1. (a) Briefly explain the following (any 5):
   - (i) **Parameter**
     - A parameter is a descriptive measure that characterizes an entire population. It is a fixed, unknown value that we often try to estimate using sample data. Examples include the population mean (μ), population standard deviation (σ), or population proportion (P).
   - (ii) **Type I error**
     - A Type I error occurs in hypothesis testing when a true null hypothesis is incorrectly rejected. It is also known as a "false positive." The probability of making a Type I error is denoted by α (alpha), which is the significance level of the test.
   - (iii) **Nominal data**
     - Nominal data is a type of qualitative (categorical) data where categories have no inherent order or ranking. Data are simply labels or names used to classify items into distinct groups. Examples include gender (male, female), marital status (single, married, divorced), or blood type (A, B, AB, O).
   - (iv) **Normal distribution**
     - The normal distribution, also known as the Gaussian distribution, is a symmetric, bell-shaped probability distribution that is frequently used in statistics. It is characterized by its mean (μ) and standard deviation (σ). In a normal distribution, the mean, median, and mode are all equal, and about 68% of the data falls within one standard deviation of the mean, 95% within two, and 99.7% within three.
   - (v) **ANOVA**

- ANOVA stands for Analysis of Variance. It is a statistical test used to compare the means of three or more groups to determine if there is a statistically significant difference between them. ANOVA works by partitioning the total variance in a dataset into different components, such as variance between groups and variance within groups.

- (vi) **Central limit theorem**
  - The Central Limit Theorem (CLT) states that, given a sufficiently large sample size from any population with a finite mean and variance, the sampling distribution of the sample mean will be approximately normally distributed, regardless of the shape of the original population distribution. This theorem is fundamental in inferential statistics, as it allows us to use normal distribution theory to make inferences about population parameters even when the population distribution is not normal.

- (vii) **Sign test**
  - The sign test is a non-parametric statistical test used to determine if there is a consistent difference between two related (paired) samples or if a single sample deviates from a hypothesized median. It works by converting quantitative data into signs (positive, negative, or zero) indicating the direction of the difference or deviation, and then analyzes the number of positive and negative signs. It is less powerful than parametric tests but useful when assumptions for parametric tests (e.g., normality) are violated.

2. (b) What is skewness? Explain its different types with the help of graphs.
   - **Skewness**

- Skewness is a measure of the asymmetry of the probability distribution of a real-valued random variable about its mean. In a perfectly symmetrical distribution (like the normal distribution), the two sides of the curve are mirror images, and the skewness is zero. When a distribution is skewed, one tail of the distribution is longer than the other, indicating that the data are concentrated more towards one side.

- **Different Types of Skewness:**
  - **Positive Skewness (Right Skewness):**
    - **Explanation:** In a positively skewed distribution, the tail on the right side of the distribution is longer or fatter than the left side. This indicates that there are more extreme values (outliers) on the higher end of the distribution. The majority of the data points are concentrated on the lower end.
    - **Relationship between Measures of Central Tendency:** For a positively skewed distribution, the mean is typically greater than the median, and the median is typically greater than the mode (Mean > Median > Mode). This is because the extreme high values pull the mean towards the right tail.
    - **Graph:** (Imagine a bell-shaped curve where the peak is shifted to the left, and the right tail extends much further than the left tail).
  - **Negative Skewness (Left Skewness):**
    - **Explanation:** In a negatively skewed distribution, the tail on the left side of the distribution is longer or fatter than the right side. This indicates that there are more extreme values (outliers) on the lower end

of the distribution. The majority of the data points are concentrated on the higher end.

- **Relationship between Measures of Central Tendency:** For a negatively skewed distribution, the mean is typically less than the median, and the median is typically less than the mode (Mean < Median < Mode). This is because the extreme low values pull the mean towards the left tail.
- **Graph:** (Imagine a bell-shaped curve where the peak is shifted to the right, and the left tail extends much further than the right tail).

- **Zero Skewness (Symmetric Distribution):**
    - **Explanation:** A distribution has zero skewness if it is perfectly symmetrical. In such a distribution, the data are evenly distributed around the mean, with no prolonged tails on either side. The normal distribution is a classic example of a perfectly symmetrical distribution.
    - **Relationship between Measures of Central Tendency:** For a perfectly symmetrical distribution, the mean, median, and mode are all equal (Mean = Median = Mode).
    - **Graph:** (Imagine a perfectly symmetrical bell-shaped curve, like a standard normal distribution curve).

3. (c) How are mutually exclusive events different from independent events? Briefly explain with suitable examples.
    - **Mutually Exclusive Events:**
        - **Definition:** Two events are mutually exclusive (or disjoint) if they cannot occur at the same time. If one event

happens, the other cannot. They have no common outcomes.

- **Probability Rule:** If A and B are mutually exclusive events, then the probability of both A and B occurring is zero: $P(A \text{ and } B) = P(A \cap B) = 0$.
- **Example:** Consider rolling a standard six-sided die.
    - Event A: Rolling an even number (2, 4, 6).
    - Event B: Rolling an odd number (1, 3, 5).
    - Events A and B are mutually exclusive because you cannot roll both an even and an odd number at the same time on a single roll. If you roll a 2, you cannot also roll a 3.

- **Independent Events:**
    - **Definition:** Two events are independent if the occurrence of one event does not affect the probability of the other event occurring. The outcome of one event has no influence on the outcome of the other.
    - **Probability Rule:** If A and B are independent events, then the probability of both A and B occurring is the product of their individual probabilities: $P(A \text{ and } B) = P(A \cap B) = P(A) \times P(B)$.
    - **Example:** Consider flipping a coin twice.
        - Event A: Getting heads on the first flip.
        - Event B: Getting heads on the second flip.
        - Events A and B are independent because the outcome of the first coin flip does not influence the outcome of the second coin flip. Getting a head on

the first flip doesn't change the probability of getting a head on the second.

- **Key Difference:**
  - The core difference lies in their relationship:
    - **Mutually Exclusive:** Events cannot happen together. They *exclude* each other.
    - **Independent:** Events happening or not happening *does not influence* each other. They occur separately without affecting probabilities.
  - If two events are mutually exclusive and have non-zero probabilities, they cannot be independent. If A and B are mutually exclusive and $P(A) > 0$ and $P(B) > 0$, then $P(A \cap B) = 0$. However, for independence, $P(A \cap B) = P(A) \times P(B)$, which would be greater than 0. Thus, they cannot satisfy both conditions simultaneously unless one of the probabilities is zero (a trivial case).

4. (a) A single 100-mg dose of a drug is administered orally to 15 adults. The time (in minutes) required by the drug to reach maximum concentration in the blood is recorded as follows: 12, 12, 16, 10, 13, 12, 14, 13, 19, 13, 13, 14, 16, 15, 14.
   - Calculate the three measures of central tendency, D_8, P_{35}, Interquartile Range and Variance for the above mentioned data.
   - **Data (sorted):** 10, 12, 12, 12, 13, 13, 13, 13, 14, 14, 14, 15, 16, 16, 19 (n = 15)
   - **Measures of Central Tendency:**
     - **Mean ($\bar{x}$):**
       - Sum of data = $10 + 12 + 12 + 12 + 13 + 13 + 13 + 13 + 14 + 14 + 14 + 15 + 16 + 16 + 19 = 208$

- $\bar{x} =$ Sum of data$/n = 208/15 \approx 13.87$ minutes

- **Median:**
  - Position of median = $(n+1)/2 = (15+1)/2 = 8^{th}$ position
  - The 8^{th} value in the sorted data is 13.
  - Median = 13 minutes

- **Mode:**
  - The value that appears most frequently.
  - The value 13 appears 4 times, which is more than any other value.
  - Mode = 13 minutes

- **D$_8$ (8th Decile):**
  - Position of $D_k = k(n+1)/10$
  - Position of $D_8 = 8(15+1)/10 = 8(16)/10 = 128/10 = 12.8^{th}$ position
  - This means D_8 is between the 12^{th} and 13^{th} values.
  - 12^{th} value = 15
  - 13^{th} value = 16
  - $D_8 = 12^{th}$ value $+ 0.8 \times (13^{th}$ value $- 12^{th}$ value)
  - $D_8 = 15 + 0.8 \times (16 - 15) = 15 + 0.8 \times 1 = 15.8$ minutes

- **P$_{35}$ (35th Percentile):**
  - Position of $P_k = k(n+1)/100$
  - Position of $P_{35} = 35(15+1)/100 = 35(16)/100 = 560/100 = 5.6^{th}$ position

- This means P_{35} is between the 5^{th} and 6^{th} values.
- 5^{th} value = 13
- 6^{th} value = 13
- $P_{35} = 5^{th}$ value $+ 0.6 \times (6^{th}$ value $- 5^{th}$ value)
- $P_{35} = 13 + 0.6 \times (13 - 13) = 13 + 0.6 \times 0 = 13$ minutes

- **Interquartile Range (IQR):**
  - **First Quartile (Q_1):**
    - Position of $Q_1 = 1(n+1)/4 = 1(15+1)/4 = 16/4 = 4^{th}$ position
    - $Q_1 = 12$ minutes
  - **Third Quartile (Q_3):**
    - Position of $Q_3 = 3(n+1)/4 = 3(15+1)/4 = 48/4 = 12^{th}$ position
    - $Q_3 = 15$ minutes
  - **IQR = Q_3 - Q_1 = 15 - 12 = 3$ minutes

- **Variance (s^2):**
  - Calculate $(x_i - \bar{x})^2$ for each data point:
    - $(10 - 13.87)^2 = (-3.87)^2 = 14.9769$
    - $(12 - 13.87)^2 = (-1.87)^2 = 3.4969$ (appears 3 times)
    - $(16 - 13.87)^2 = (2.13)^2 = 4.5369$ (appears 2 times)
    - $(13 - 13.87)^2 = (-0.87)^2 = 0.7569$ (appears 4 times)

        - $(14 - 13.87)^2 = (0.13)^2 = 0.0169$ (appears 3 times)
        - $(19 - 13.87)^2 = (5.13)^2 = 26.3169$
        - $(15 - 13.87)^2 = (1.13)^2 = 1.2769$
    - Sum of squared differences ($\Sigma(x_i - \bar{x})^2$):
        - $14.9769 + 3 \times 3.4969 + 2 \times 4.5369 + 4 \times 0.7569 + 3 \times 0.0169 + 26.3169 + 1.2769$
        - $= 14.9769 + 10.4907 + 9.0738 + 3.0276 + 0.0507 + 26.3169 + 1.2769 = 65.2135$
    - $s^2 = \Sigma(x_i - \bar{x})^2/(n - 1)$
    - $s^2 = 65.2135/(15 - 1) = 65.2135/14 \approx 4.6581$ minutes²

5. (b) In a study concerning the blood pressure of 60-year-old women with glaucoma, 200 women with glaucoma were randomly selected. The sample mean systolic blood pressure was found to be 140 mm Hg and standard deviation as 25 mm Hg. Calculate 95% confidence interval for the true mean systolic blood pressure among this population of women with glaucoma. Interpret your result if the mean systolic blood pressure for normal women of this age is 120 mm Hg.

    - **Given:**
        - Sample size (n) = 200
        - Sample mean ($\bar{x}$) = 140 mm Hg
        - Sample standard deviation (s) = 25 mm Hg
        - Confidence level = 95%
    - **1. Find the Critical Z-value:**
        - For a 95% confidence interval, $\alpha = 0.05$. The critical Z-value corresponds to $\alpha/2 = 0.025$ in each tail.

  - The Z-value for 95% confidence is $Z_{\alpha/2} = 1.96$.

- **2. Calculate the Standard Error of the Mean (SEM):**
  - $SEM = s/\sqrt{n} = 25/\sqrt{200} = 25/14.142 \approx 1.7678$

- **3. Calculate the Margin of Error (ME):**
  - $ME = Z_{\alpha/2} \times SEM = 1.96 \times 1.7678 \approx 3.4649$

- **4. Calculate the Confidence Interval:**
  - Confidence Interval = $\bar{x} \pm ME$
  - Lower Limit = $140 - 3.4649 = 136.5351$ mm Hg
  - Upper Limit = $140 + 3.4649 = 143.4649$ mm Hg

- **95% Confidence Interval for the true mean systolic blood pressure:** $(136.54$ mm Hg, 143.46 mm Hg$)$

- **Interpretation:**
  - We are 95% confident that the true mean systolic blood pressure for 60-year-old women with glaucoma lies between 136.54 mm Hg and 143.46 mm Hg.
  - **Comparison with normal women:** The mean systolic blood pressure for normal women of this age is given as 120 mm Hg. Since the entire 95% confidence interval (136.54 mm Hg to 143.46 mm Hg) for women with glaucoma is *above* 120 mm Hg, we can conclude that the true mean systolic blood pressure for 60-year-old women with glaucoma is significantly higher than that of normal women of the same age. This suggests that glaucoma may be associated with elevated systolic blood pressure in this age group.

6. (a) For ten states in India, an index of arsenic exposure X and the cancer mortality Y (deaths per 100,000 persons for 1990) were calculated as given in the table.

   - Is there any evidence of a relationship between arsenic exposure and cancer mortality?
   - Find the regression line of X on Y. Estimate the cancer mortality (Y) associated with arsenic exposure index value of 8.0.
   - **Data (from Table 1):**

State	Arsenic Exposure (X)	Cancer Mortality (Y)
A	8.3	210
B	6.4	180
G	3.4	130
H	3.8	170
I	2.6	130
J	11.6	210
K	1.2	120
L	2.5	150
M	1.6	140
N	3.5	150

     - n = 10

   - **Calculations for Pearson Correlation Coefficient (r) to check relationship:**
     - $\Sigma X = 8.3 + 6.4 + 3.4 + 3.8 + 2.6 + 11.6 + 1.2 + 2.5 + 1.6 + 3.5 = 44.9$

- $\Sigma Y = 210 + 180 + 130 + 170 + 130 + 210 + 120 + 150 + 140 + 150 = 1690$
- $\Sigma X^2 = 8.3^2 + 6.4^2 + 3.4^2 + 3.8^2 + 2.6^2 + 11.6^2 + 1.2^2 + 2.5^2 + 1.6^2 + 3.5^2 = 301.99$
- $\Sigma Y^2 = 210^2 + 180^2 + 130^2 + 170^2 + 130^2 + 210^2 + 120^2 + 150^2 + 140^2 + 150^2 = 296900$
- $\Sigma XY = (8.3 \times 210) + (6.4 \times 180) + \cdots + (3.5 \times 150) = 8013$
- $S_{XY} = \Sigma XY - (\Sigma X \Sigma Y)/n = 8013 - (44.9 \times 1690)/10 = 8013 - 7598.1 = 414.9$
- $S_{XX} = \Sigma X^2 - (\Sigma X)^2/n = 301.99 - (44.9)^2/10 = 301.99 - 2016.01/10 = 301.99 - 201.601 = 100.389$
- $S_{YY} = \Sigma Y^2 - (\Sigma Y)^2/n = 296900 - (1690)^2/10 = 296900 - 2856100/10 = 296900 - 285610 = 11290$
- Correlation Coefficient (r) = $S_{XY}/\sqrt{S_{XX} \times S_{YY}}$
  - $r = 414.9/\sqrt{100.389 \times 11290} = 414.9/\sqrt{1133405.81} = 414.9/1064.615 \approx 0.390$

- **Evidence of a relationship:**
  - The Pearson correlation coefficient $r = 0.390$. This indicates a weak to moderate positive linear relationship between arsenic exposure and cancer mortality. Since r is positive, it suggests that as arsenic exposure increases, cancer mortality tends to increase, but the relationship is not very strong. To determine statistical significance, one would typically perform a hypothesis test for r, but based solely on the magnitude, it suggests some evidence of a relationship.

- **Regression line of Y on X (not X on Y, as requested to estimate Y given X):**
  - The question asks for "regression line of X on Y" and then "Estimate the cancer mortality (Y) associated with arsenic exposure index value of 8.0." This implies a regression of Y on X (where Y is the dependent variable and X is the independent variable), which is $Y = a + bX$.
  - **Calculate slope (b):**
    - $b = S_{XY}/S_{XX} = 414.9/100.389 \approx 4.1329$
  - **Calculate Y-intercept (a):**
    - $\bar{X} = 44.9/10 = 4.49$
    - $\bar{Y} = 1690/10 = 169$
    - $a = \bar{Y} - b\bar{X} = 169 - (4.1329 \times 4.49) = 169 - 18.5529 = 150.4471$
  - **Regression Equation (Y on X):**
    - $Y = 150.4471 + 4.1329X$
- **Estimate cancer mortality (Y) for arsenic exposure index (X) of 8.0:**
  - Substitute $X = 8.0$ into the regression equation:
    - $Y = 150.4471 + 4.1329 \times 8.0$
    - $Y = 150.4471 + 33.0632$
    - $Y = 183.5103$
  - **Estimated Cancer Mortality (Y) for X=8.0:** Approximately **183.51 deaths per 100,000 persons**.

7. (b) In a population, the average IQ is 100. A team of scientists want to test a new medication to see if it has any effect on intelligence or not. A sample of 25 participants who have taken the medication for a 'desired duration, recorded a mean IQ of 130 with standard deviation of 20. Did the medication affect intelligence? Justify the selection of the statistical test used to draw your conclusion.

- **Hypothesis Testing:**
  - **Null Hypothesis (H_0):** The medication has no effect on intelligence. The true mean IQ of participants who took the medication is 100 ($\mu = 100$).
  - **Alternative Hypothesis (H_1):** The medication does affect intelligence. The true mean IQ of participants who took the medication is not 100 ($\mu \neq 100$). (This is a two-tailed test as we are checking for "any effect").
- **Given Data:**
  - Population mean under H_0 (μ_0) = 100
  - Sample size (n) = 25
  - Sample mean ($\bar{x}$) = 130
  - Sample standard deviation (s) = 20
- **Selection of Statistical Test:**
  - We should use a **one-sample t-test**.
  - **Justification:**
    - We are comparing a sample mean ($\bar{x}$) to a known population mean (μ_0).
    - The population standard deviation (σ) is *unknown*, and we are using the sample standard deviation (s) as an estimate.

- The sample size ($n = 25$) is less than 30, although for larger samples, the t-distribution approaches the Z-distribution. For $n = 25$, using the t-distribution is appropriate.
- The data is quantitative (IQ scores).

- **Calculation of Test Statistic (t-statistic):**
  - Standard Error of the Mean (SEM) = $s/\sqrt{n} = 20/\sqrt{25} = 20/5 = 4$
  - $t = (\bar{x} - \mu_0)/SEM = (130 - 100)/4 = 30/4 = 7.5$
- **Degrees of Freedom (df):**
  - $df = n - 1 = 25 - 1 = 24$
- **Critical Value (for $\alpha = 0.05$, two-tailed test):**
  - For $df = 24$ and $\alpha = 0.05$ (two-tailed), the critical t-value is approximately ± 2.064.
- **Decision Rule:**
  - Reject H_0 if $|t_{calculated}| > |t_{critical}|$.
  - $|7.5| > |2.064|$
- **Conclusion:**
  - Since the calculated t-statistic (7.5) is much greater than the critical t-value (2.064), we reject the null hypothesis.
  - **Interpretation:** There is statistically significant evidence to conclude that the medication *did* affect intelligence, as the mean IQ of participants who took the medication (130) is significantly different from the average IQ of the general population (100). The medication appears to have increased intelligence.

8. (a) In an anti-malaria campaign in a certain area, quinine was administered to 170 persons out of a total population of 250. The number of fever cases are shown below. Check whether quinine usage is associated with controlling malaria. Check at significance level 0.05. Justify the selection of the statistical test used.

    - **Data (from Table 2):**

Treatment	Fever (F)	No fever (F')	Total
Quinine (Q)	140	30	170
No Quinine (Q')	60	20	80
Total	200	50	250

- **Hypothesis Testing:**
    - **Null Hypothesis (H_0):** Quinine usage is *not* associated with controlling malaria (i.e., there is no association between quinine usage and fever cases).
    - **Alternative Hypothesis (H_1):** Quinine usage *is* associated with controlling malaria (i.e., there is an association between quinine usage and fever cases).
- **Selection of Statistical Test:**
    - We should use a **Chi-Square Test of Independence (χ^2 test)**.
    - **Justification:**
        - We are analyzing the association between two categorical variables: "Quinine Usage" (with two categories: Yes/No) and "Fever Cases" (with two categories: Yes/No).
        - The data is presented in a contingency table.

- We want to determine if there is a statistically significant relationship between these two variables.

- **Calculations for Chi-Square Test:**
  - **Calculate Expected Frequencies (E_{ij}):**
    - E_{ij} = (Row Total × Column Total)/Grand Total
    - $E_{Q,F} = (170 \times 200)/250 = 34000/250 = 136$
    - $E_{Q,F'} = (170 \times 50)/250 = 8500/250 = 34$
    - $E_{Q',F} = (80 \times 200)/250 = 16000/250 = 64$
    - $E_{Q',F'} = (80 \times 50)/250 = 4000/250 = 16$
  - **Observed (O) and Expected (E) Frequencies Table:**

Treatment	Fever (O)	Fever (E)	No fever (O)	No fever (E)
Quinine (Q)	140	136	30	34
No Quinine (Q')	60	64	20	16

  - **Calculate Chi-Square Test Statistic (χ^2_{calc}):**
    - $\chi^2_{calc} = \Sigma(O_{ij} - E_{ij})^2/E_{ij}$
    - $\chi^2_{calc} = (140 - 136)^2/136 + (30 - 34)^2/34 + (60 - 64)^2/64 + (20 - 16)^2/16$
    - $\chi^2_{calc} = (4)^2/136 + (-4)^2/34 + (-4)^2/64 + (4)^2/16$
    - $\chi^2_{calc} = 16/136 + 16/34 + 16/64 + 16/16$
    - $\chi^2_{calc} = 0.1176 + 0.4706 + 0.25 + 1.0 = 1.8382$

- **Degrees of Freedom (df):**

    - $df =$ (number of rows $- 1) \times$ (number of columns $- 1)$
    - $df = (2 - 1) \times (2 - 1) = 1 \times 1 = 1$
  - **Critical Chi-Square Value (χ^2_{crit}):**
    - At $\alpha = 0.05$ and $df = 1$, the critical χ^2 value is 3.841.
  - **Decision Rule:**
    - Reject H_0 if $\chi^2_{calc} > \chi^2_{crit}$.
    - $1.8382 < 3.841$
  - **Conclusion:**
    - Since the calculated χ^2 value (1.8382) is less than the critical χ^2 value (3.841), we *fail to reject* the null hypothesis.
    - **Interpretation:** At a significance level of 0.05, there is **no statistically significant evidence** to conclude that quinine usage is associated with controlling malaria in this area based on the given data. The observed differences in fever cases between the groups could be due to random chance.

9. (b) What is the OR rule of probability? Explain with the help of an example. An office has 60 female and 40 male employees, out of which 24 females and 16 males wear eyeglasses. What is the probability that an employee picked at random:
  - (i) will be a male and wear eyeglasses,
  - (ii) will wear eyeglasses given that the employee is a male,
  - (iii) will be a female given that the employee eyeglasses
  - **OR Rule of Probability (Addition Rule):**

- The OR rule (or Addition Rule) of probability is used to find the probability that at least one of two or more events occurs.
- **For two events A and B:**
  - $P(A \text{ or } B) = P(A \cup B) = P(A) + P(B) - P(A \cap B)$
  - Where $P(A \cap B)$ is the probability that both A and B occur. This term is subtracted to avoid double-counting the outcomes that are common to both events.
- **For mutually exclusive events A and B:**
  - If A and B are mutually exclusive (cannot happen at the same time), then $P(A \cap B) = 0$.
  - In this case, the rule simplifies to: $P(A \text{ or } B) = P(A) + P(B)$.
- **Example:** Consider drawing a single card from a standard 52-card deck.
  - Event A: Drawing a King. $P(A) = 4/52$.
  - Event B: Drawing a Heart. $P(B) = 13/52$.
  - Event A and B (drawing a King of Hearts): $P(A \cap B) = 1/52$.
  - Probability of drawing a King OR a Heart:
    - $P(\text{King or Heart}) = P(\text{King}) + P(\text{Heart}) - P(\text{King of Heart})$
    - $P(\text{King or Heart}) = 4/52 + 13/52 - 1/52 = 17/52 - 1/52 = 16/52 = 4/13$.

- **Office Employee Problem:**

- Total employees = 60 (Female) + 40 (Male) = 100
- Females who wear eyeglasses = 24
- Males who wear eyeglasses = 16
- Let:
  - M = event of being a male
  - F = event of being a female
  - E = event of wearing eyeglasses
  - E' = event of not wearing eyeglasses
- **Table of counts:**

	Eyeglasses (E)	No Eyeglasses (E')	Total
Male (M)	16	24	40
Female (F)	24	36	60
Total	40	60	100

- (i) **Probability that an employee picked at random will be a male and wear eyeglasses:**
  - This is $P(M \cap E)$.
  - Number of males who wear eyeglasses = 16
  - Total employees = 100
  - $P(M \cap E) = 16/100 = 0.16$
- (ii) **Probability that an employee will wear eyeglasses given that the employee is a male:**
  - This is $P(E|M)$, a conditional probability.

- The sample space is reduced to only males (40 employees).
- Number of males who wear eyeglasses = 16
- $P(E|M) =$ (Number of males who wear eyeglasses)/ (Total number of males)
- $P(E|M) = 16/40 = 0.40$

▪ (iii) **Probability that an employee will be a female given that the employee wears eyeglasses:**

- This is $P(F|E)$, a conditional probability.
- The sample space is reduced to only employees who wear eyeglasses (40 employees).
- Number of females who wear eyeglasses = 24
- $P(F|E) =$ (Number of females who wear eyeglasses)/ (Total number of employees who wear eyeglasses)
- $P(F|E) = 24/40 = 0.60$

10. (a) The weights of a certain population of young adult females are approximately normally distributed with a mean of 132 pounds and a standard deviation of 15. Find the probability that a subject selected at random from this population will weigh:

○ Given: Mean (μ) = 132 pounds, Standard Deviation (σ) = 15 pounds.

○ (i) **More than 155 pounds**

▪ Find Z-score for $X = 155$:

- $Z = (X - \mu)/\sigma = (155 - 132)/15 = 23/15 \approx 1.53$

    - Find $P(Z > 1.53)$:
        - Using a Z-table, $P(Z < 1.53) = 0.9370$
        - $P(Z > 1.53) = 1 - P(Z < 1.53) = 1 - 0.9370 = 0.0630$
    - The probability that a subject will weigh more than 155 pounds is **0.0630**.
- (ii) **100 pounds or less**
    - Find Z-score for $X = 100$:
        - $Z = (100 - 132)/15 = -32/15 \approx -2.13$
    - Find $P(Z \leq -2.13)$:
        - Using a Z-table, $P(Z \leq -2.13) = 0.0166$
    - The probability that a subject will weigh 100 pounds or less is **0.0166**.

- (iii) **Between 105 and 145 pounds**
    - Find Z-score for $X_1 = 105$:
        - $Z_1 = (105 - 132)/15 = -27/15 = -1.80$
    - Find Z-score for $X_2 = 145$:
        - $Z_2 = (145 - 132)/15 = 13/15 \approx 0.87$
    - Find $P(-1.80 < Z < 0.87)$:
        - $P(Z < 0.87) = 0.8078$
        - $P(Z < -1.80) = 0.0359$
        - $P(-1.80 < Z < 0.87) = P(Z < 0.87) - P(Z < -1.80) = 0.8078 - 0.0359 = 0.7719$

    - The probability that a subject will weigh between 105 and 145 pounds is **0.7719**.

  - (iv) **More than 120 pounds**

    - Find Z-score for $X = 120$:

      - $Z = (120 - 132)/15 = -12/15 = -0.80$

    - Find $P(Z > -0.80)$:

      - Using a Z-table, $P(Z < -0.80) = 0.2119$

      - $P(Z > -0.80) = 1 - P(Z < -0.80) = 1 - 0.2119 = 0.7881$

    - The probability that a subject will weigh more than 120 pounds is **0.7881**.

11. (b) Based on data, an estimate of adults who have hypertension is 24%. If we select a simple random sample of 20 adults, find the probability that the number of people in the sample who have been told that they have hypertension will be:

  - This is a binomial probability problem.

  - Given:

    - Probability of success (hypertension) $p = 0.24$

    - Number of trials (sample size) $n = 20$

    - Probability of failure (no hypertension) $q = 1 - p = 1 - 0.24 = 0.76$

  - The binomial probability formula is $P(X = k) = C(n, k) \times p^k \times q^{(n-k)}$, where $C(n, k) = n!/(k! \times (n - k)!)$.

  - (i) **Exactly three**

    - $k = 3$

  - $P(X = 3) = C(20,3) \times (0.24)^3 \times (0.76)^{(20-3)}$
  - $C(20,3) = 20!/(3! \times 17!) = (20 \times 19 \times 18)/(3 \times 2 \times 1) = 1140$
  - $P(X = 3) = 1140 \times (0.24)^3 \times (0.76)^{17}$
  - $P(X = 3) = 1140 \times 0.013824 \times 0.01336 \approx 0.2109$
  - The probability of exactly three people having hypertension is **0.2109**.

- (ii) **Three or more**
  - $P(X \geq 3) = 1 - P(X < 3) = 1 - [P(X = 0) + P(X = 1) + P(X = 2)]$
  - $P(X = 0) = C(20,0) \times (0.24)^0 \times (0.76)^{20} = 1 \times 1 \times 0.00693 \approx 0.0069$
  - $P(X = 1) = C(20,1) \times (0.24)^1 \times (0.76)^{19} = 20 \times 0.24 \times 0.00912 \approx 0.0438$
  - $P(X = 2) = C(20,2) \times (0.24)^2 \times (0.76)^{18} = 190 \times 0.0576 \times 0.0120 \approx 0.1312$
  - $P(X < 3) = 0.0069 + 0.0438 + 0.1312 = 0.1819$
  - $P(X \geq 3) = 1 - 0.1819 = 0.8181$
  - The probability of three or more people having hypertension is **0.8181**.

- (iii) **Fewer than three**
  - $P(X < 3) = P(X = 0) + P(X = 1) + P(X = 2)$
  - From the calculations above:
    - $P(X = 0) \approx 0.0069$
    - $P(X = 1) \approx 0.0438$

- $P(X = 2) \approx 0.1312$

- $P(X < 3) = 0.0069 + 0.0438 + 0.1312 = 0.1819$
- The probability of fewer than three people having hypertension is **0.1819**.

- (iv) **Between three and seven (both inclusive)**
  - $P(3 \le X \le 7) = P(X = 3) + P(X = 4) + P(X = 5) + P(X = 6) + P(X = 7)$
  - We already calculated $P(X = 3) \approx 0.2109$.
  - $P(X = 4) = C(20,4) \times (0.24)^4 \times (0.76)^{16} = 4845 \times 0.00331776 \times 0.01758 \approx 0.2223$
  - $P(X = 5) = C(20,5) \times (0.24)^5 \times (0.76)^{15} = 15504 \times 0.00079626 \times 0.02313 \approx 0.1706$
  - $P(X = 6) = C(20,6) \times (0.24)^6 \times (0.76)^{14} = 38760 \times 0.0001911 \times 0.03043 \approx 0.1009$
  - $P(X = 7) = C(20,7) \times (0.24)^7 \times (0.76)^{13} = 77520 \times 0.00004586 \times 0.04004 \approx 0.0435$
  - $P(3 \le X \le 7) = 0.2109 + 0.2223 + 0.1706 + 0.1009 + 0.0435 = 0.7482$
  - The probability of between three and seven people (inclusive) having hypertension is **0.7482**.

12. (a) A research team wished to evaluate a proposed screening test for a neurological disease. Assume that the rate of the disease in the general population is 10%. The test was given to a random sample of 400 patients with the disease and an independent random sample of 600 patients without symptoms of the disease. The two samples were drawn from populations of subjects who were 75 years or older. The results are as follows: Calculate the specificity and

sensitivity of the test. What is the predictive value positive of the symptom and the predictive value negative of the symptom?

- **Data (from Table 3):**

	Disease (D)	No Disease (D')	Total
Positive Result (T+)	350	10	360
Negative Result (T-)	50	590	640
Total	400	600	1000

- **Definitions:**
  - **True Positive (TP):** Positive test, has disease (350)
  - **False Positive (FP):** Positive test, no disease (10)
  - **False Negative (FN):** Negative test, has disease (50)
  - **True Negative (TN):** Negative test, no disease (590)

- **1. Sensitivity:**
  - **Definition:** The probability that a test correctly identifies those with the disease (True Positive Rate).
  - Sensitivity = $TP/(TP + FN)$
  - Sensitivity = $350/(350 + 50) = 350/400 = 0.875$ or **87.5%**
- **2. Specificity:**
  - **Definition:** The probability that a test correctly identifies those without the disease (True Negative Rate).
  - Specificity = $TN/(TN + FP)$
  - Specificity = $590/(590 + 10) = 590/600 \approx 0.9833$ or **98.33%**
- **3. Predictive Value Positive (PVP):**

- **Definition:** The probability that a subject with a positive test result actually has the disease. This depends on the prevalence of the disease in the general population.
- First, we need to adjust the table based on the population prevalence of 10%.
  - Assume a population of 1000 people aged 75+.
  - Number with disease = 10% of $1000 = 100$
  - Number without disease = 90% of $1000 = 900$
  - Using Sensitivity (0.875) and Specificity (0.9833):
    - **Positive Test (T+):**
      - True Positives (TP) = $0.875 \times$ Diseased $= 0.875 \times 100 = 87.5$
      - False Positives (FP) = $(1 -$ Specificity$) \times$ Non-Diseased $= (1 - 0.9833) \times 900 = 0.0167 \times 900 \approx 15.03$
      - Total Positives = $87.5 + 15.03 = 102.53$
    - **Negative Test (T-):**
      - False Negatives (FN) = $(1 -$ Sensitivity$) \times$ Diseased $= (1 - 0.875) \times 100 = 0.125 \times 100 = 12.5$
      - True Negatives (TN) = Specificity $\times$ Non-Diseased $= 0.9833 \times 900 = 884.97$
      - Total Negatives = $12.5 + 884.97 = 897.47$
    - **New table based on general population prevalence:**

	Disease (D)	No Disease (D')	Total
Positive Result (T+)	87.5	15.03	102.53
Negative Result (T-)	12.5	884.97	897.47
Total	100	900	1000

- PVP = $TP/(TP + FP)$
- PVP = $87.5/102.53 \approx 0.8534$ or **85.34%**

- **4. Predictive Value Negative (PVN):**
  - **Definition:** The probability that a subject with a negative test result actually does not have the disease.
  - PVN = $TN/(TN + FN)$
  - PVN = $884.97/(884.97 + 12.5) = 884.97/897.47 \approx 0.9861$ or **98.61%**

13. (b) In an air pollution study, a random sample of 200 families was selected from each of two communities. A person in each family was asked whether or not anyone in the family was affected by air pollution. Based on the following responses, can the researchers conclude that the two communities differ with respect to the variable of interest?

- **Data (from Table 4):**

	Community I (C1)	Community II (C2)	Total
Affected by Air Pollution (A)	8	43	51

	Community I (C1)	Community II (C2)	Total
Not Affected (A')	157	119	276
Total	165	162	327

- **Correction to Table 4 provided in prompt context:** The total in the provided table is wrong. Let's re-calculate from the individual values, assuming the individual values are correct.
    - Community I affected: 8, Not affected: 157. Total Community I = $8 + 157 = 165$
    - Community II affected: 43, Not affected: 119. Total Community II = $43 + 119 = 162$
    - Total Affected: $8 + 43 = 51$
    - Total Not Affected: $157 + 119 = 276$
    - Grand Total = $165 + 162 = 327$. (The problem states "A random sample of 200 families was selected from each of two communities", implying total of 400, but the numbers given sum to 327. I will proceed with the numbers given in the table, assuming it's a sub-sample or the "200 families" was an error in the question phrasing itself if the table numbers are to be used.)
    - Let's assume the question meant 200 families were selected, and the *given data* is the subset of results. Given the structure of the question, a Chi-Square Test of Independence is appropriate.
    - **Revised Corrected Table Based on Input Data:**

	Community I (C1)	Community II (C2)	Total
Affected by Air Pollution (A)	8	43	51
Not Affected (A')	157	119	276
Total	165	162	327

- **Hypothesis Testing:**
  - **Null Hypothesis (H_0):** There is no difference between the two communities with respect to families affected by air pollution (i.e., being affected by air pollution is independent of the community).
  - **Alternative Hypothesis (H_1):** There is a difference between the two communities with respect to families affected by air pollution (i.e., there is an association between being affected by air pollution and the community).

- **Selection of Statistical Test:**
  - A **Chi-Square Test of Independence (χ^2 test)** is appropriate because we are comparing two categorical variables (Community and Air Pollution Effect) to see if they are associated.
- **Calculations for Chi-Square Test:**
  - **Degrees of Freedom (df):**
    - $df = (\text{number of rows} - 1) \times (\text{number of columns} - 1)$
    - $df = (2 - 1) \times (2 - 1) = 1 \times 1 = 1$
  - **Expected Frequencies (E_{ij}):**

- $E_{C1,A}$ = (Row A Total × Col C1 Total)/ Grand Total $= (51 \times 165)/327 \approx 25.75$
- $E_{C1,A'}$ = (Row A’ Total × Col C1 Total)/ Grand Total $= (276 \times 165)/327 \approx 139.25$
- $E_{C2,A}$ = (Row A Total × Col C2 Total)/ Grand Total $= (51 \times 162)/327 \approx 25.25$
- $E_{C2,A'}$ = (Row A’ Total × Col C2 Total)/ Grand Total $= (276 \times 162)/327 \approx 136.75$

- **Observed (O) and Expected (E) Frequencies Table:**

	Community I (O)	Community I (E)	Community II (O)	Community II (E)
Affected by Air Pollution (A)	8	25.75	43	25.25
Not Affected (A')	157	139.25	119	136.75

- **Calculate Chi-Square Test Statistic (χ^2_{calc}):**
  - $\chi^2_{calc} = (8 - 25.75)^2/25.75 + (157 - 139.25)^2/139.25 + (43 - 25.25)^2/25.25 + (119 - 136.75)^2/136.75$
  - $\chi^2_{calc} = (-17.75)^2/25.75 + (17.75)^2/139.25 + (17.75)^2/25.25 + (-17.75)^2/136.75$

- $\chi^2_{calc} = 315.0625/25.75 + 315.0625/139.25 + 315.0625/25.25 + 315.0625/136.75$
- $\chi^2_{calc} \approx 12.23 + 2.26 + 12.48 + 2.30 \approx 29.27$

- **Critical Chi-Square Value (χ^2_{crit}):**
  - At significance level $\alpha = 0.05$ and $df = 1$, the critical χ^2 value is 3.841.
- **Decision Rule:**
  - Reject H_0 if $\chi^2_{calc} > \chi^2_{crit}$.
  - $29.27 > 3.841$
- **Conclusion:**
  - Since the calculated χ^2 value (29.27) is much greater than the critical χ^2 value (3.841), we **reject the null hypothesis**.
  - **Interpretation:** At a significance level of 0.05, the researchers can conclude that the two communities **do differ significantly** with respect to the proportion of families affected by air pollution. Community II appears to have a higher proportion of families affected by air pollution compared to Community I.