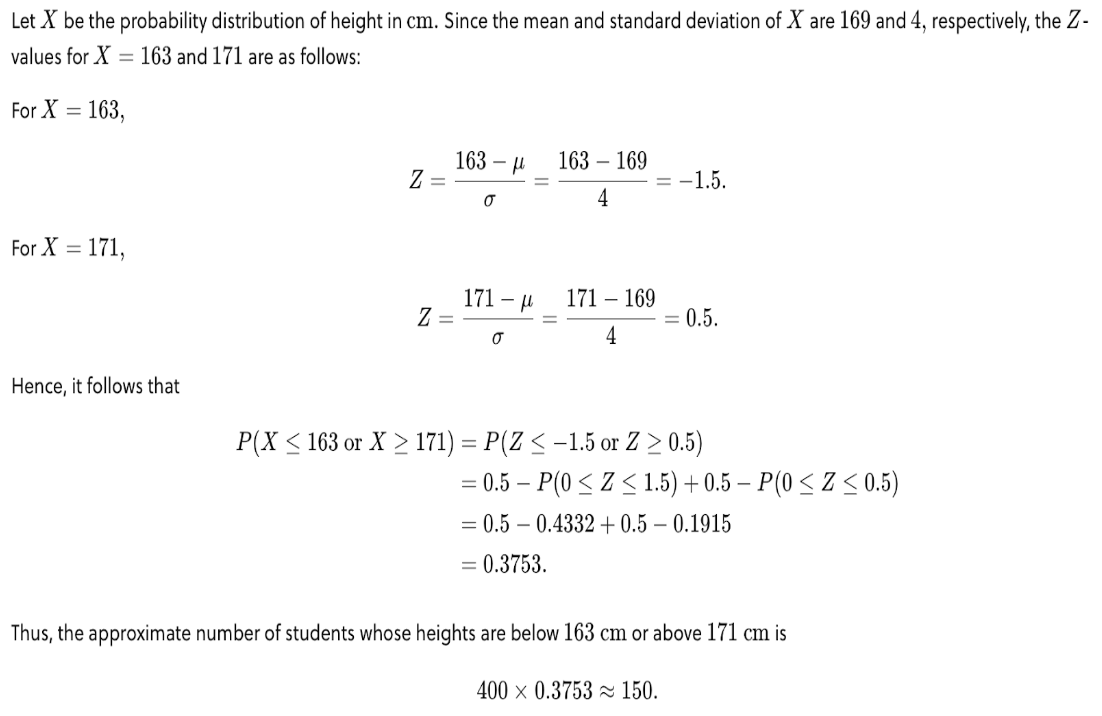
a. 146

b. 148

c. 150

d. 152

Solution:



Let X be the probability distribution of height in cm. Since the mean and standard deviation of X are 169 and 4, respectively, the Z-values for $X = 163$ and 171 are as follows:

For $X = 163$,

$$Z = \frac{163 - \mu}{\sigma} = \frac{163 - 169}{4} = -1.5.$$

For $X = 171$,

$$Z = \frac{171 - \mu}{\sigma} = \frac{171 - 169}{4} = 0.5.$$

Hence, it follows that

$$\begin{aligned} P(X \leq 163 \text{ or } X \geq 171) &= P(Z \leq -1.5 \text{ or } Z \geq 0.5) \\ &= 0.5 - P(0 \leq Z \leq 1.5) + 0.5 - P(0 \leq Z \leq 0.5) \\ &= 0.5 - 0.4332 + 0.5 - 0.1915 \\ &= 0.3753. \end{aligned}$$

Thus, the approximate number of students whose heights are below 163 cm or above 171 cm is

$$400 \times 0.3753 \approx 150.$$

2. Suppose a researcher wanted to investigate the relationship between car color and speeding violations, with the null hypothesis that there is no relationship between car color and number of speeding violations.

   The researcher tested 20 different car colors individually, with each test using a p-value of 0.05. The researcher only found evidence that a bright yellow car was connected to an increase in speeding violations.

   What is the best explanation for this? **(#MEDIUM #MCQ)**

   a. Bright yellow cars definitely result in an increase in speeding violations.
   b. Because the researcher tested so many colors individually, chances are one of them would be significant due to random variation.
   c. Car color is a significant factor for the number of speeding violations.
   d. None of the above

   Solution:

   This set of experiments provides no statistical evidence for a relationship between car color and speeding violations. Over the course of 20 different experiments, we should expect one to display a relationship with a p-value of 0.05 simply due to random fluctuations in the data.

   In fact, that is what a p-value of 0.05 means! It means that if the null hypothesis is true, we would expect to see results like this in 5% of the trials. Since that's exactly what has happened in this set of experiments, we have absolutely no justification for rejecting the null hypothesis.

   Note that it is precisely for this reason that p-values are most useful when they are used in the context of a single question applied to a single set of data. To ask multiple questions of the same data (or, as in the question above, the same question of multiple sets of data) is known as p-hacking and, if done often enough, guarantees us "statistically significant" results even if there is no real relationship to find.

3. What is the interquartile range of the list of positive integers from 1 to 27 inclusive? **(#MEDIUM #MCQ)**

    a. 1
    b. 7
    c. 14
    d. 21

    Solution:
    First, we find the median of the list. The median is (1 + 27)/2 = 14
    Next, out of the list of numbers from 15 to 27, we find the median which is (15 + 27) / 2 = 21 This is the upper quartile.
    Next, out of the list of numbers from 1 to 13, we find the median which is (1 + 13)/2 = 7 this is lower quartile
    Hence, the interquartile range is 21-7 = 14

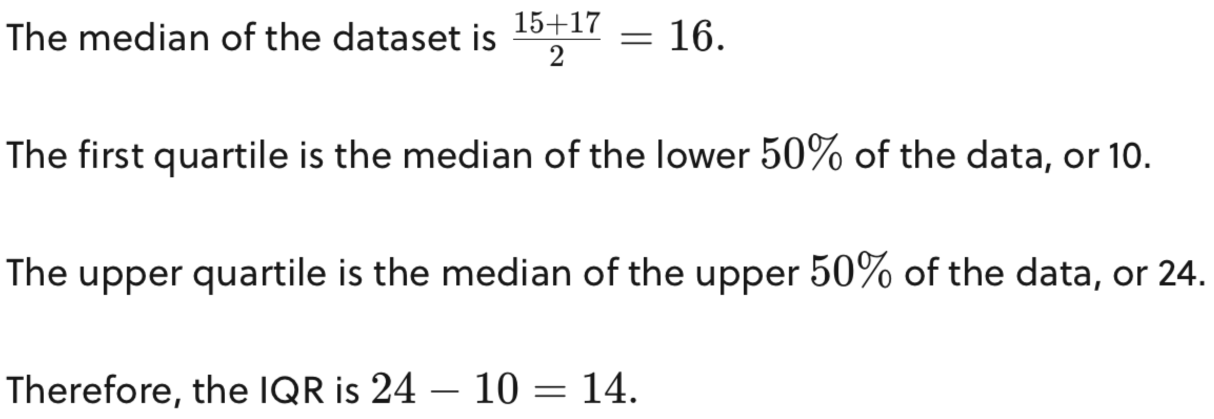
4. We can use the IQR to identify outliers in a dataset. An outlier in a dataset has a value that is significantly larger or smaller than most of the rest of the set. For example, in the set {1, 45, 50, 52, 57, 61}, the value 1 is an outlier because it is significantly lower than the other numbers. We can use the IQR for a common mathematical definition of an outlier. An outlier is any data point that is:
    Smaller than (First Quartile) - (1.5*IQR)
    Larger than (Third Quartile) + (1.5* IQR)

    The data set is shown below. Which of the data points are outliers using the method above?

5, 6, 10, 11, 15, 17, 20, 24, 46, 47 **(#MEDIUM #MSQ)**

a. 5
b. 6
c. 46
d. 47

Solution:



The median of the dataset is $\frac{15+17}{2} = 16$.

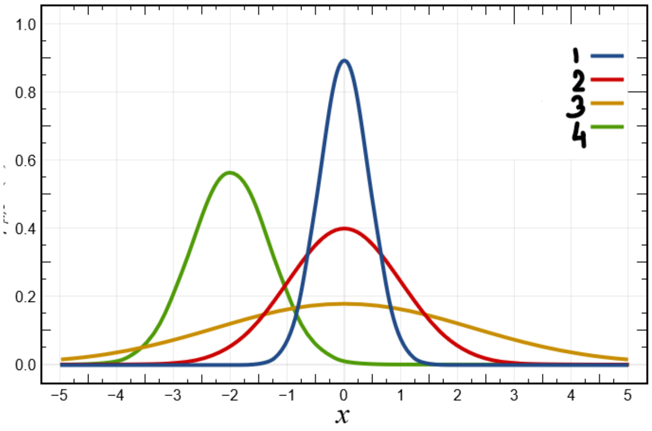
The first quartile is the median of the lower 50% of the data, or 10.

The upper quartile is the median of the upper 50% of the data, or 24.

Therefore, the IQR is $24 - 10 = 14$.

Let's begin by determining that IQR *1.5 = 14 * 1.5 = 21
Lower outliers will be values below 10 - 21 = -11 Upper outliers will be values above 24 + 21 = 45
Therefore, this dataset has two outliers: 46 and 47.

5. Consider the image given below:



Given these four normally distributed curves for four different random variables, which of the following holds true **(#MEDIUM #MCQ)**

a. The variance is maximum for 1 and is minimum for 3.
b. The standard deviation is maximum for 2 and minimum for 3.
c. The standard deviation is maximum for 3 and minimum for 1
d. The variance is maximum for 2.

Solution:

Variance is the spread of the data. Hence, the wider and lesser in height the normal curve is the more will be variance. From the above curves, we can clearly see that height of 3 is the minimum and the height of 1 is the maximum. Hence, the variance of 3 is maximum, and a variance of 1 is minimum. Variance is the square of standard deviation, hence whatever holds true for variance also holds true for standard deviation.

6. which of the following statement(s) is/are correct **(#EASY #MSQ)**

a. TSNE is a non-convex optimization.
b. We can produce similar results as PCA using TSNE with perplexity=number of data points
c. Using the P-value approach the comparison "P-values < alpha" reaches the conclusion: "reject H0" (reject the null hypothesis)
d. We can use Box-cox transformations to check if a random variable is coming from Normal distribution

Solution:

7. Which of the following statement(s) is/are true? **(#EASY #MCQ)**

a. Kurtosis helps us to measure the peakedness of the given distribution of the data
b. Kurtosis tells us if there are any outliers present in the data
c. The Q-Q plot between a log-normal distribution and a normal distribution should be a straight line with a slope approximately equal to 1.

d. If a person's rank correlation between two distributions is close to 1 then they are sure to have a linear relation between them.

Solution:

Explanation: Kurtosis gives us the measure of tailedness and outliers in the distribution of the data. It does not give us the measure of the peak.

The Q-Q plot between the log-normal distribution and normal distribution won’t be a straight line. For getting a straight line we will first have to take the log of each value in the log-normal distribution and then the Q-Q plot of these logarithmically transformed values from log-normal distribution and normal distribution will be a straight line.

For a data distribution like the one in the picture shown below also we can get a good Pearson correlation score but with that, we cannot assume that the relationship between the data is linear.

8. Assume data A=np.array([23.76318668, 27.15338608, 3.37866817, 21.09207851, 1.62700401, 14.46955243, 27.86130518, 3.69529287, 17.6142863 , 3.43226588, 1.12047069, 23.55792867, 11.18580995, 16.89633421, 13.25504754,10.42776142, 4.82765035, 6.21961274, 29.44926291, 12.10018172])
   Which of these techniques will be used to test if these the distribution A normal distribution? **(#EASY #MSQ)**

   a. Standardization
   b. Q-Q plot
   c. KS test
   d. KL divergences

9. Consider a random variable X with values [1,2,3,4,5].Assume we have applied boxcox transforms on X with lambda=2
   Y = boxcox(X,lambda=2)

The mean of the distribution Y is **(#MEDIUM #MCQ)**

a. 3.6
b. 4.3
c. 4.8
d. 5

Solution:

10. Consider the following array

array= [1, 2, 3, 5, 10,11, 12, 13, 14, 15, 21,23, 25, 26, 27, 29, 30,31, 35, 51]

and we have used np.histogram(array) to plot its histogram with bins=5. Find the number of data points will be present in 4th bin are ___ (#mcq #medium)

a. 1
b. 2
c. 5
d. 7

Solution:

11. A disease test is advertised as being 99% accurate: if you have the disease, you will test positive 99% of the time, and if you don't have the disease, you will test negative 99% of the time. If 1% of all people have this disease and you test positive, what is the probability that you actually have the disease? **(#mcq #medium)**

a. 33%
b. 99%
c. 50%
d. 100%

Solution:

You either don't have the disease and tested correctly (1%×99%), or you don't have the disease and tested incorrectly (99%×1%). By Bayes' Theorem, the probability you have the disease is

(1%×99%)/(1%×99%+99%+1%)=50%.

12. In a certain game of tennis, Nadel has a 60% probability to win any given point against Federer. The player who gets to 4 points first wins the game, and points cannot end in a tie.What is Nael's probability to win the game? Try to use your intuition, rather than making a calculation.

  a. Exactly 60%
  b. Less than 60%
  c. Greater than 60%
  d. None of the above

Solution:

Nadel has an advantage over Federer to win any given point. Because the game is played over many points, it becomes less and less likely that Federer will be able to win. Intuitively, you can think about the idea that Federer stringing together many "upsets" is less likely than a single upset (40%). Thus, Nadel has a greater than 60% probability to win the game. (This phenomenon might be even more clear if you think about a game with, say, 1000 points.)

13. Which of the following statements is correct for t-SNE and PCA? **(#EASY #MCQ)**

  a. t-SNE is linear whereas PCA is non-linear
  b. t-SNE and PCA both are linear
  c. t-SNE and PCA both are nonlinear
  d. t-SNE is nonlinear whereas PCA is linear

Solution:

t-Distributed Stochastic Neighbor Embedding is a non-linear dimensionality reduction algorithm used for exploring high-dimensional data. It maps multi-dimensional data to two or more dimensions suitable for human observation. With help of the t-SNE algorithms, you may have to plot fewer exploratory data analysis plots next time you work with high dimensional data.

14. {**medium msq**}

Chebyshev's inequality can be expressed as:

A. $P(|x - \mu| >= k\sigma) <= 1/k^2$
B. $P(|\mu - x| < k\sigma) > 1 - 1/k^2$
C. $P((\mu - k\sigma) < x < (\mu + k\sigma)\ ) > 1 - 1/k^2$
D. None of the above

Ans: Option A is the original expression of Chebyshev's inequality. Options B & C are just rearrangements of the original equation. So all options except D are correct.

15. {**medium #MSQ**}

For a given μ, as the parameter **σ** of log normal distribution is increases, it becomes more:

A. Right skewed.
B. Left Skewed
C. Positively skewed
D. None of the above

Ans: Lognormal distributions are positively skewed.
Positively skewed = Right skewed
Hence options A & C are correct.
https://en.wikipedia.org/wiki/File:PDF-log_normal_distributions.svg

16. {**easy MCQ**}

The length of comments posted on online social media forums follows

A. Gaussian distribution
B. Uniform distribution
C. Log Normal distribution
D. Bernoulli distribution

Ans: Comments contain outlier elements where a few people post extremely long comments. This implies a right skewed long tailed distribution. Gaussian and Uniform distributions are symmetric, hence they are wrong options.

Bernoulli distribution can have only 2 values hence this option is wrong.

Lognormal is right skewed long tailed distribution and hence option C is correct.

17. {**easy #MCQ**}

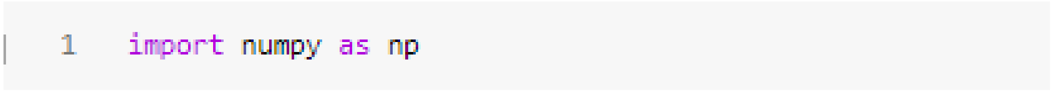
The income of people in the world can be most closely modelled by:

A. Log normal distribution
B. Pareto distribution
C. Exponential distribution
D. None of the above

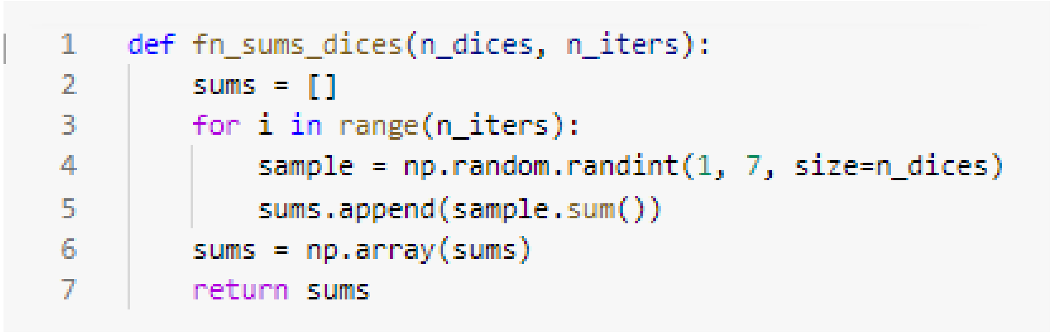
Ans: In general it is observed that 80% of a country's wealth is concentrated in the hands of only 20% of the population. The pareto distribution models this property

18. {**hard**}

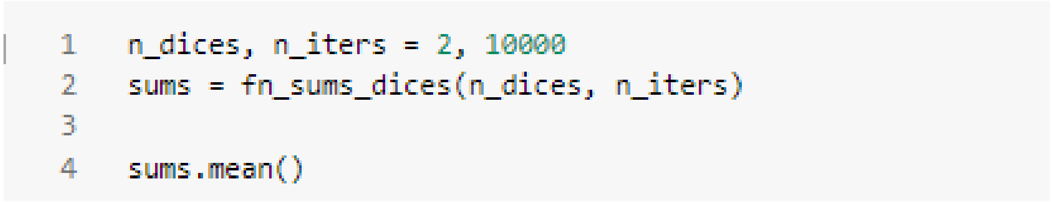
The value returned by the code below will be closest to:



```
import numpy as np
```



```
def fn_sums_dices(n_dices, n_iters):
    sums = []
    for i in range(n_iters):
        sample = np.random.randint(1, 7, size=n_dices)
        sums.append(sample.sum())
    sums = np.array(sums)
    return sums
```



```
n_dices, n_iters = 2, 10000
sums = fn_sums_dices(n_dices, n_iters)

sums.mean()
```

A. 6
B. 3.5
C. 7
D. None of the above

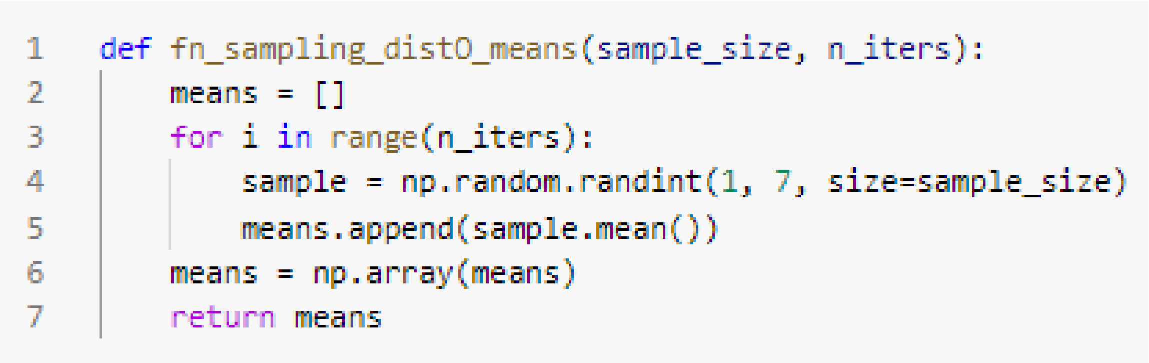
Ans: The function **fn_sums_dice** basically simulates simultaneous dice throws of multiple six faced dies.

The code below the above function simulates 2 dice thrown simultaneously 10000 times.

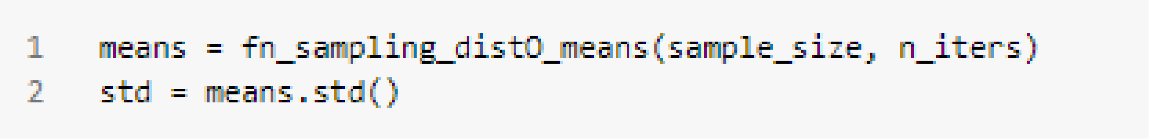
Since dice throws are completely random (ie: uniform distribution), each dice given enough iterations will have an average value of (1+2+3+4+5+6)/6 = 3.5. Since there are 2 dice, the sum of their values will be = 7.

19. {**hard**}

Given the code shown below:



```python
def fn_sampling_dist0_means(sample_size, n_iters):
    means = []
    for i in range(n_iters):
        sample = np.random.randint(1, 7, size=sample_size)
        means.append(sample.mean())
    means = np.array(means)
    return means
```



```python
means = fn_sampling_dist0_means(sample_size, n_iters)
std = means.std()
```

A. As the **sample_size** parameter is increased, the value of **std** generally remains the same.
B. As the **sample_size** parameter is increased value of **std** increases
C. As the **n_iters** parameter is increased, the value of **std** generally remains the same.
D. As the **sample_size** parameter is increased value of **std** decreases

Ans: The larger the sample size, the more representative it will be of the population and the lesser the variation there will be between the means of various samples. Hence option D is correct. The n_iters has no influence on the same size and hence no influence of the variability of the means of the samples. Hence option C is also correct.

20. {**easy**}

Given that we have two samples and we want to know if they come from the same Population we use:

A. Bootstrap resampling
B. Proportional sampling
C. Permutation sampling
D. None of the above

21. {**easy**}

Point the odd one out:

a. log-log plot

b. Q-Q plot
c. K-S test
d. Box-Cox Transform.

Ans: Box-Cox transform is used to transform non normal distributions to normal, whereas the other plots/tests are used to compare 2 distributions. Hence D is the correct option.

22. {**easy**}

The p value is

A. The probability of the Null hypothesis given a particular observation
B. The probability that the Null hypothesis is true
C. The probability of seeing a particular observation if the null hypothesis is true
D. None of the above

Ans: p value = P(observation | null hypothesis)

23. {**easy**}

Which of the following statements are true:

A. AB testing is an example of hypothesis testing.
B. Resampling is used when we have few samples from which to make inferences
C. Concurrent AB testing is preferred for clinical trials
D. None of the above

Ans:
AB testing follows the same principles as hypothesis testing hence only option A is correct.
We can use resampling techniques to simulate the population and sample from it. Hence option B is correct
Concurrent AB testing is not as thorough as plain AB testing and hence not preferred for clinical trials.

24. {**easy**}

We check if a sample is power law distribution using:

A. KS test
B. Log-log plot
C. Q Q plot
D. Parento Test.

25. {**medium**}

Given that we want to check for the similarity of two samples using the absolute difference of the means of two samples for a significance level of 5%, we use:

A. Two tailed test where p_val = sum(sampling_distribution >= 0.05)/n_samples
B. Two tailed test where p_val = sum(sampling_distribution >= 0.025)/n_samples

C. One tailed test where p_val = sum(sampling_distribution >= 0.05)/n_samples

D. None of the above

Ans: Absolute difference of means will yield only positive values and hence the distribution will be one tailed.

26. **{easy}**

Which of the following is used to test a Null Hypothesis?

A. Variance statistic

B. Null statistic

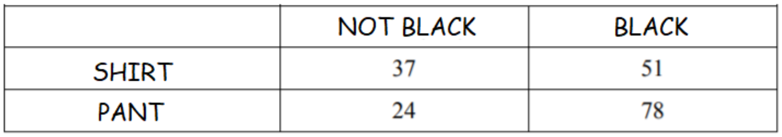
C. Test statistic

D. Population statistic

Ans: Test statistic refers to the specific measurement of the sample that is being studied.

27. {**medium**}

Given the chart below:



	NOT BLACK	BLACK
SHIRT	37	51
PANT	24	78

Which of the following statements are true:

A. P(pant) =0.53

B. P(black|shirt) =0.58

C. P(shirt|not black) = 0.60

D. P(pant|black) = 0.43

Ans:

P(pant) = total n_pants/total n_clothes = (24+78)/190 = 0.53

P(black | shirts) = total n_black_shirts/total n_shirts = 51/(37+51) = 0.58

P(shirt | not black) = total n_non_black_shirts/Total n_non_black_clothes = 37/(37+24) = 0.60

P(pant | black) = total n_black_pants/total n_black_clothes = 78/(78+51) = 0.60

28. Parametric test unlike non-parametric test make certain assumptions about

A. Sample size

B. Population size

C. Population distribution

D. None of the above

Ans: Parametric tests assume that the distribution of the population that the samples are drawn from to are normally distributed.

29. {**easy**}

Which of the following are true:

A. Dependent variable = Features

B. Independent variable = Features

C. Independent variable = labels

D. None of the above

30. {**easy**}

Given a 2D dataset, if we want to visualize the separability of the classes it contains we can use:

A. CDF plot

B. Scatter plot

C. Contour plot

D. None of the above

31. {**easy**}

If we want to know the frequency count of the classes in the output variable, we can use:

A. CDF

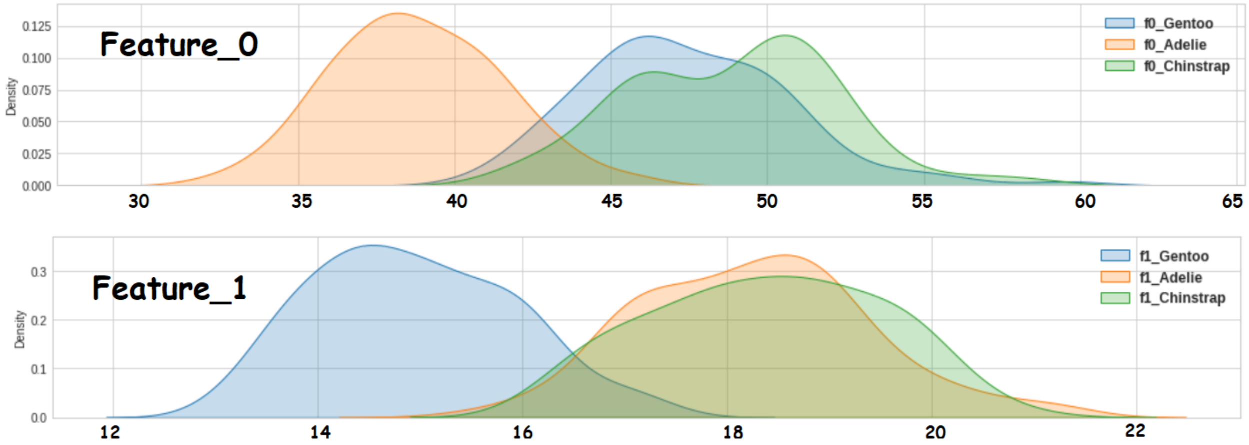
B. PDF

C. Histogram

D. Scatter plot

32. {**medium**}

Given the distribution shown below:



Which of the approximate thresholds are the best:

A. We can separate the Gentoo class from others by using feature_1 <= 18

B. We can separate the Gentoo class from others by using feature_1 <= 16.5

C. We can separate the Adelie class from others by using feature_0 <= 43

D. We can separate the Adelie class from others by using feature_0 <= 47.5

33. {**medium**}

Given that the 1st, 2nd and 3rd quartiles are equi-spaced, the distribution could be:

A. Normal distribution
B. Uniform distribution
C. Log normal distribution
D. None of the above

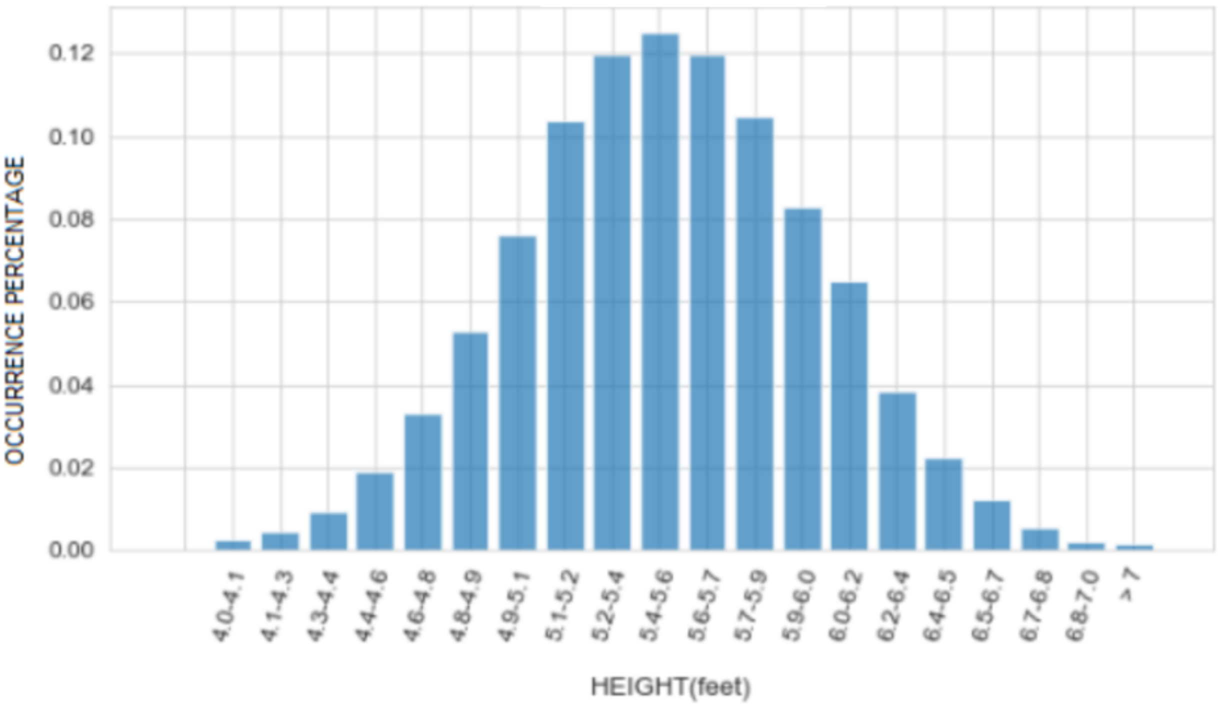
34. {**medium**}

With respect to the thresholds chosen in box plots for determining outliers, which of the following statements are true:

A. Lower threshold = Q1 + (1.5 x IQR)
B. Higher threshold = Q2 + (1.5 x IQR)
C. Lower threshold = Q1 **-** (1.5 x IQR)
D. Higher threshold = Q3 + (1.5 x IQR)

35. {**easy**}

Assuming that the plot shown below accurately depicts the percentages of height ranges for the human population, which of the following statements are True:



A. The probability that the height of the next person you meet will be in the range 4.3 to 4.6 is around 3%.

B. The probability that the height of the next person you meet will be in the range 5.7 to 6.0 is around 18%.

C. The probability that the height of the next person you meet will not be in the range 6.5 to 6.7 is around 97%.

D. None of the above

36. {**medium**}

Given a normal distribution with mean (μ) = 5.5 and std deviation (σ) = 0.5, which of the following statements are true:

A. Around 48% of its values lie between -2σ and μ

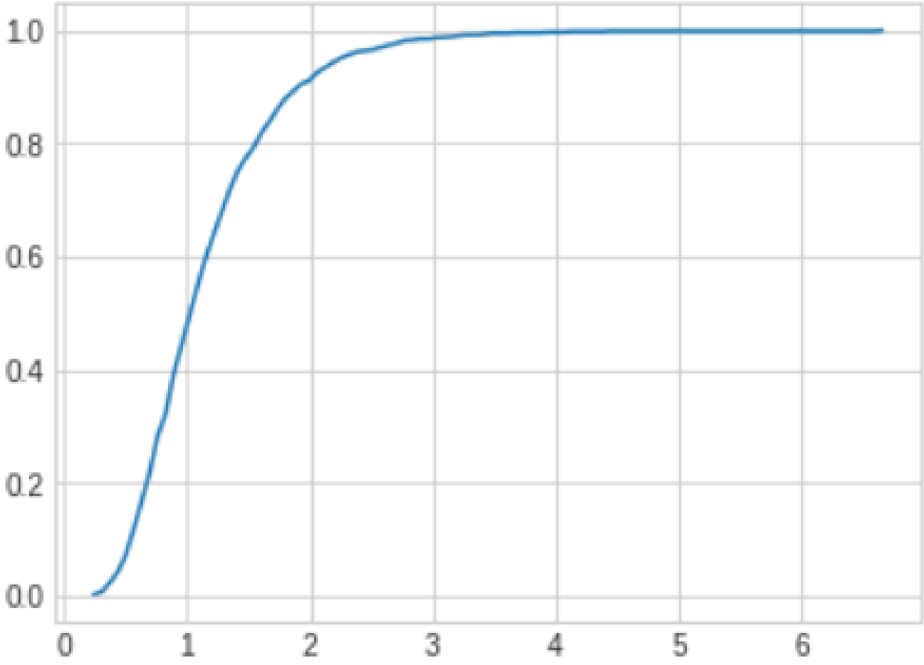
B. Around 4% of its values lie above +2σ

C. Around 34% of its values lie between μ and +1σ

D. None of the above

37. {**medium**}

Given the CDF shown below:



Which of the following statements are true:

A. Around 20% of the values are <= 0.7

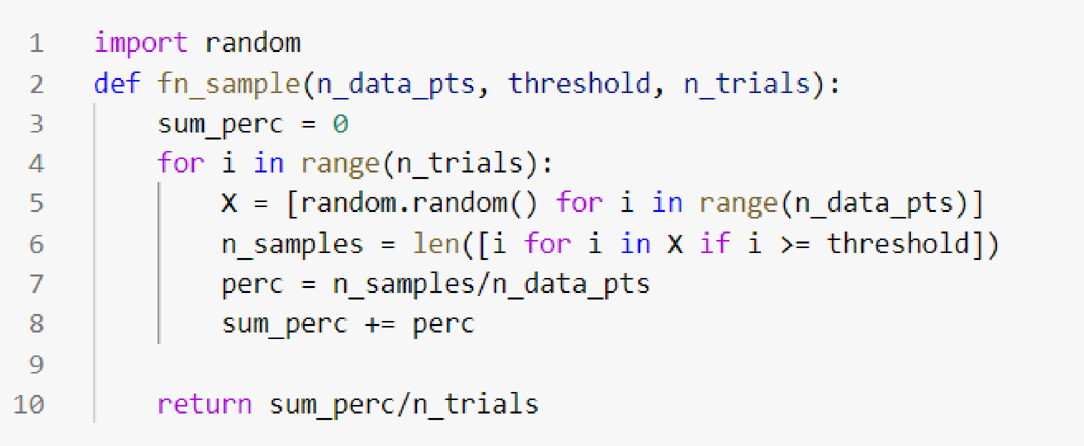
B. Around 10% of the values are >= 3

C. Around 40% of the values are between 1 & 2.

D. None of the above

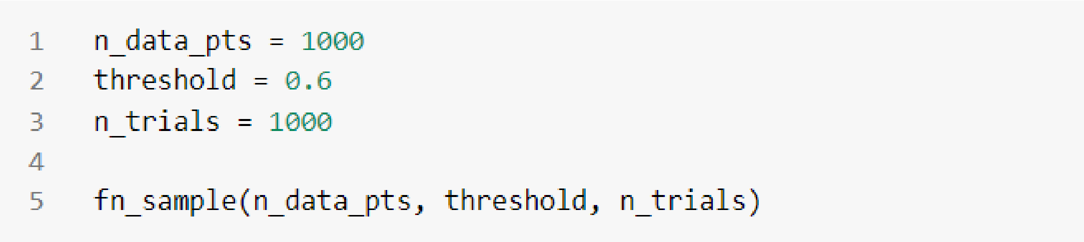
38. {**hard**}

The output of code shown below will be:



```python
import random
def fn_sample(n_data_pts, threshold, n_trials):
    sum_perc = 0
    for i in range(n_trials):
        X = [random.random() for i in range(n_data_pts)]
        n_samples = len([i for i in X if i >= threshold])
        perc = n_samples/n_data_pts
        sum_perc += perc

    return sum_perc/n_trials
```



```python
n_data_pts = 1000
threshold = 0.6
n_trials = 1000

fn_sample(n_data_pts, threshold, n_trials)
```

A. Around 0.5

B. Around 0.4

C. Around 0.6

D. None of the above

39. {**easy**}

Which of the following equations are True:

A. Mean Absolute Deviation $= \frac{1}{n}\sum_{i} \text{abs}\left(\bar{x} - x_i\right)$

B. Standard Deviation $= \frac{1}{n}\sum_{i} \left(\bar{x} - x_i\right)^2$

C. Median Absolute Deviation, MAD $=$ median $\left(\text{abs}\left(\bar{x} - x_i\right) \text{ for all i}\right)$

D. None of the above

40. {**easy MCQ**}

Which of the following statements are true:

A. After standardization all the values of the dataset will lie within a unit cube.

B. The mean of a standardized dataset will be at 0.

C. Standardization reduces the variance of all columns of the dataset.
D. None of the above.

Ans:
Standardization scales all columns to have mean = 0 and std_deviation = 1, hence option A is wrong and Option B is correct.
Standardization increases or decreases the variance of columns depending upon whether their original std_deviation is > or < 1, so option C is wrong.

41. {**easy MSQ**}
The covariance matrix S of column standardized data matrix X:

A. Has the same dimensionality as the dataset
B. Is always a symmetric matrix.
C. Has a diagonal which contains variances of each column of X
D. Each cell represents a dot product.

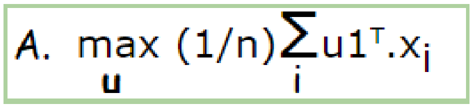
Ans:
S will always be a square matrix, this need not be true of X, hence option A is wrong.
Since S contains covariances of all column pairs of X, it will be a symmetric matrix, hence option B is correct
The diagonal of S will contain covariance of columns with themselves hence option C is correct.
Since X is column standardized, mean of each column will be zero, hence covariances between each pair of column will represent a dot product

42. {**easy MCQ**}
V is an principal component of data matrix X having covariance matrix S, if :

A. Its magnitude does not change after being multiplied with S.
B. Its direction does not change after being multiplied with S.
C. It becomes a unit vector after being multiplied with S.
D. None of the above

Ans: V is an principal component of data matrix X having covariance matrix S only if its direction does not change after being multiplied with S - hence only option B is correct

43. {**medium MCQ**}
Given a d dimensional column standardized matrix $X_{n \times d}$ containing n vectors, the objective function of PCA can be defined as (u is unit vector):


A. $\max_{\mathbf{u}} (1/n)\sum_{i} u1^{T}.x_{i}$

B. $\max\limits_{\mathbf{u}} \sum\limits_{i} (d_i)^2$ where $d_i = \sqrt{(x_i{}^T x_i - u^T x_i\ )}$

C. $\max\limits_{\mathbf{u}} X.u$

D. None of the above

Ans: Option A represents finding that u that maximises the variance in X, hence this option is correct.
Otiona B & C are other ways to represent option A hence they too are correct.

44. {**easy MCQ**}

Given that S is the covariance matrix of some dataset X, if S[1, 3] = 12, then
S[3, 1] will be:

A. 1/12
B. 12
C. 144
D. None of the above

Ans: Covariance matrix will be a symmetric square matrix, henace option B is correct.

45. {**easy MSQ**}

Given that λ = [λ1, λ2….λn] are the set of eigenvalues of corresponding to the covariance matrix of dataset X, then the most important eigen vectors is given by the vector corresponding to the following eigenvalue:

A. max(λ)
B. Sorted(λ)[-1]
C. Sorted(λ)[::-1][1]
D. None of the above

Ans: The eigenvector corresponding to the largest eigenvalue explained the max variance contained in the data. Hence option A, B & C are correct.

46. {**easy MSQ**}

Given a data matrix X having d features, which of the following is true:

A. Covariance matrix of X will have d eigenvectors.
B. The dot product of ant 2 eigenvectors = 1
C. Each eigenvector will be d dimensional
D. None of the above.

Ans:

Number of eigenvectors is always = number of dimensions, hence option A is correct

All eigenvectors are orthogonal to each other hence dot product between any of them will be = 0, hence option B is wrong.

Eigenvectors always have the same dimensionality as the original dataset, hence option C is correct

47. {**medium MSQ**}

After performing PCA on dataset X:

A. The amount of information contained within the data remains the same if the dimensionality of the X is maintained .

B. The interpretability of the dataset's features is lost.

C. The new features are ordered as per their importance.

D. None of the above

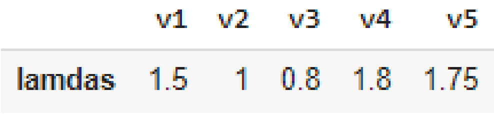
Ans: As long as the dimensionality is not reduced information is not lost hence option A is correct

The new features obtained after PCA are not a linear combination of the original feature, hence interpretability is lost- option B is correct

The matrix containing the eigen vectors must be ordered based on the amount of variance they explain before matrix multiplication with X - option C is correct.

48. {**medium MCQ**}

Given below are the eigenvectors (v1, v2...v5) of the covariance matrix of a dataset X and their corresponding eigenvalues.



	v1	v2	v3	v4	v5
lamdas	1.5	1	0.8	1.8	1.75

Say we want to preserve 50% of the information of the dataset after dimensionality reduction using PCA. We will use the following principal components:

A. v1, v2 & v3

B. v3, v4, & v5

C. v4 & v5

D. None

Ans:

I. Find the % values of each lambda wrt to the sum of the lambdas

II. We then sort these values in decreasing order of magnitude

III. Find out how many values from the beginning need to be added to get a sum >= 0.5

IV. The eigenvectors corresponding to those values will give us the principal components that explain at least 50% of variance in the data. These turn out to be those mentioned in option C.

49. {**medium MSQ**}

Given a dataset $X_{n \times d}$ and matrix $V_{d \times d}$ containing the eigenvectors of its covariance matrix in increasing order of variance they explain, then we can perform dimensionality reduction to m (m< d) dimensions by:

A. X^T V[:m]

B. XV’ where V’ = V[:, :m]

C. Z[:, :m] where Z = XV

D. None of the above

Ans: X^T will have d rows and n columns, so matrix multiplication shown in option A won’t work, hence option A is wrong.

Options B & C mean the same thing and result is a matrix of n rows and n columns and hence these options are correct.

50. {**medium MCQ**}

For Dimensionality reduction using tSNE:

A. As we increase perplexity the quality of the embedding keeps improving.

B. As we increase the number of iterations the quality of the embedding generally keeps improving.

C. The distances between clusters in the embedding gives an idea as to how different they are.

D. None of the above

Ans:

As we increase perplexity, the quality of the embedding improves upto a certain point after which it gets worse. Hence option A is wrong

Since tSNE is a stochastic process where each iteration tries to improve on the previous one, option B is correct.

tSNE tries to preserve local information and does not give importance to global information, hence option C is wrong.

51. {**easy**}

To check which website design yields a better conversion rate, we use:

A. AB testing

B. Hypothesis testing

C. Randomized controlled trials

D. Permutation test.

Ans: A, B & C mean the same thing and they are used for the above mentioned task