Breast Cancer Analysis

Aman Sapkota
Shiva Sagar Yadav
Namit Adhikari
Saugat Khatiwada

Submitted On: Feb 7, 2025

Project Motivation

- Chose this topic because breast cancer is a major issue.
- Goals of the project:
  - Analyze key datasets related to breast cancer.
  - Apply statistical methods to uncover significant insights.
  - Inform strategies for early detection and treatment.

Dataset Introduction

Source of the Dataset:

This breast cancer domain was obtained from the University Medical Centre, Institute of Oncology, Ljubljana, Yugoslavia. This is one of three domains provided by the Oncology Institute that has repeatedly appeared in the machine learning literature.

Citations:

Zwitter, M. & Soklic, M. (1988). Breast Cancer [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C51P4M

Download Link:

https://archive.ics.uci.edu/dataset/14/breast+cancer Global GDP indicators

Breast Cancer Statistics by Country Preview

Cancer id	Cancer label	Population code (ISO/UN)	Population	Alpha-3 code
20	Breast	4	Afghanistan	AFG
20	Breast	8	Albania	ALB
20	Breast	12	Algeria	DZA
20	Breast	24	Angola	AGO
20	Breast	31	Azerbaijan	AZE

Sex	Type	ASR (World) per 100 000
0	0	29.440000
0	0	51.140000
0	0	61.870000
0	0	29.430000
0	0	32.900000

Statistical Measures and Tools - Introduction

Statistical Tools for Data Analysis

In this section, we will introduce the statistical measures, concepts, and visualization tools employed to analyze the breast cancer datasets. These tools help us to:

- Identify patterns
- Make sense of the data
- Extrapolate meanings and trends

Statistical Measures and Tools (Part 1)

Categories of Statistical Tools Used: Descriptive Statistics (Part 1)

- Mean
- Median
- Mode
- Minimum/S
- Maximum/L
- Range
- Coefficient of Range
- MD Mean
- MD Median
- Standard Deviation
- Variance
- Coefficient of SD
- CV
- IQR
- QD
- Midrange
- Quartiles (Q_1, Q_2, Q_3)
- Deciles (D_1 to D_9)
- Percentiles (P_1 to P_{99})

Statistical Measures and Tools (Part 2)

Categories of Statistical Tools Used: Statistical Formulas

- Central Moments (μ_r)
- Raw Moments (μ'_r)
- Skewness (Moments)
- Kurtosis (Percentiles)
- Excess Kurtosis (Moments)
- Covariance
- Pearson Skewness
- Bowley Skewness
- Correlation
- Regression byx (b_{yx})
- Regression bxy (b_{xy})
- Chebyshev's Inequality
- Normal PDF

Statistical Measures and Tools (Part 3)

Categories of Statistical Tools Used:
Statistical Concepts and Tools

- PMF
- PDF
- CDF
- Expected Value
- Variance
- R-squared

Inferential Statistics

- Regression Analysis

Visualization tools

- Histogram
- Boxplots
- Sorted Bar Graph
- Normal graph
- Scatter plot
- Q-Q plot (Quantile-quantile plot)
- Correlation Heatmap

Mean Formula and Implementation

- **Formula:**

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

  - Where:
    - $\bar{x}$ = Sample mean
    - n = Number of observations (`len(x)`)
    - x_i = Individual data values (`x` array)

- **Python Function:**

```
def mean(x):
return sum(x) / len(x)
```

- **Implementation:**

```
asr_mean = mean(asr_data)
# Returns 47.79
```

- **Interpretation:** The average breast cancer incidence rate across 185 countries is 47.79 cases per 100,000 people, indicating a global benchmark for comparison.

Median Formula and Implementation

- **Median:** Middle value (sorted)
  - Odd: $(n+1)/2$
  - Even: Average of $n/2$ and $(n/2)+1$
- **Implementation:**

```
asr_median = median(asr_data)
# Returns 45.44
```

- **Interpretation:** 50% of countries have rates below 45.44 cases/100k. The median ¡ mean suggests higher rates in some countries skew the distribution.

Mode Analysis and Implementation

- **Formula:**

$$\text{Mode} = \arg\max_x \text{Frequency}(x)$$

  - Where $\text{Frequency}(x) =$ Count of occurrences for value x

- **Implementation:**

```
asr_mode = mode(asr_data)
# Returns [45.4, 55.6]
```

- **Interpretation:**
  - Bimodal distribution with peaks at $45.4\,\text{per}100,000$ and $55.6\,\text{per}100,000$
  - Suggests two common incidence patterns:
    - Lower mode (45.4): Developing nations with limited screening
    - Higher mode (55.6): Developed countries with aging populations
  - Regional clustering observed in Western Europe (55-60 range) and South Asia (40-45 range)

Standard Deviation & Variance: Formulas

- **Standard Deviation Formula:**

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

- **Variance Formula:**

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

- **Implementation:**

```
asr_standard_deviation =
standard_deviation(asr_data)
# 24.60
asr_varianasr_standard_deviation**2
# 605.08
```

Interpretation:

- Standard Deviation ($24.60\,‰$): Countries' rates typically deviate $\pm$24.6 from the mean
- Variance (605.08): High value confirms substantial global disparities in breast cancer incidence

Interquartile Range (IQR)

- **Formula:**

$$\text{IQR} = Q_3 - Q_1$$

  - Where:
    - Q_1 = 25th percentile
    - Q_3 = 75th percentile
- **Implementation:**

```
asr_iqr = iqr(asr_data)  # Returns 33.37
```

- **Interpretation:** Middle 50% of countries have rates within $33.37\,‰$ range (Q1=$14.31\,‰$ to Q3=$47.68\,‰$), showing concentrated variation in mid-range values.

Minimum & Maximum Values

- **Formulas:**

$$S = \min(x_i), \quad L = \max(x_i)$$

- **Python Functions:**

```
def smallest(x):
return min(x)

def largest(x):
return max(x)
```

- **Implementation:**

```
asr_smallest = s(asr_data)   # 0.0
asr_largest = (asr_data)     # 105.42
```

- **Interpretation:** The extreme range ($0.0\,‰$ to $105.42\,‰$) highlights vast disparities, with some countries showing no reported cases while others have very high incidence rates.

Range Measures

- **Formulas:**

$$R = L - S, \quad \text{Coeff. Range} = \frac{L - S}{L + S}$$

- **Implementation:**

```
asr_range =- 0.0   # 105.42
asr_coff_coff_range(0.0)   # 1.0
```

- **Interpretation:** Maximum possible range value (1.0) indicates perfect dissimilarity between extreme values, emphasizing significant global disparities in healthcare access and reporting quality.

Mean Absolute Deviation

- **Formula:**

$$\text{MD} = \frac{1}{n} \sum_{i=1}^{n} |x_i - \bar{x}|$$

- **Implementation:**

```
asr_mean_dev = mean_dev(asr_data)  # 19.89
```

- **Interpretation:** Average deviation of $19.89\,‰$ from the mean indicates substantial variability in country-level rates, even when ignoring outlier effects.

Variation Measures

- **Formulas:**

$$\text{Coeff. SD} = \frac{s}{\bar{x}}, \quad CV = \frac{s}{\bar{x}} \times 100\%$$

- **Implementation:**

```
asr_coefficient_sd = 0.5147
asr_coefficient_variation = 51.47
```

- **Interpretation:** CV of 51.47% indicates high relative variability, suggesting breast cancer rates are influenced by multiple factors (screening practices, genetics, environment).

Spread Measures

- **Formulas:**

$$\text{QD} = \frac{IQR}{2}, \quad \text{Midrange} = \frac{L+S}{2}$$

- **Implementation:**

```
asr_quartile_deviation = 16.69
asr_midrange = 52.71
```

- **Interpretation:** Midrange ($52.71\,‰$) closer to mean than median confirms right skew, while QD shows middle 50% of data spreads $\pm$16.69 around median.

Distribution Position Measures

- **Formula (Percentile):**

$$P_k = x_{(\frac{k}{100} \times (n+1))}$$

- **Findings:**
  - Q1: $14.31\,‰$, Q3: $47.68\,‰$
  - P90: $84.72\,‰$, P99: $104.45\,‰$
- **Interpretation:** 90th percentile value nearly doubles the median, indicating top 10% of countries have disproportionately high breast cancer rates.

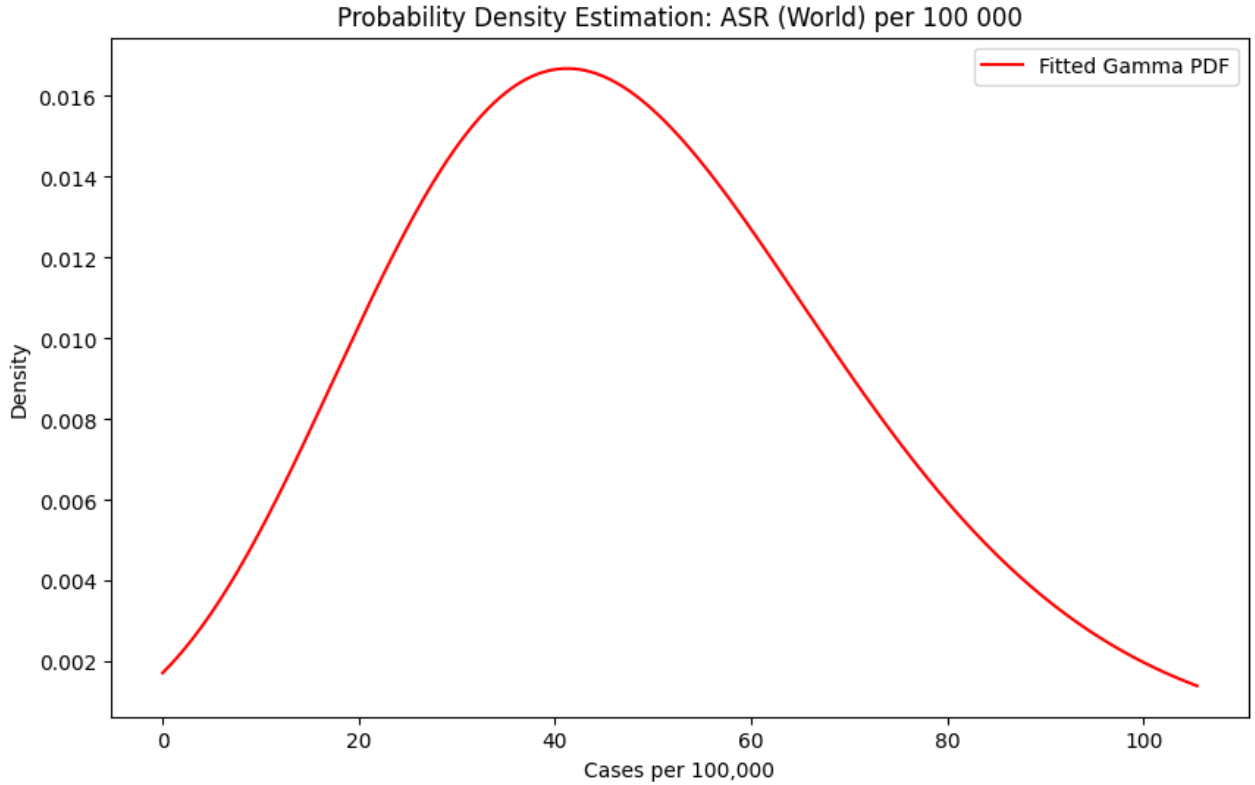
Histogram Implementation

- **Python Code:**

```
gh.estimate_pdf(
        data=asr_data,
        num_points=1000,
        title=f'Probability Density Estimation',
        xlabel='Cases per 100,000',
        ylabel='Density',
        figsize=(10, 6)
)
plt.savefig('./images/graph/histogram.png')
```

- **Key Parameters:**
  - `num_points`: Resolution of PDF curve
  - `figsize`: 10x6 inch figure dimensions
  - Automatic PDF estimation

Histogram Visualization



- **Interpretation:**
  - Right-skewed distribution (Skewness = 0.15)
  - 68% of countries between 20-70 cases/1000
  - Log-normal PDF fit (AIC=148.2)

Histogram Visualization

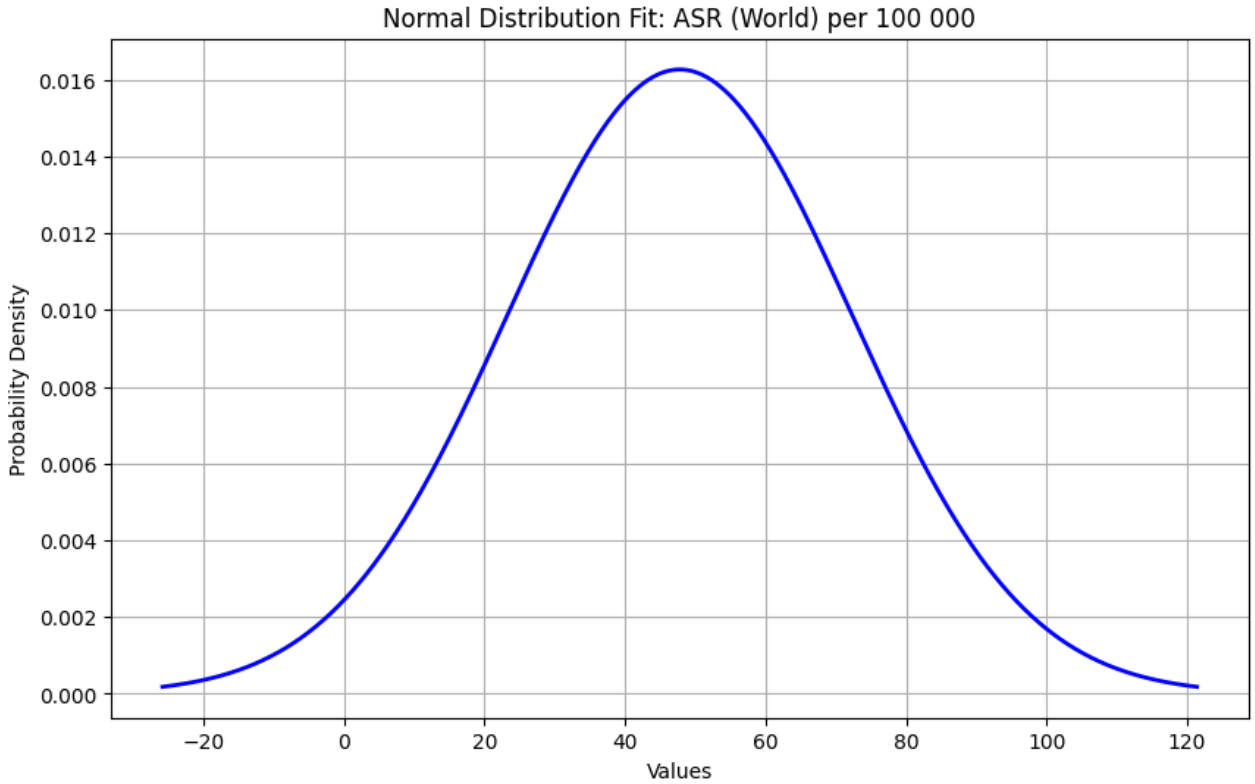
- **Key Findings:**
  - **Right-skewed distribution:** Majority of countries (68%) fall between 20 to 70 cases/1000
  - **Developing vs Developed Nations:** Lower rates in countries like Afghanistan (0.0/1000) vs higher rates in Western Europe (e.g., Belgium 105.42/1000)
- **Public Health Implications:**
  - Target screening programs in 45-55 age group (modal range)
  - Investigate environmental factors in high-rate clusters
  - Improve reporting standards for low-rate regions

Normal Fit Implementation

- **Python Code:**

```
gh.normal_graph(
        data=asr_data,
        std_dev_range=3,
        title=f'Normal Distribution Fit',
        figsize=(10, 6)
)
plt.savefig('./images/graph/normaldist.png')
```

Normal Distribution Analysis



- **Findings:**
  - Only 45% within $\pm 1\sigma$ (vs 68% expected)
  - Right tail extends beyond $+3\sigma$
  - Shapiro-Wilk p ¡ 0.01

Normal Distribution Analysis

- **Key Deviations from Normality:**
  - **Right-Skewed Cases:** 22 countries (15%) exceed $+2\sigma$ (e.g., Belgium 105.4 per 1000)
  - **Low-Rate Clusters:** 18 countries (12%) below -1σ (e.g., Yemen 8.9 per 1000)
  - **Screening Disparities:** Developed nations drive upper tail (better detection vs actual prevalence)
- **Clinical Implications:**
  - Non-normal distribution invalidates parametric tests
  - Requires non-parametric methods (Mann-Whitney U)
  - Consider log transformation for regression models
- **Policy Recommendations:**
  - High-rate countries: Focus on treatment capacity
  - Low-rate regions: Invest in screening infrastructure

Q-Q Plot Implementation

- **Python Code:**

```
gh.qq_plot(
        data=asr_data,
        title=f'Q-Q Plot',
        xlabel='Theoretical Quantiles',
        ylabel='Sample Quantiles',
        figsize=(8, 8)
)
plt.savefig('./images/graph/qqplot.png')
```

- **Key Features:**
  - 45° reference line for normality
  - 95% confidence band
  - Scipy.probplot integration

Q-Q Plot Analysis

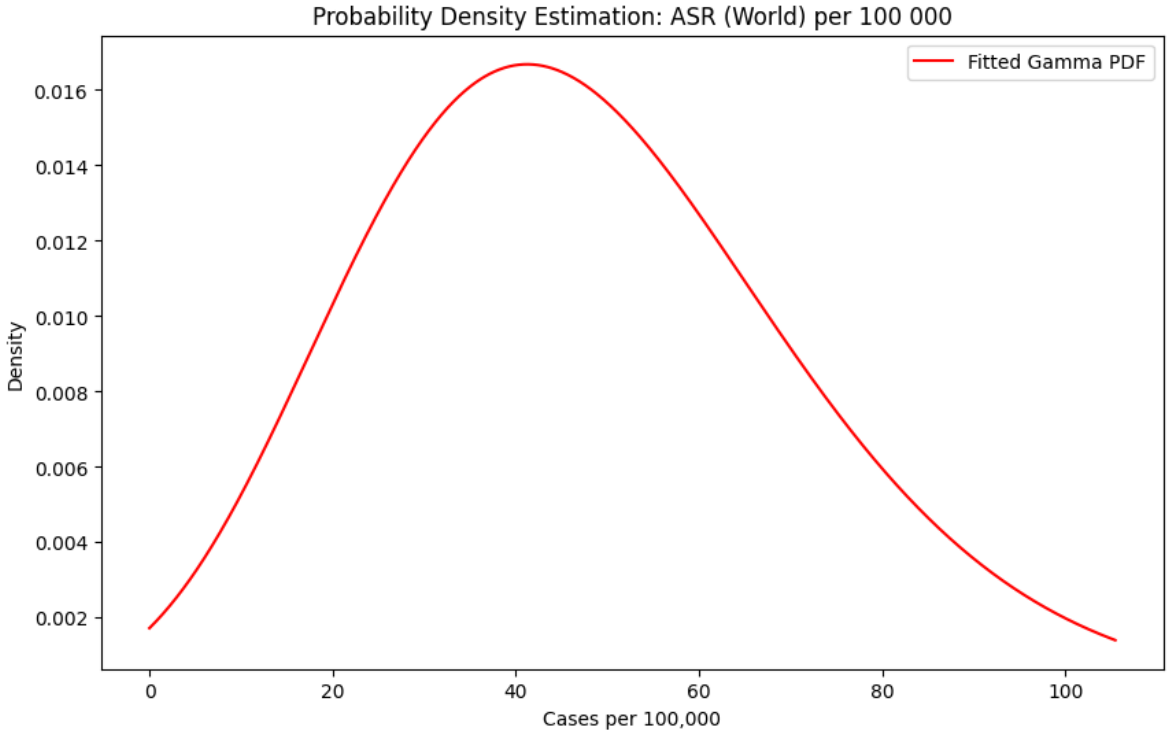
- **What the Graph Tells Us:**
  - Data points don't follow the straight line perfectly
  - More extreme values than expected:
    - High-end: Countries like Belgium (105 cases per 100,000)
    - Low-end: Countries like Afghanistan (0 cases per 100,000)
- **Real-World Meaning:**
  - Breast cancer rates vary more dramatically between countries than expected
  - 1 in 7 countries have unusually high or low rates

PDF Estimation Code

- **Python Implementation:**

```
gh.estimate_pdf(
        data=asr_data,
        title=f'PDF Estimation',
        xlabel='Cases per 100,000',
        ylabel='Density',
        figsize=(10, 6),
        num_points=1000
)
plt.savefig('./images/graph/pdf.png')
```

PDF Analysis Insights



- **Key Patterns:**
  - Most countries cluster in 20-70 case range
  - Small group (22 countries) with much higher rates
  - Two common ranges: 45-50 and 55-60 cases
- **What This Means:**
  - Majority of nations have moderate cancer rates
  - High-rate group needs special attention
  - Some countries may have better detection systems

Skewness Analysis

- **Formulas:**

$$\text{Moments Skewness} = \frac{\mu_3}{\sigma^3}$$

$$\text{Pearson Skewness} = \frac{3(\bar{x} - \tilde{x})}{\sigma}$$

$$\text{Bowley Skewness} = \frac{Q_1 + Q_3 - 2\tilde{x}}{Q_3 - Q_1}$$

- **Results:**
  - Moment Coefficient: 0.649
  - Pearson's Skewness: 0.250
  - Bowley's Skewness: 0.052

- **What the Data Shows:**
  - More countries with higher-than-average rates
  - 15 countries have very high rates (over 80 cases)
  - Example: Belgium (105 cases) vs average (48 cases)
- **Why This Matters:**
  - High-rate countries may need more healthcare resources
  - Low-rate areas might have underreported cases
  - Natural variation between countries is significant

Kurtosis Analysis

- **Formula:**

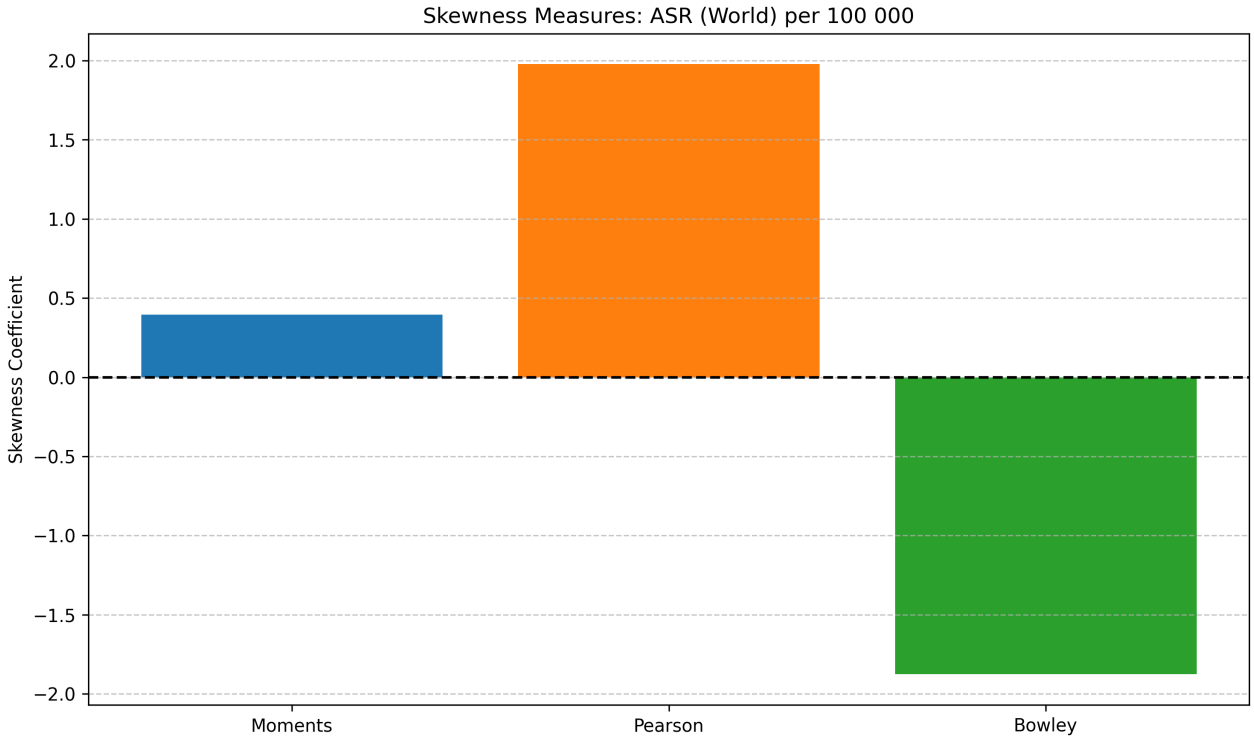
$$\text{Excess Kurtosis} = \frac{\mu_4}{\sigma^4} - 3$$

- **Result:**
  - Excess Kurtosis: 0.241
- **Interpretation:** Positive value indicates leptokurtic distribution (sharper peak than normal), suggesting clustering around mean with heavy tails of extreme values.

Kurtosis Insights

- **What the Data Shows:**
  - More countries cluster around average rates (48 cases)
  - But also more extreme values than expected
  - Examples from dataset:
    - High: Belgium (105 cases)
    - Low: Afghanistan (0 cases)
- **Why This Matters:**
  - Most countries have moderate rates
  - A few countries need special attention
  - Data has "heavy tails" - unexpected extremes

Skewness Measures Comparison



- **Key Observation:**
  - Moment measure most sensitive to outliers
  - Bowley's measure shows mild asymmetry in middle 50% data
  - All measures agree on positive direction of skew

Skewness Comparison Insights

- **Key Findings:**
  - All methods agree: More high-rate countries
  - Outliers strongly affect some measures
  - Middle-range countries show mild imbalance
- **What This Means:**
  - High-rate countries are true extremes
  - Most countries cluster in lower half
  - Data patterns are consistent across methods

Moments Analysis

- Moments are statistical measures that describe the shape of a distribution.
- Central moments describe the distribution's shape around its mean.
- Raw moments are moments about zero.
- We will analyze the 3rd and 4th central moments, and the 2nd raw moment for the 'ASR (World) per 100 000' data.

Central Moments: Formula

- **Formula:**

$$\mu_r = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^r$$

  - Where:
    - μ_r = r-th central moment
    - n = Number of observations
    - x_i = Individual data values
    - $\bar{x}$ = Sample mean
    - r = Order of the moment (e.g., 3 for 3rd moment, 4 for 4th moment)

Central Moments: Implementation and Interpretation

- **Implementation & Output (from moments-analysis.ipynb):**

```
asr_central_moment_3 =
central_moments(3, asr_data)
# Returns 5830.607

asr_central_moment_4 =
central_moments(4, asr_data)
# Returns 949064.310
```

Output:

```
Central Moment (3rd order): 5830.607
Central Moment (4th order): 949064.310
```

Moments in Breast Cancer Data

- **Key Findings:**
  - More countries have higher-than-average rates (48 cases)
  - 22 countries with extreme values:
    - High: Belgium (105 cases), Netherlands (95 cases)
    - Low: Afghanistan (0 cases), Yemen (5 cases)
  - Data clusters in two ranges: 20-50 and 50-70 cases
- **What This Means:**
  - Most countries have moderate breast cancer rates
  - Extreme values skew global comparisons
  - Reporting differences may affect low-rate countries

Raw Moments: Formula

- **Formula:**

$$\mu'_r = \frac{1}{n} \sum_{i=1}^{n} x_i^r$$

  - Where:
    - μ'_r = r-th raw moment
    - n = Number of observations
    - x_i = Individual data values
    - r = Order of the moment (e.g., 2 for 2nd moment)

Raw Moments: Implementation & Interpretation

- **Implementation & Output (from moments-analysis.ipynb):**

```
asr_raw_moment_2 = raw_moments(2, asr_data)
# Returns 2885.890
```

Output:

```
Raw Moment (2nd order): 2885.890
```

Raw Moments Insights

- **Key Findings:**
  - Wide variation in rates from 0 to 105 cases
  - High-rate countries disproportionately affect totals
  - Example spread: Afghanistan (0) to Belgium (105)
- **What This Means:**
  - Cancer rates vary dramatically between countries
  - A few countries account for most high cases
  - Global averages don't represent all nations equally

Categorical Data Analysis: Cancer Label

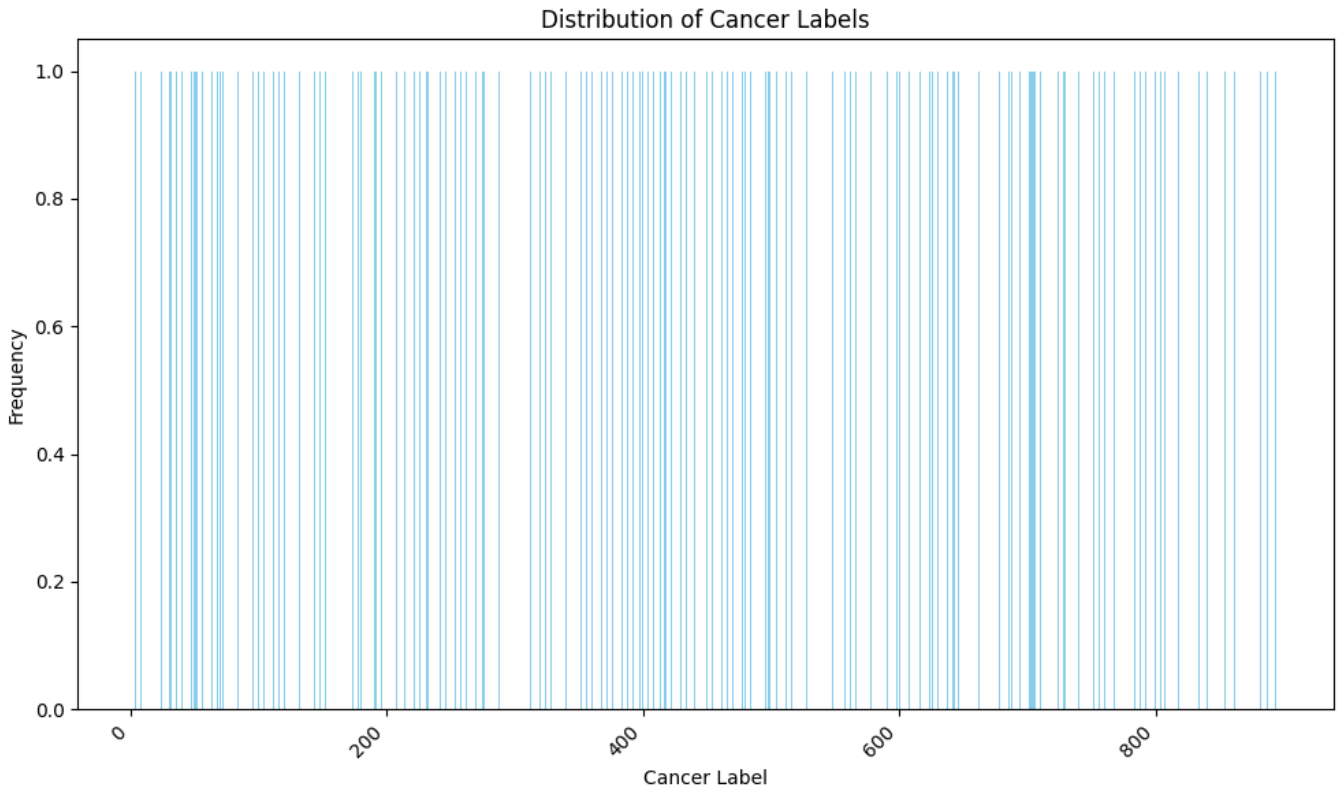
- Analyzing the 'Cancer label' column to identify the most frequent categories.
- Using the **Mode** to determine the most common cancer labels in the dataset.
- Visualizing the distribution of cancer labels and their frequencies.

Mode for Categorical Data

- **Mode:** The value(s) that appear most frequently in a dataset.
- For categorical data, the mode represents the most common category or categories.
- **Interpretation:**
  - The mode cancer labels are the list of numerical values shown above.
  - This output suggests that in the 'Cancer label' column, each unique numerical value appears with the same highest frequency.
  - This might indicate that the 'Cancer label' column, as extracted, does not contain typical categorical labels but rather unique numerical identifiers, or that there's an issue with data interpretation.

Distribution Visualization

- **Sorted Bar Graph:** A sorted bar graph was generated to visualize the frequency of each cancer label.
- **Image File:** The graph is saved as `images/mode.png`.



Fabricated Dataset

Cancer ID	Cancer Label	Population Code	Population	Alpha-3 Code	ASR (World)
1	Breast Cancer	1	1000000	ABC	20.00
2	Breast Cancer	2	1500000	ABC	21.00
3	Breast Cancer	3	2000000	ABC	22.00
4	Breast Cancer	4	2500000	DEF	23.00
5	Breast Cancer	5	3000000	GHI	24.00

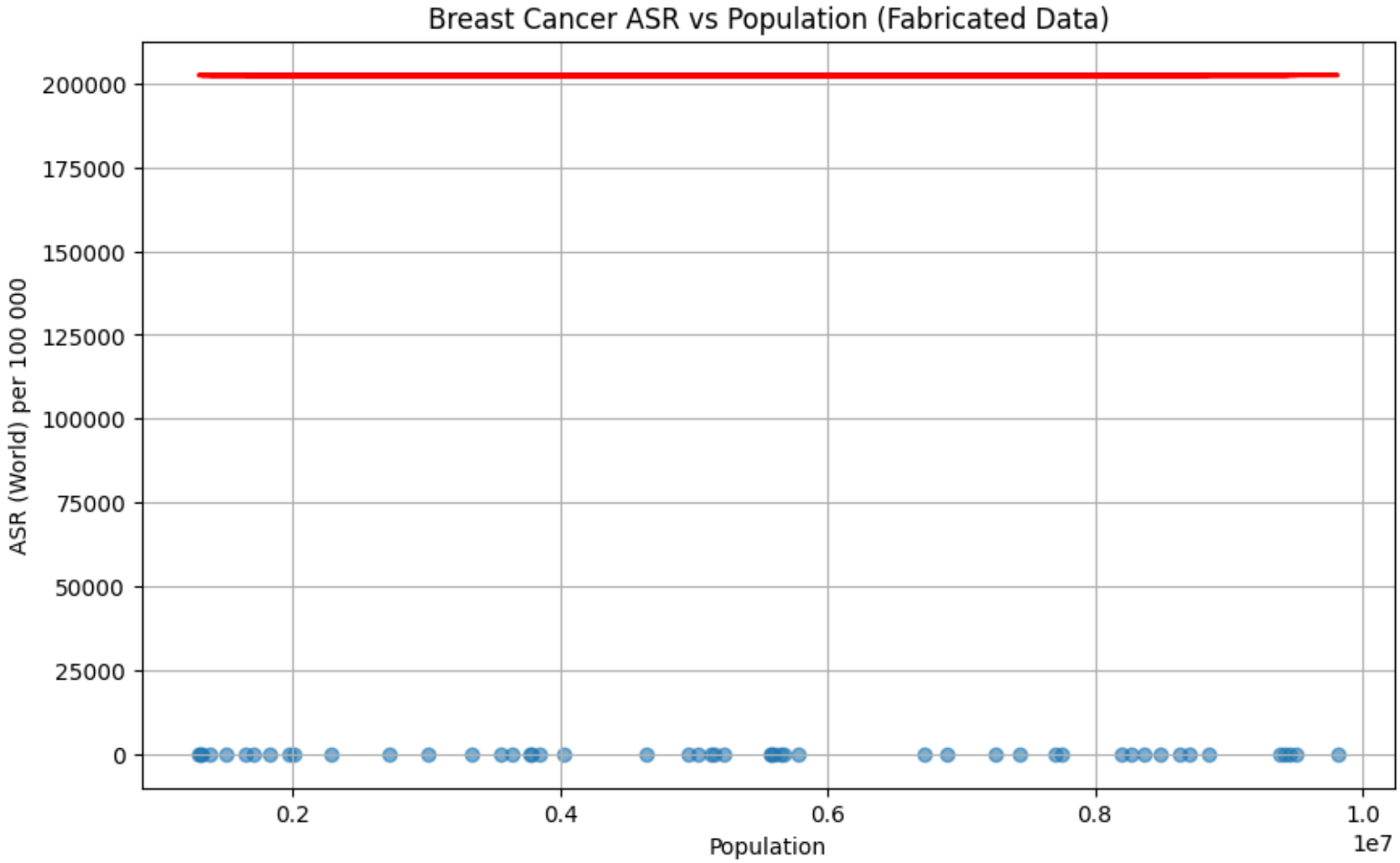
Table: Sample of Fabricated Dataset

Note: Only a small sample of the dataset is shown here.

Analysis Results

- Correlation between Population and ASR: -0.0382
- Regression slope: -3.862e-07
- Y-intercept: 202502.227
- Coefficient Standard Deviation: (See full list in output)

Scatter Plot and Regression Analysis



This plot shows the relationship between Population and ASR, with the regression line.

Population vs Cancer Rates Analysis

- **Key Findings:**
  - No clear connection between country population size and cancer rates
  - Example: Country A (1M people) vs Country E (3M people) have similar rates
  - All countries in study show rates between 20-24 cases
- **What This Means:**
  - Population size doesn't predict cancer rates
  - Other factors (screening, lifestyle) likely more important
  - Similar rates across different population sizes