**CSC 362E REVISION PAPER**

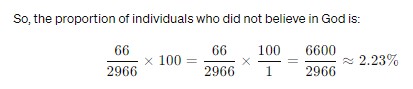
**QUESTION ONE [COMPULSORY]**

[30 MARKS]

**a) In the 2006 survey, 2,966 individuals reported their belief in the existence of God. Sixty- six respondents said they did not believe in God. What proportion did not believe?**

**[2 marks]**

To find the proportion of individuals who did not believe in God, you need to divide the number of respondents who did not believe (66) by the total number of respondents (2966) and then multiply by 100 to convert it to a percentage.

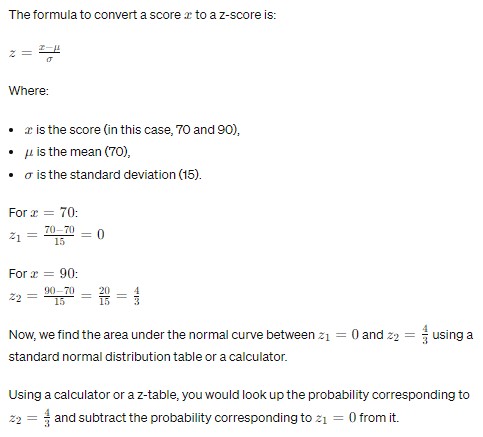


Therefore, approximately 2.23% of the respondents did not believe in God.

**b) If a statistics class has a distribution of grades with a mean of 70 and standard deviation of**

**15, what is the probability of a student scoring between a 70 and 90? [4 marks]**

To find the probability of a student scoring between 70 and 90 in a normal distribution with a mean of 70 and a standard deviation of 15, we'll first convert these scores to z-scores and then use a z-table or calculator to find the area under the normal curve between the corresponding z-scores.



This probability represents the likelihood of a student scoring between 70 and 90.

**c) If a statistics class has a distribution of grades with a mean of 70 and standard deviation of 15, what is the z-score associated with a raw score of 85?**

**[3 marks]**

85-70/15 = 1

So, the z-score associated with a raw score of 85 is 1.

**d) Suppose we have data on employees who earn an average wage of KES.2000.00 per hour. If the standard deviation of wages is KES400.00 per hour, what percentage of workers earn between KES. 1900.00 and KES.2500.00? [4 marks]**

*μ* as the mean wage (KES 2000.00),

*σ* as the standard deviation (KES 400.00),

1*x*1​ as the lower wage boundary (KES 1900.00),

*x*2​ as the upper wage boundary (KES 2500.00).

z= x- *μ/ σ*

for x =1900.00

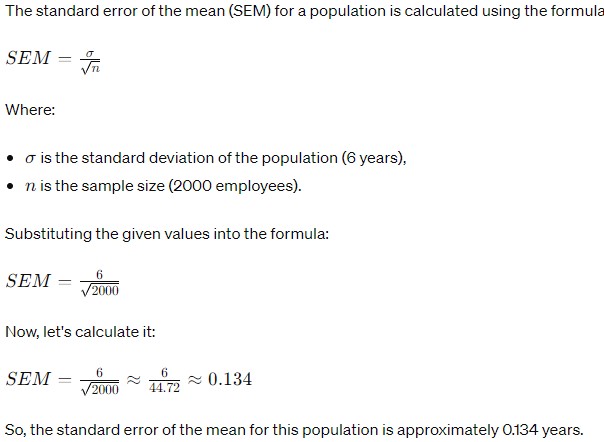
z1 = 1900.00-2000.00/400.00 ANS **= -0.25**

for x2 = 2500.00

z2 = 2500.00-2000.00/400.00 ANS **= 1.25**

are under the normal curve =   
�1=−0.25*z*1​=−0.25 and �2=1.25*z*2​=1.25 using a standard normal distribution table or a calculator

**e) Suppose a corporation has 2,000 employees with an average age of 36. If the standard deviation of age is 6 years, what is the standard error of the mean for this population? [3 marks]**



**1) Use the following data to answer questions i-iv:**

**15, 29, 20, 14, 20, 30, 49, 20, 63**

1. **What is the mode for the distribution?**

**Mode = 20**

1. **What is the median of the above distribution?**

**Median = 20**

**What is the mean of the above distribution?**

**Mean ≈ 28.89**

**iv. Is this distribution symmetrical, negatively skewed, or positively you determine this? - skewed? How did you determine this?**

To determine whether the distribution is symmetrical, negatively skewed, or positively skewed, we can compare the mean, median, and mode.

- If the mean, median, and mode are approximately equal, the distribution is symmetrical.

- If the mean is less than the median, the distribution is negatively skewed.

- If the mean is greater than the median, the distribution is positively skewed.

Since the mean is greater than the median and mode, the distribution is **positively skewed.**

**g) Discuss the primary difference between a parameter and a statistic. [3 marks]**

The primary difference between a parameter and a statistic lies in what they represent and how they are calculated:

**1. Parameter:**

- A parameter is a descriptive measure of a population.

- It represents a characteristic or feature of the entire population.

- Parameters are typically denoted using Greek letters (e.g., μ for population mean, σ for population standard deviation).

- Parameters are often unknown and are estimated using statistics.

- Examples of parameters include the population mean, population standard deviation, population proportion, etc.

- Parameters are fixed values for a specific population and do not change unless the population changes.

**2. Statistic:**

- A statistic is a descriptive measure of a sample.

- It represents a characteristic or feature of a subset of the population (i.e., the sample).

- Statistics are typically denoted using Roman letters (e.g., x̄ for sample mean, s for sample standard deviation).

- Statistics are calculated from sample data and are used to estimate population parameters.

- Examples of statistics include the sample mean, sample standard deviation, sample proportion, etc.

- Statistics can vary from sample to sample, as they depend on the specific individuals or elements included in the sample.

**h) Identify the independent and the dependent variable, as well as the unit of analysis, in the following hypothesis: Urban residents are more likely than rural residents to be overweight:**

- Independent Variable: Type of residency (urban vs. rural)

- Dependent Variable: Likelihood of being overweight

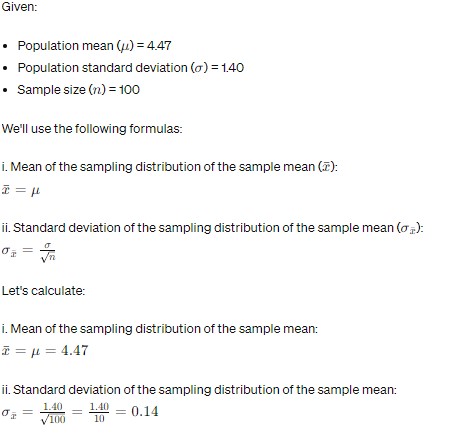
- Unit of Analysis: Individual residents

QUESTION TWO:

[20 MARKS]

a) The sampling distribution of sample means is from a highly skewed population with μ=4.47 and 6-1.40. For repeated random samples of size 100 from this population:

i. Find the mean and standard deviation of the sampling distribution of the sample mean. [5 marks]



1. **Explain why the sampling distribution of the sample mean is bell-shaped, even though the population is highly skewed**

The sampling distribution of the sample mean tends to become approximately normally distributed regardless of the shape of the population distribution. This is due to the Central Limit Theorem (CLT), which states that as the sample size increases, the sampling distribution of the sample mean approaches a normal distribution.

Even though the population distribution is highly skewed, the sampling distribution of the sample mean becomes bell-shaped because:

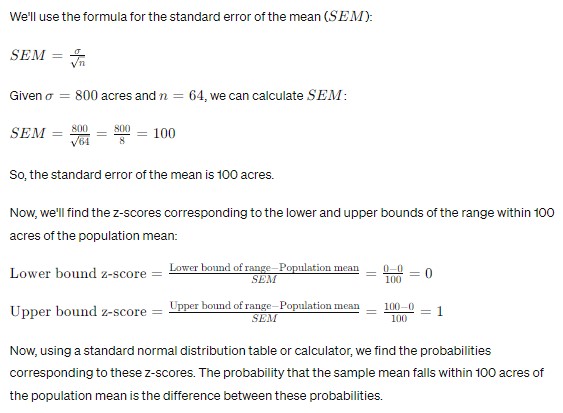
- The CLT applies to any population distribution, including highly skewed ones.

- With a large enough sample size (in this case, \( n = 100 \)), the individual sample means tend to cluster around the population mean, resulting in a bell-shaped distribution.

- The bell-shaped distribution of the sampling distribution of the sample mean is symmetric, with the mean of the sampling distribution equal to the population mean (\( \mu \)) and the standard deviation of the sampling distribution (\( \sigma\_{\bar{x}} \)) getting smaller as the sample size increases.

**b) To estimate the mean acreage of ranches in a certain province a researcher plans to obtain the acreage for a random sample of 64 farms. Results from an earlier study suggest that 800 acres is a reasonable guess for the standard deviation of the ranch size.**

**i. Find the probability that the sample mean acreage falls within 100 acres of the population mean acreage. [5 marks]**

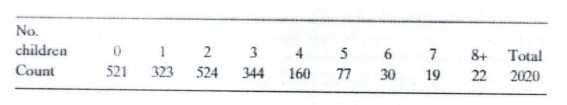


**ii. If the researcher can increase n above 64, will the probability that the sample mean falls within 100 acres of the population mean increase or decrease? Why? [6 marks]**

ii. If the researcher can increase \( n \) above 64, the probability that the sample mean falls within 100 acres of the population mean will increase. This is because, as \( n \) increases, the standard error of the mean (\( SEM \)) decreases. A smaller \( SEM \) means that the range within which the sample mean is likely to fall around the population mean becomes narrower, leading to a higher probability of the sample mean falling within 100 acres of the population mean.

**QUESTION THREE [20 MARKS]**

a) In a certain survey, 2020 respondents answered the question, "How many children have you ever had?" The results were



**1. Is the variable, number of children, categorical or quantitative?**

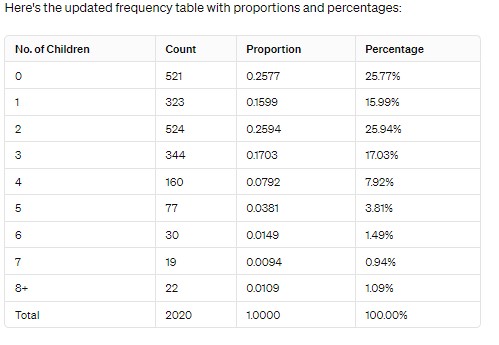
**[1 mark]**

The variable "number of children" is quantitative because it represents a count or measurement of a characteristic.

**ii. Is the variable, number of children, discrete or continuous?**

The variable "number of children" is **discrete** because it consists of distinct, separate categories (0, 1, 2, 3, 4, 5, 6, 7, 8+ children). Each category represents a whole number count of children, and there cannot be fractions or decimals of children.

**Add proportions and percentages to this frequency table. [6 marks]**



1. **Which response is the mode? [1 mark]**

The mode is the response with the highest frequency. Looking at the frequency table, the mode is the category with the highest count, which is "2 children" with 524 respondents.

**b) A historian wants to estimate the average age at marriage of women in New England in the early 19th century. Within her state archives she finds marriage records for the years 1800- 1820, which she treats as a sample of all marriage records from the early 19th century. The average age of the women in the records is 24.1 years. Using the appropriate statistical method, she estimates that the average age of brides in early 19th-century New England was between 23.5 and 24.7.**

**Which part of this example gives a descriptive summary of the data? [3 marks]**

The part of the example that gives a descriptive summary of the data is when the historian states that the average age of the women in the marriage records is 24.1 years. This statement provides a summary statistic (average) of the sample data.

**ii. Which part of this example draws an inference about a population? [3 marks]**

The part of the example that draws an inference about a population is when the historian estimates that the average age of brides in early 19th-century New England was between 23.5 and 24.7 years. This estimation is based on the sample data and is used to make a conclusion about the entire population of brides in early 19th-century New England.

**iii. To what population does the inference in part b refer? [3 marks]**

The inference in part b refers to the population of brides in early 19th-century New England. This is the group about which the historian is making conclusions based on the sample data.

**iv. The average age of the sample was 24.1 years. Is 24.1 a statistic or a parameter. [2 marks]**

The average age of 24.1 years is a statistic. A statistic is a characteristic of a sample, in this case, the average age of the women in the marriage records from 1800-1820. If the historian had access to data for all brides in early 19th-century New England, the average age would be a parameter, which describes a characteristic of the entire population.

**QUESTION FOUR [20 MARKS]**

**a) Distinguish between the following pairs of variables:**

**[4 marks]**

**i. Nominal variable and ordinal variable**

Nominal variables are categorical variables where the categories represent names, labels, or qualities with no inherent order or ranking. These categories are typically mutually exclusive and exhaustive. Examples include gender (male, female), marital status (single, married, divorced), and eye color (blue, brown, green).

**Ordinal variable:** Ordinal variables are categorical variables where the categories have a natural order or ranking. The categories represent qualitative differences that can be ranked or ordered but do not have a consistent magnitude between them. Examples include education level (elementary, high school, bachelor's, master's, Ph.D.), socioeconomic status (low, middle, high), and survey ratings (poor, fair, good, excellent).

**ii. Categorical variable and quantitative variable**

- **Categorical variable:** A categorical variable is a variable that represents categories or groups. It can take on values that are names, labels, or other qualitative descriptors. Categorical variables are often used to classify data into distinct groups or to represent qualitative differences between observations. Examples include gender, ethnicity, marital status, and type of car.

- **Quantitative variable:** A quantitative variable is a variable that represents numerical quantities or measurements. It can take on values that are numerical and have a meaningful order or magnitude. Quantitative variables are often used to represent quantities, measurements, or counts. Examples include age, height, weight, income, and temperature.

**b) What are the main ways of gathering data, and what are their advantages and disadvantages? [4 marks]**

There are several main ways of gathering data, each with its own advantages and disadvantages:

**1. Surveys and Questionnaires:**

**Advantages:**

- Surveys and questionnaires are flexible and can be administered in various formats, such as online, by mail, or in-person.

- They allow for standardized data collection, making it easier to compare responses.

- Surveys and questionnaires can collect data from a large number of respondents relatively quickly and cost-effectively.

**Disadvantages:**

- Response rates can be low, leading to potential non-response bias.

- Responses may be subject to social desirability bias, where respondents provide answers they think are socially acceptable rather than true reflections of their beliefs or behaviors.

- Designing effective surveys and questionnaires requires careful attention to wording, formatting, and question order to avoid bias and ensure clarity.

**2. Interviews:**

**Advantages:**

- Interviews allow for in-depth exploration of topics and follow-up questions, providing rich qualitative data.

- They can be more flexible than surveys, allowing interviewers to adapt their approach based on the respondent's answers.

- Interviews can build rapport and trust between the interviewer and respondent, leading to more honest and detailed responses.

**Disadvantages:**

- Interviews can be time-consuming and resource-intensive, especially if conducted in-person.

- Interviewer bias may occur if interviewers unintentionally influence respondents' answers through their tone, body language, or wording of questions.

- The qualitative nature of interview data can make analysis more subjective and time-consuming compared to quantitative data.

**3. Observational Studies:**

**Advantages:**

- Observational studies allow researchers to directly observe and record behaviors, actions, or events in natural settings, providing a more accurate picture of real-world behavior.

- They can capture data on phenomena that may be difficult to measure through other methods, such as non-verbal communication or environmental factors.

- Observational studies can be useful for generating hypotheses and exploring relationships between variables.

**Disadvantages:**

- Observer bias may occur if researchers unintentionally interpret or record data in a way that aligns with their expectations or beliefs.

- Observer effect, where participants modify their behavior because they know they are being observed, can impact the validity of the data.

- Observational studies may lack control over extraneous variables, making it challenging to establish causality.

**4. Experiments:**

**Advantages:**

- Experiments allow researchers to manipulate variables and control conditions to establish cause-and-effect relationships.

- They provide a high level of control over extraneous variables, enhancing the internal validity of the study.

- Experiments can be replicated to test the robustness of findings and ensure their generalizability.

**Disadvantages:**

- Experimental settings may lack ecological validity, meaning findings may not generalize to real-world settings.

- Ethical considerations may limit the types of experiments that can be conducted, particularly those involving deception or harm to participants.

- Experiments can be resource-intensive and time-consuming to set up and conduct, particularly for large-scale studies.

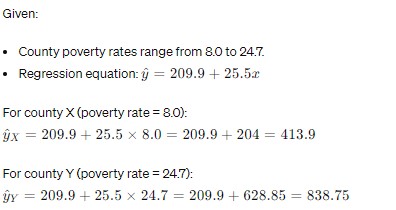
**c) An analysis of data for the 47 Kenyan counties on y = violent crime rate (measured as number of violent crimes per 100,000 people in the county) and x = poverty rate (percent of people in the county living at or below the poverty level) yielded the regression equation,**

**ŷ 209.9+ 25.5x.**

**i. Interpret the slope.** **[3 marks]**

The slope of the regression equation represents the change in the dependent variable (violent crime rate, y) for a one-unit change in the independent variable (poverty rate, x). In this case, the slope is 25.5, which means that for every one-unit increase in the poverty rate, the predicted violent crime rate increases by 25.5 violent crimes per 100,000 people.

**ii. The county poverty rates ranged from 8.0 (for county X) to 24.7 (for county Y). Over this range, find the range of predicted values for the violent crime rate. [5 marks]**



So, over the range of poverty rates from 8.0 to 24.7, the predicted violent crime rate ranges from approximately 413.9 to 838.75 violent crimes per 100,000 people.

**Would the correlation between these variables be positive or negative? Why? [4 marks]**

The correlation between the violent crime rate and poverty rate would be positive. This is because the regression equation has a positive slope coefficient (25.5). A positive slope indicates that as the poverty rate increases, the predicted violent crime rate also increases. Therefore, there is a positive relationship between these variables, and the correlation would be positive

2021/22 okumu wilkistar sec

Merry madline

97