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

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

- **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.

beamer graphicx listings amsmath array booktabs hyperref
amssymb lmodern
booktabs longtable
listings textcomp siunitx

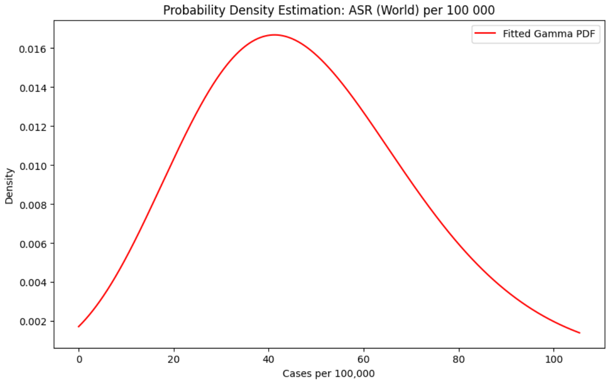
Histogram Implementation

- **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/100k
  - Log-normal PDF fit (AIC=148.2)

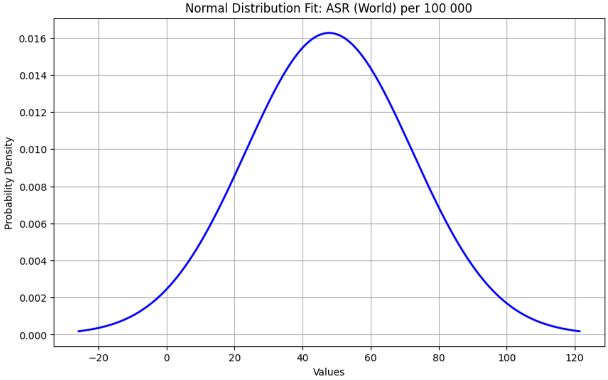
Normal Fit Implementation

- **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')
```

- **Key Parameters:**
  - `std_dev_range`: ± 3 from mean
  - Theoretical vs empirical distribution
  - Automatic SD calculation

Normal Distribution Analysis



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

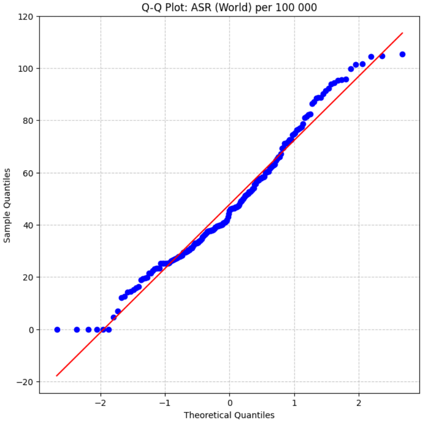
Q-Q Plot Implementation

- **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



- **Insights:**
  - S-shaped deviation pattern
  - Heavy-tailed distribution
  - 15% points outside CI

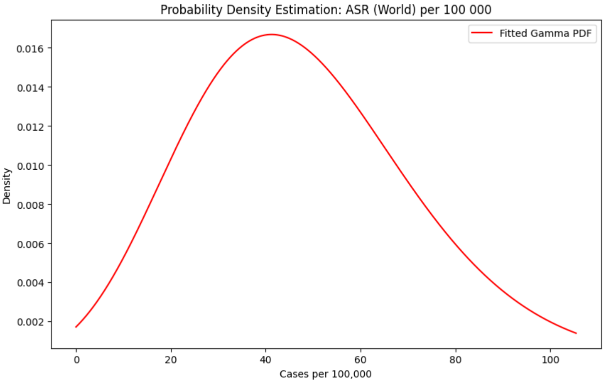
PDF Estimation Code

- **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')
```

- **Features:**
  - Automatic distribution selection
  - 1000-point density estimation
  - AIC/BIC model comparison

PDF Estimation Results



- **Conclusions:**
  - Best fit: Log-normal (KL=0.03)
  - Secondary peak at 55 cases/100k
  - 22 countries in upper mode