**Name: Dishant gupta**

**Course: Full stack data science pro**

**Assignment: continue from question 20:**

1. **What do you mean by Measure of Central Tendency and Measures of Dispersion .How it can be calculated.**

**Measures of central tendency** are statistical metrics that describe the center or typical value of a dataset. The three main measures of central tendency are:

1. Mean (Average): The sum of all data points divided by the number of data points.
2. Median: The middle value of a dataset when it is ordered from smallest to largest. If there is an even number of observations, the median is the average of the two middle numbers.
3. Mode: The value that appears most frequently in a dataset.

### **Measures of dispersion** describe the spread or variability of a dataset. The main measures of dispersion include:

### Range: The difference between the highest and lowest values in a dataset.

### Variance: The average of the squared differences from the mean.

### Standard Deviation: The square root of the variance, representing the average distance from the mean.

### Interquartile Range (IQR): The range between the first quartile (25th percentile) and the third quartile (75th percentile).

**2. What do you mean by skewness? Explain its types. Use graphs to show.**

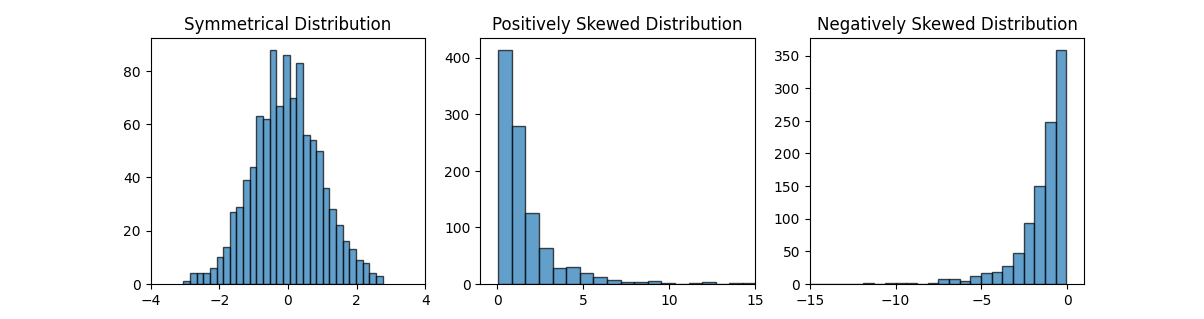
**Skewness** is a measure of the asymmetry of the probability distribution of a real-valued random variable about its mean. It indicates whether the data points are spread out more on one side of the mean than the other. Skewness can be positive, negative, or zero.

#### **Types of Skewness:**

#### **Positive Skewness (Right Skewed)**:

* + The right tail (higher values) is longer or fatter than the left tail (lower values).
  + Most of the data points are concentrated on the left side of the distribution, with fewer larger values stretching out to the right.
  + Example: Income distribution where a few people have very high incomes.

1. **Negative Skewness (Left Skewed)**:
   * The left tail (lower values) is longer or fatter than the right tail (higher values).
   * Most of the data points are concentrated on the right side of the distribution, with fewer smaller values stretching out to the left.
   * Example: Age at retirement, where most people retire around a certain age, but a few retire much earlier.
2. **Zero Skewness (Symmetrical Distribution)**:
   * The tails on both sides of the mean are equally balanced.
   * The mean, median, and mode of the distribution are all equal.

****

**3. Explain PROBABILITY MASS FUNCTION (PMF) and PROBABILITY DENSITY FUNCTION (PDF). and what is the difference between them?**

1. Probability mass function (pmf): pmf is used for discrete random variable X gives the probability that X takes on a specific value x. P(X = x) lies in interval 0 and 1.
2. Probability density function (pdf): pdf is used for continuous random variables. A continuous random variable is one that can take on an infinite number of values within a given range. PDF of a continuous random variable X gives the relative likelihood of X taking on a particular value. Here f(x)>= 0 for all x and integration of f(x) wrt x from -inf to +inf is +1.

The probability that X lie within interval [a, b] given by

P(a <= X <= b) = integration of f(x) wrt x from a to b.

**4. What is correlation? Explain its types in detail. What are the methods of determining correlation?**

Correlation is used as it covers the limitations of covariance. Correlation tell us magnitude and direction of relationship. It is a dimensionless quantity and its range is -1 to +1. Correlation between two variables x and y let us say alpha. Then physically, alpha represents how much sure is it if there is 1 unit change in x, other unit will also change by 0.4. Types of correlation are:

1. Positive correlation: ranges from 0 to 1
2. Negative correlation: ranges from -1 to 0.
3. Zero correlation: means independent features.

**Pearson correlation coefficient = Covariance(X, Y) / ((std(X) \* std(Y))**

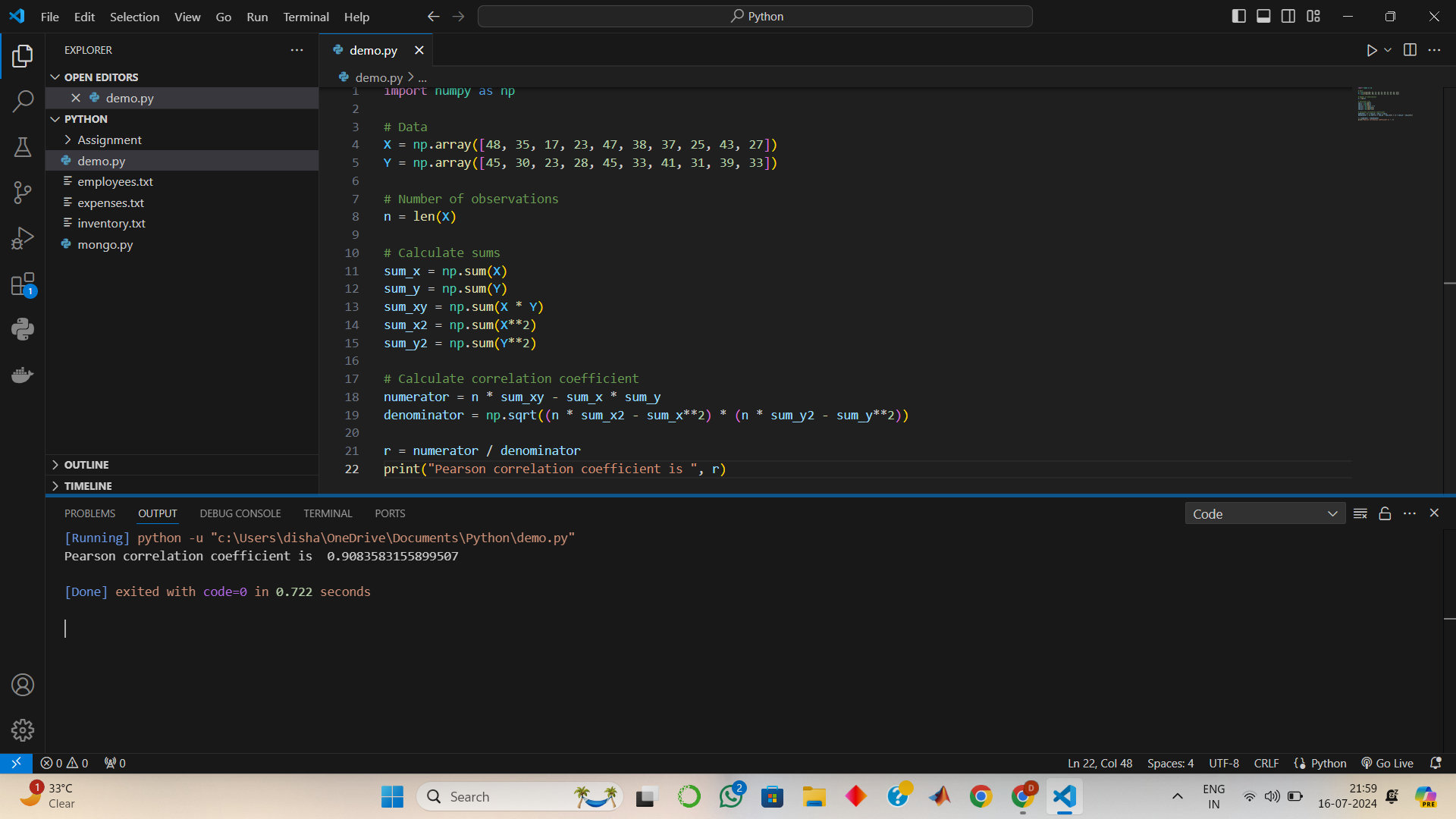
**Pearson correlation always measures linear relationship. For other type of curves, we have spearman rank correlation formula.**

**Spearman rank correlation = Cov(R(x), R(y)) / (sigma(R(x)) \* sigma(R(y)))**

R(x) represents rank of matrix x and R(y) is rank of y.

**5. Calculate coefficient of correlation between the marks obtained by 10 students in Accountancy and statistics:**

| **Student** | **Accountancy** | Statistics |
| --- | --- | --- |
| 1 | 45 | 35 |
| 2 | 70 | 90 |
| 3 | 65 | 70 |
| 4 | 30 | 40 |
| 5 | 90 | 95 |
| 6 | 40 | 40 |
| 7 | 50 | 60 |
| 8 | 75 | 80 |
| 9 | 85 | 80 |
| 10 | 60 | 50 |



**6. Discuss the 4 differences between correlation and regression.**

| **Correlation** | **Regression** |
| --- | --- |
| It measures the strength and direction of linear relationship between two variables. | It estimates the relationship between a dependent variable and one or more independent variables. |
| It is symmetric in nature. | It is anti symmetric in nature. |
| Its purpose is to quantify the degree of association between two variables. | Its purpose is to predict the value of the dependent variable based on the independent variables. |
| It is unitless. It is pure number from -1 to +1. | It has units, the slope of regression line has units dependent on the units of variables. |
|  |  |

**7. Find the most likely price at Delhi corresponding to the price of Rs. 70 at Agra from the following data: Coefficient of correlation between the prices of the two places +0.8.**

Here, physical meaning of correlation of (x, y) = 0.8 that with 1 unit change in x (or y) , how much sure the change in y (or x) will be 0.8

Thus most likely price at delhi will be 70 + (70 \**0.8) or 70 - (70*\*0.8)

75.6 or 64.4

**8. In a partially destroyed laboratory record of an analysis of correlation data, the following results only are legible: Variance of x = 9, Regression equations are: (i) 8x−10y = −66; (ii) 40x − 18y = 214. What are (a) the**

**mean values of x and y, (b) the coefficient of correlation between x and y, (c) the σ of y.**

**9. What is Normal distribution? What are the four assumptions of normal distribution ? Explain it in detail.**

Normal distribution is a continuous probability distribution characterized by a bell shaped curve, symmetric about mean. It is defined by its mean (mu) and standard deviation (sigma).

Four Assumptions of normal distribution:

1. The distribution is symmetric about mean
2. Mean, median, mode are equal.
3. The tails of distribution extend inf in both directions without touching the x axis.
4. Approximately 68% of the data falls within one standard deviation of the mean.
5. Approximately 95% of the data falls within two standard deviations of the mean.
6. Approximately 99.7% of the data falls within three standard deviations of the mean.
7. The distribution has a single peak (mode).

**10. Write all characteristics or properties of Normal distribution curve.**

1. It is a type of continuous probability distribution for a real valued random variable. The general form of its PDF is

f(x) = 1/(sigma \* sqroot(2\*pi)) exp(-(0.5\*(x-mu)/sigma)\*2)

1. These are symmetric in nature about mean.
2. It follows empirical rule as mentioned above.
3. Mean, median, mode are all equal.
4. It has bell shaped curve.

**11. Which of the following options are correct about Normal Distribution Curve.**

**(a) Within a range 0.6745 of σ on both sides the middle 50% of the observations occur i,e. mean ±0.6745σ**

**covers 50% area 25% on each side.**

**(b) Mean ±1S.D. (i,e.μ ± 1σ) covers 68.268% area, 34.134 % area lies on either side of the mean.**

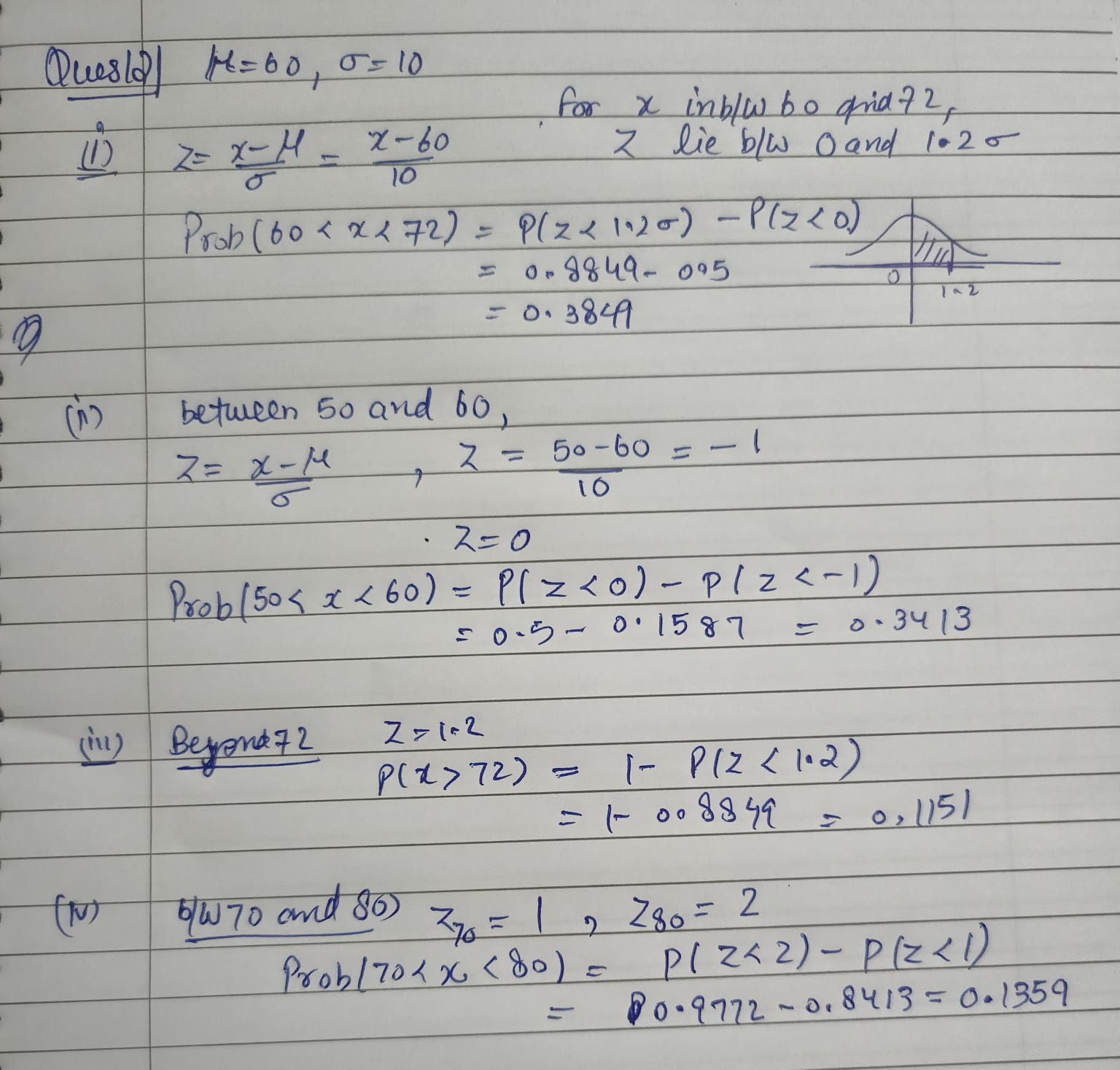
**(c) Mean ±2S.D. (i,e. μ ± 2σ) covers 95.45% area, 47.725% area lies on either side of the mean.**

**(d) Mean ±3 S.D. (i,e. μ ±3σ) covers 99.73% area, 49.856% area lies on the either side of the mean.**

**(e) Only 0.27% area is outside the range μ ±3σ.**

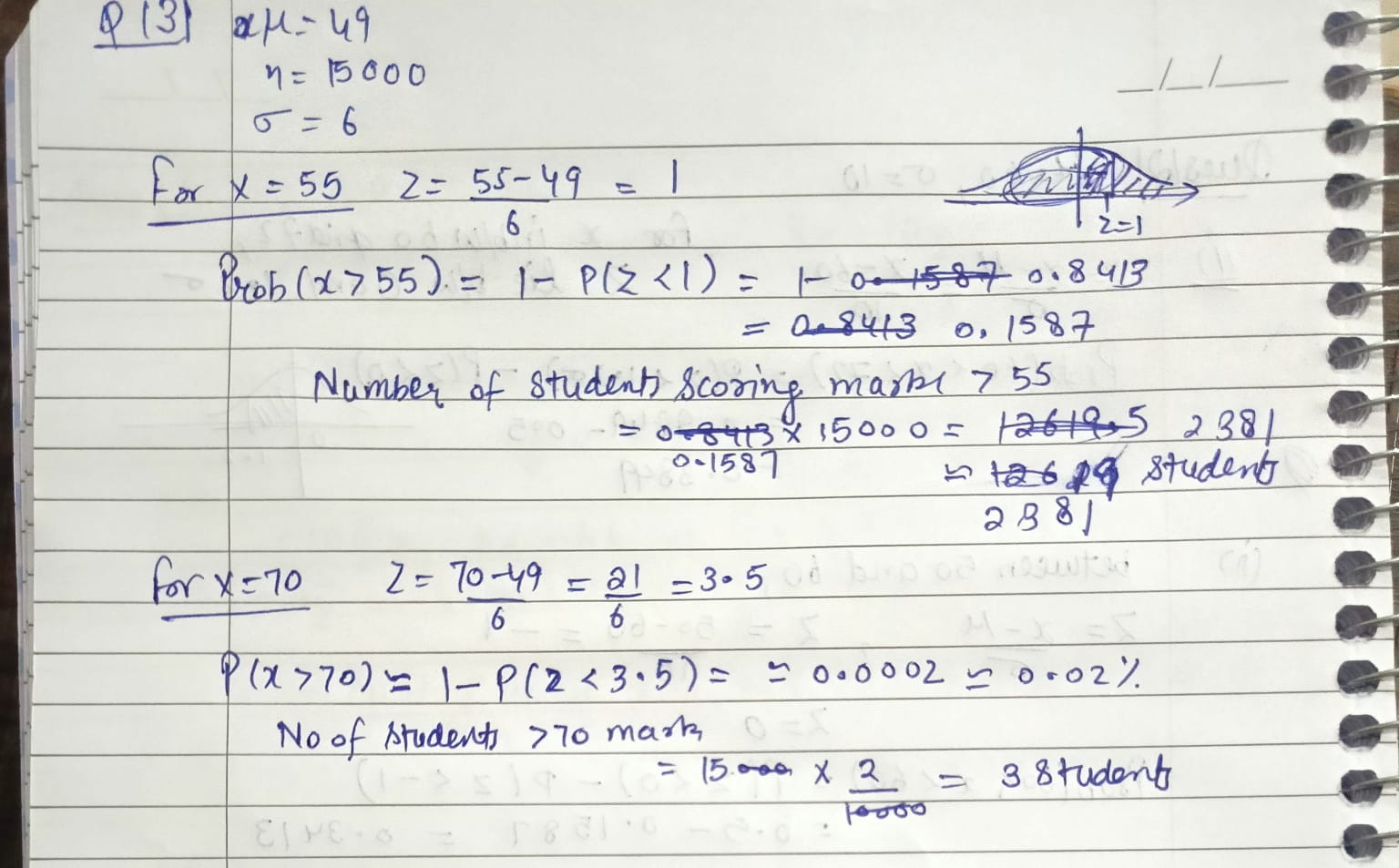
All are correct options.

**12. The mean of a distribution is 60 with a standard deviation of 10. Assuming that the distribution is normal, what percentage of items be (i) between 60 and 72, (ii) between 50 and 60, (iii) beyond 72 and (iv) between 70 and 80?**

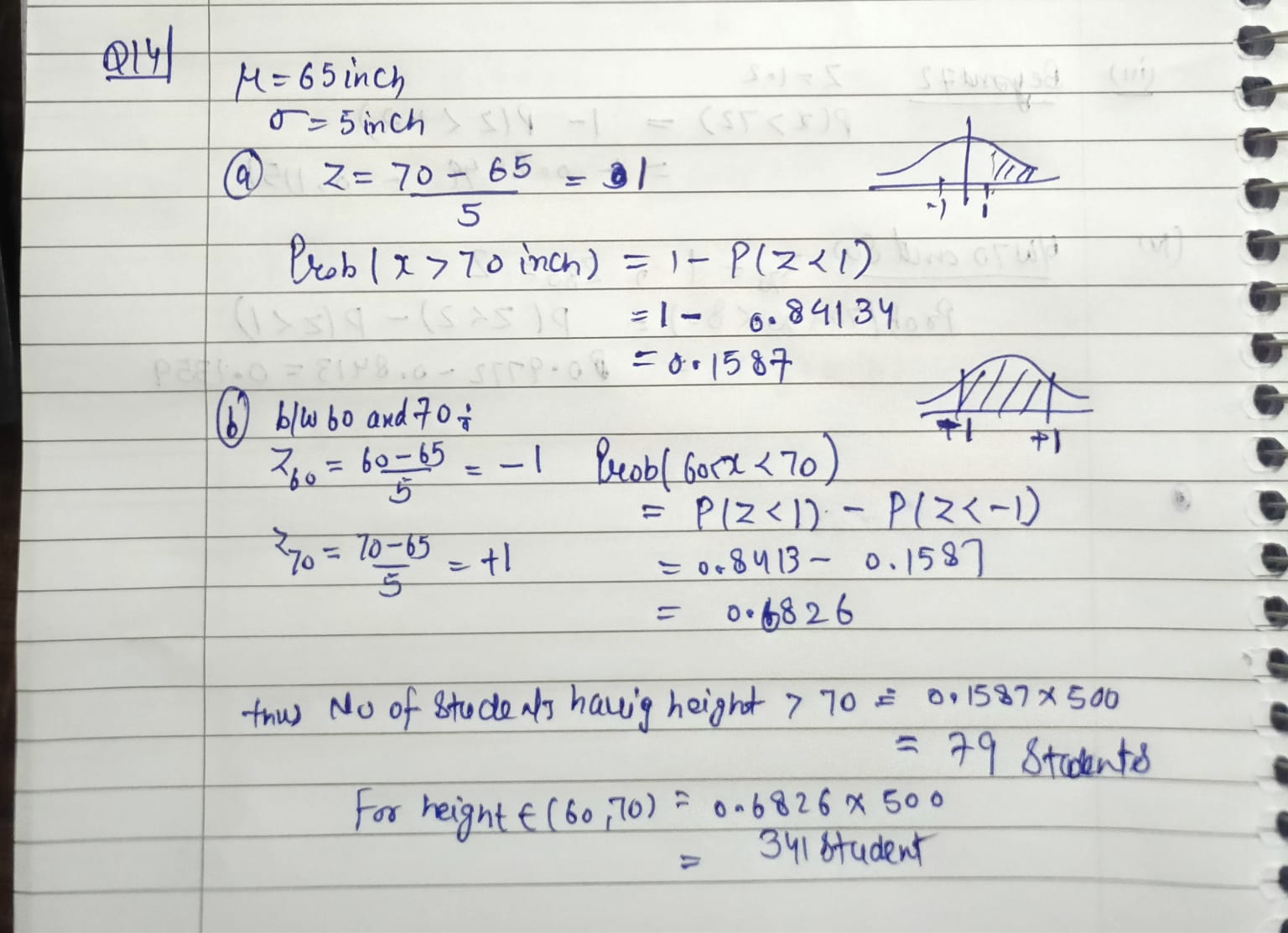


Here my handwritten solution of question.

13. **15000 students sat for an examination. The mean marks was 49 and the distribution of marks had a standard deviation of 6. Assuming that the marks were normally distributed what proportion of students scored (a) more than 55 marks, (b) more than 70 marks.**

****

**14. If the height of 500 students are normally distributed with mean 65 inch and standard deviation 5 inch. How many students have height : a) greater than 70 inch. b) between 60 and 70 inch**



**15. What is the statistical hypothesis? Explain the errors in hypothesis testing.b)Explain the Sample. What are Large Samples & Small Samples?**

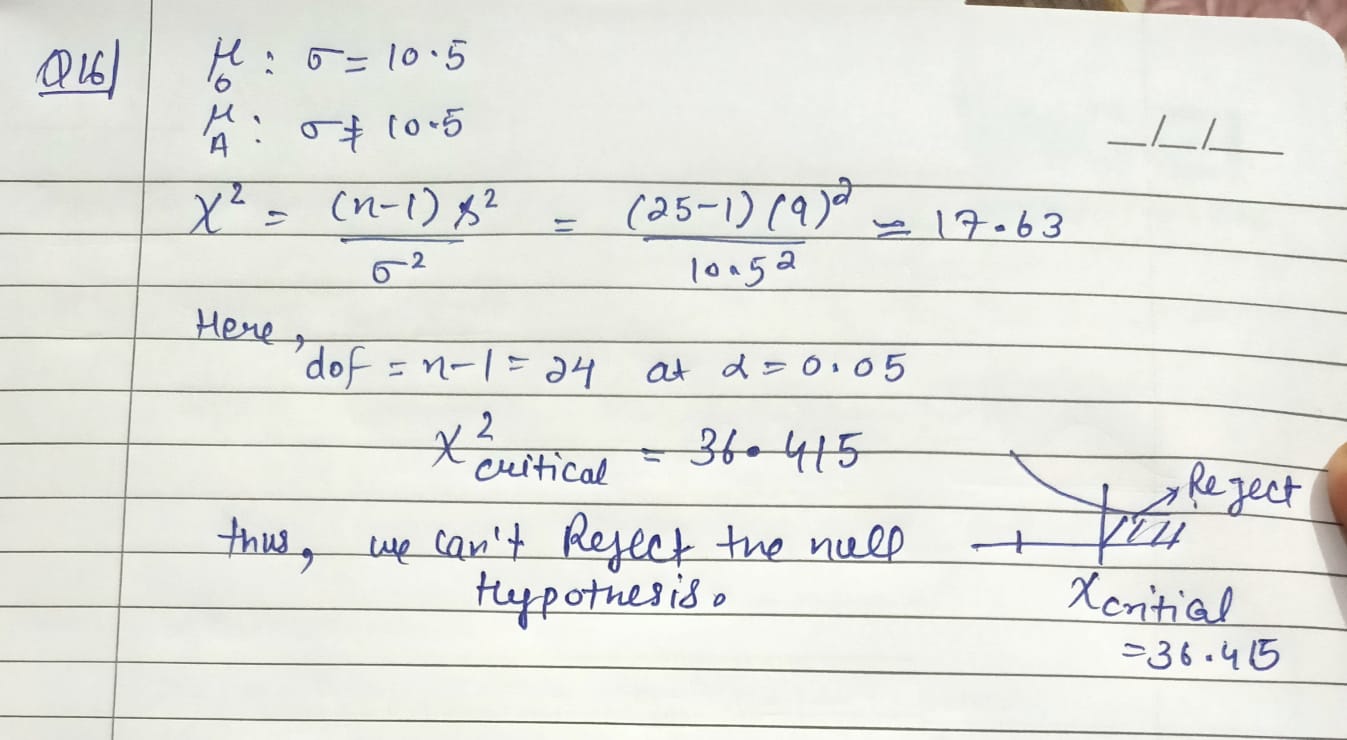
A statistical hypothesis is an assumption or claim about a population parameter (e.g., mean, variance) that can be tested using statistical methods. There are two types of hypotheses:

* Null hypothesis (H0H\_0H0​): It assumes no effect or no difference. It's a statement of no change.
* Alternative hypothesis (H1H\_1H1​): It assumes some effect or difference. It’s what you want to prove.

**Errors in Hypothesis Testing:**

* + Type I Error: Rejecting the null hypothesis when it is actually true.
  + Type II Error: Failing to reject the null hypothesis when it is actually false.
* **Sample:** A sample is a subset of a population used to represent the entire group. It's used to make inferences about the population.
  + **Large Samples:** Generally, a sample size greater than 30 is considered large. Large samples are more likely to represent the population accurately.
  + **Small Samples:** A sample size of 30 or fewer. Special statistical methods may be needed to analyze small samples due to less reliable representation of the population.

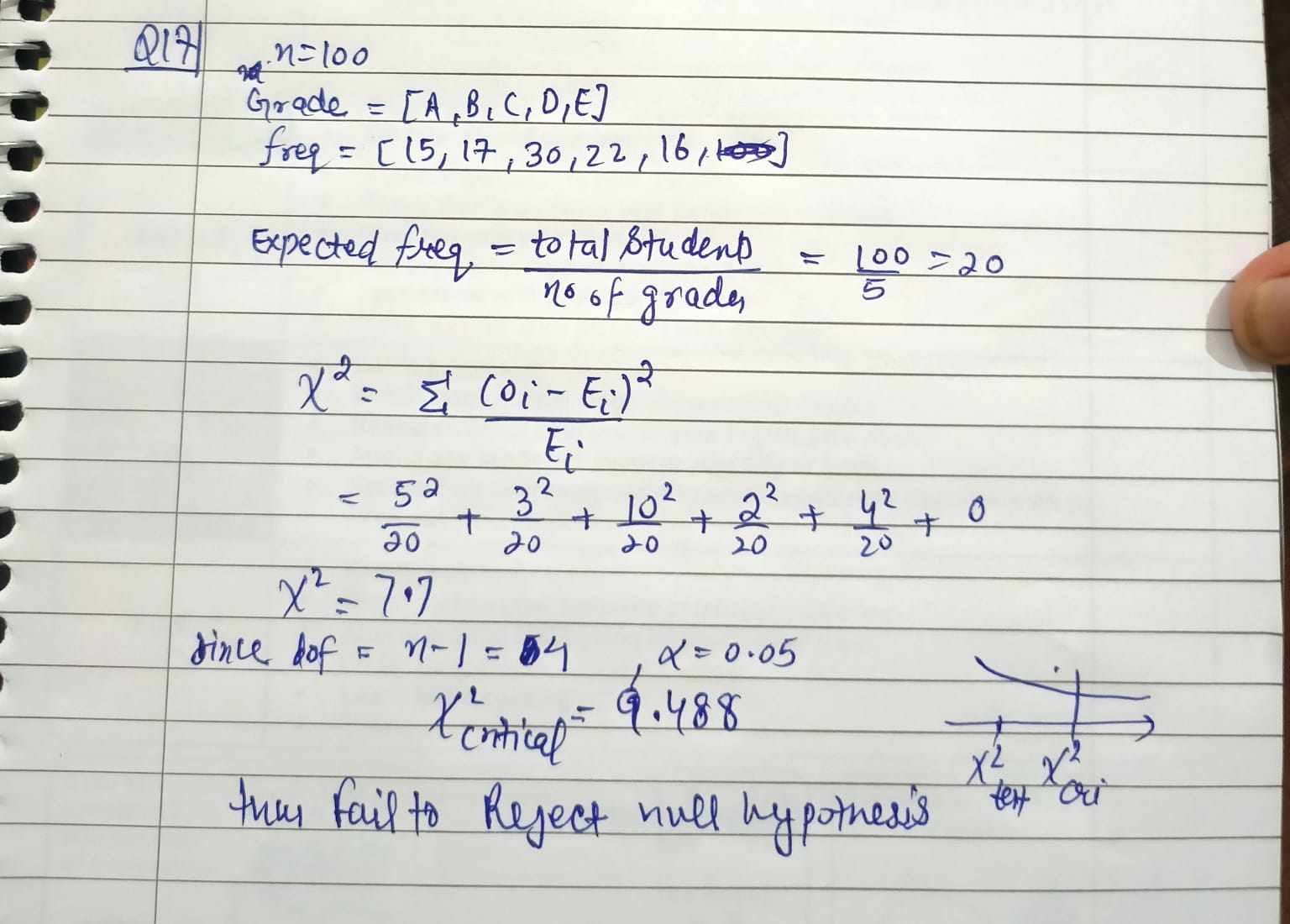
**16. A random sample of size 25 from a population gives the sample standard deviation to be 9.0. Test the hypothesis that the population standard deviation is 10.5. Hint(Use chi-square distribution).**

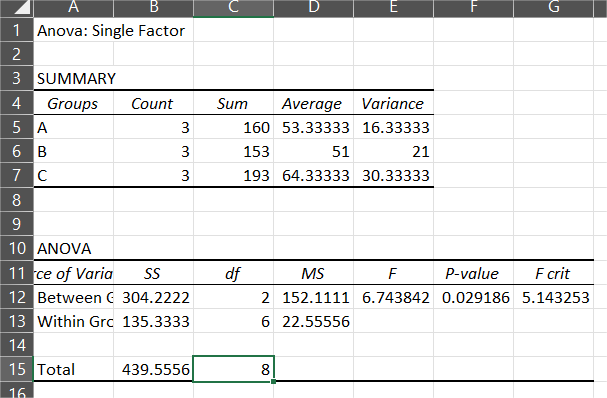


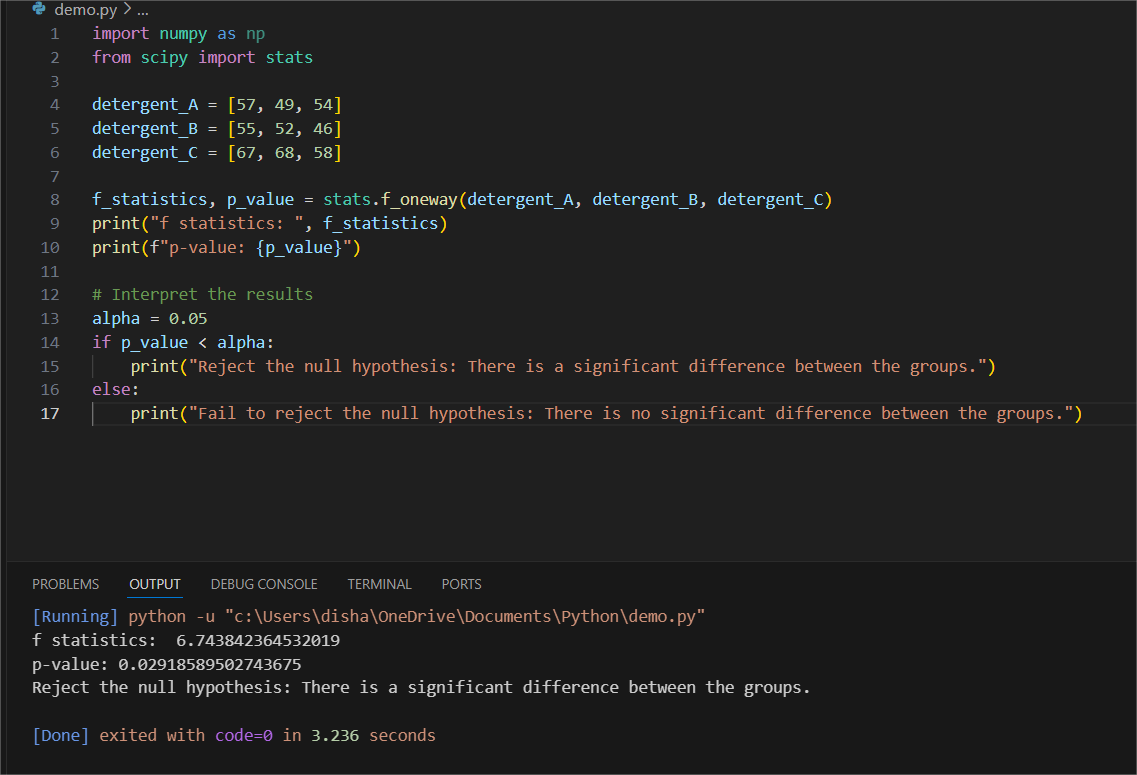
**17. 100 students of a PW IOI obtained the following grades in Data Science paper : Grade :[A, B, C, D, E]**

**Total Frequency :[15, 17, 30, 22, 16, 100]**

**Using the χ 2 test , examine the hypothesis that the distribution of grades is uniform.**

****

**18. Anova test question**

****