



Corridor of Comprehensive Knowledge

STATISTICAL ANALYSIS OF STROKE DATA BASED ON THE WEIBULL DISTRIBUTION



cck@theccck.com

www.theccck.com

Statistical Analysis of Stroke Data Based on the Weibull Distribution

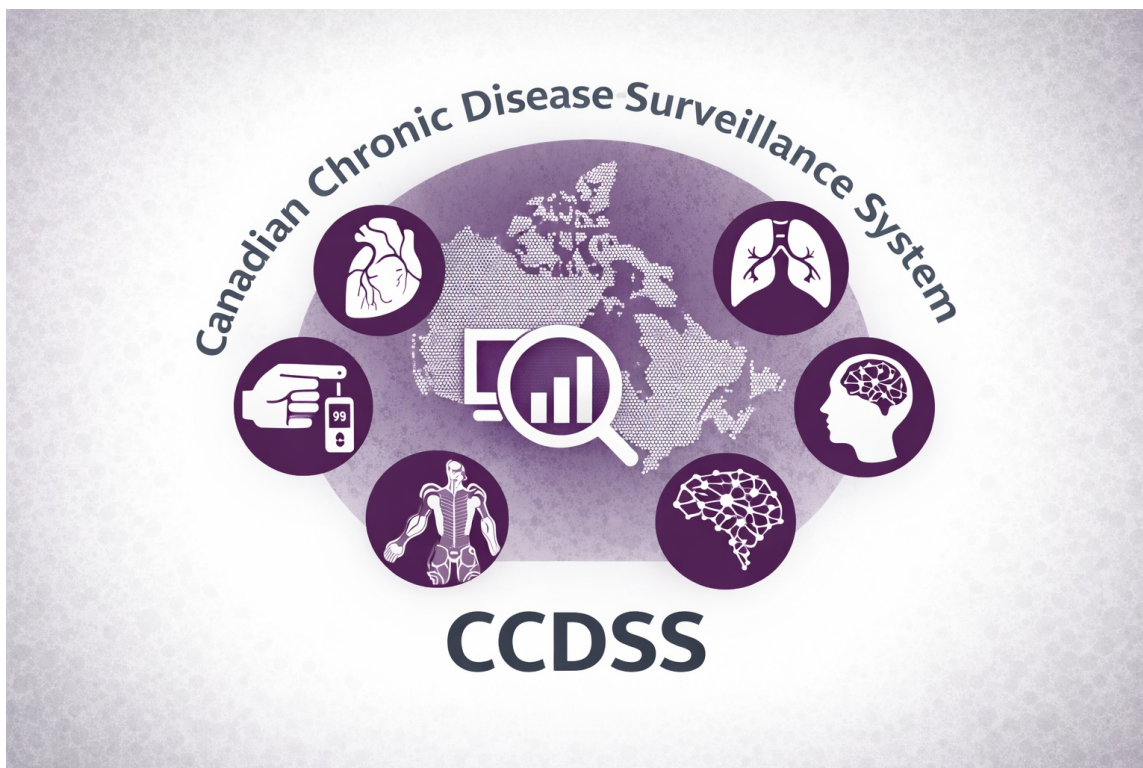
1) Background

The Canadian Chronic Disease Surveillance System (CCDSS) is a national surveillance system that reports indicators such as the incidence and prevalence rates of several important diseases and health conditions in Canada using aggregated data.

Part of the outputs of this system relates to cardiovascular diseases, including stroke. These results are typically presented as annual rates (for example, within five-year intervals) and are stratified by age groups.

The data generated by this system contain highly valuable information for examining changes and developments in population health in Canada across different years.

In this study, and in the second section in particular, we examine a portion of the data obtained from CCDSS in order to demonstrate how such data can be analyzed from a statistical and probabilistic perspective and how a probability density function consistent with the observed data can be constructed.



2) Problem Statement

In the second column of Table 1, the number of strokes occurring in Canada per 100,000 individuals during the 2023–2024 period is presented for age groups defined in five-year intervals (Column 1) [1].

In the third column, the population size of each age group is provided. These values are approximate estimates obtained from the official Statistics Canada website [2].

Table 1) Stroke-related data in Canada during the 2023–2024 period

Group	Age Group	Rate per 100000	Population
1	<20	0	8,000,000
2	20–24	19	2,600,000
3	25–29	22	2,900,000
4	30–34	31	3,000,000
5	35–39	48	2,900,000
6	40–44	76	2,700,000
7	45–49	116	2,500,000
8	50–54	179	2,400,000
9	55–59	246	2,500,000
10	60–64	353	2,700,000
11	65–69	483	2,400,000
12	70–74	690	1,900,000
13	75–79	972	1,500,000
14	80–84	1,391	950,000
15	85–89	1,852	550,000
16	90+	2,034	350,000

Using this table, answer the following questions:

- Assuming that stroke cases are uniformly distributed within each age group, determine the probability of stroke occurrence in each age group. Note that the sample space consists of individuals who have experienced a stroke. In other words, the objective is to determine the probability that a stroke occurred in a given age group among the total population of stroke cases.
- Draw the bar chart representing the probability of stroke occurrence by age.
- If the random variable X denotes the age of an individual who has experienced a stroke, determine the mean and variance of X .
- Using the bar chart or Table 2, determine the percentage of individuals aged 72 years or younger who experienced a stroke during the 2023–2024 period. Repeat the same calculation for ages 22, 32, 42, 52, 62, 82, and 92.
- A Weibull random variable is a random variable whose probability density function is given by:

$$f(x) = \frac{k}{\lambda} \left(\frac{x}{\lambda}\right)^{k-1} \exp\left(-\left(\frac{x}{\lambda}\right)^k\right), \quad x > 0$$

where λ and k are positive constants and represent the parameters of the Weibull probability density function.

If X is a Weibull random variable, it can be shown that its mean and variance are given, respectively, by

$$E(X) = \lambda \Gamma\left(1 + \frac{1}{k}\right)$$

and

$$\text{Var}(X) = \lambda^2 \left(\Gamma\left(1 + \frac{2}{k}\right) - \Gamma^2\left(1 + \frac{1}{k}\right) \right).$$

In these relationships,

$$\Gamma(\alpha) = \int_0^{+\infty} x^{\alpha-1} e^{-x} dx$$

Using these results, and with the aim of selecting the most appropriate density function, determine the values of λ and k such that the resulting Weibull density function provides the best behavioral fit to the bar chart drawn in response to part b). Use the results obtained in part (c).

- f) Now, using the discussion above and the probability density function obtained in part (e), determine the probability that an individual who experienced a stroke during the 2023–2024 period was 72 years of age or younger. Repeat the same calculation for individuals aged 22, 32, 52, 62, 82, and 92 years or younger. Compare the resulting values with those obtained in part (d). What conclusions can be drawn?
- g) A probability density function helps us analyze the occurrence of stroke during the 2023–2024 period from both probabilistic and statistical perspectives. In effect, by having such a function, we summarize the statistical behavior of the phenomenon using only two parameters (λ and k). With this in mind, and considering the procedure followed in the previous parts, explain which factors may affect the accuracy of determining a probability density function that provides an adequate fit to the observed data. Also explain why having a density function that is well aligned with the observations can be useful in data analysis. Finally, discuss whether the Weibull distribution was an appropriate choice for this problem.



Ernst Hjalmar Waloddi Weibull
(18 June 1887 – 12 October 1979)

Source: Porträttarkiv (Swedish Portrait Archive), via Wikimedia Commons, licensed under CC BY 4.0.

3) Solution

- a) Based on the number of stroke cases per 100,000 individuals in each age group, the number of stroke cases in the entire population of each age group can be determined by proportion. The number of stroke cases in age group 1 is zero ($\text{case}_1 = 0$). The number of stroke cases in age group 2 is obtained by multiplying $\frac{19}{100000}$ by 2,600,000 that is,

$$\text{case}_2 = \frac{19 \times 2,600,000}{100000} = 494.$$

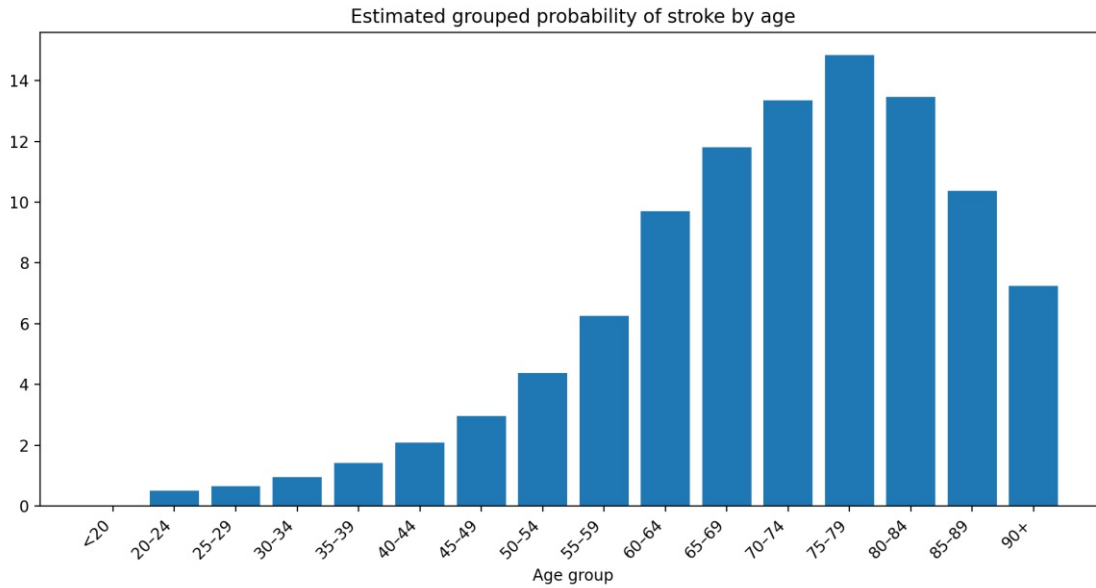
The values for the remaining groups are calculated in the same way (Table 2, Column 3). The probability of stroke for group i is then determined by (table 2, Column 4).

$$P_i = \frac{\text{case}_i}{\sum_{k=1}^{16} \text{case}_k} = \frac{\text{case}_i}{494 + 638 + \dots + 7,119} = \frac{\text{case}_i}{98,184.5}$$

Table 2. Calculation of the probability of stroke occurrence based on age-group data

Group	Age Group	Expected Strokes	Probability
1	<20	0.0	0.000000
2	20–24	494.0	0.005031
3	25–29	638.0	0.006498
4	30–34	930.0	0.009472
5	35–39	1,392.0	0.014177
6	40–44	2,052.0	0.020899
7	45–49	2,900.0	0.029536
8	50–54	4,296.0	0.043754
9	55–59	6,150.0	0.062637
10	60–64	9,531.0	0.097072
11	65–69	11,592.0	0.118063
12	70–74	13,110.0	0.133524
13	75–79	14,580.0	0.148496
14	80–84	13,214.5	0.134588
15	85–89	10,186.0	0.103743
16	90+	7,119.0	0.072506

- b) The bar chart is constructed using the data presented in Column 4 of Table 2. The midpoint of each age interval is taken as the representative value for that interval.



- c) For individuals younger than 20 years, the value 10 is used, and for individuals older than 90 years, the value 92 is used as the representative age. Accordingly, the mean of X and its variance are given, respectively, by:

$$E(X) = (10 \times 0) + (22 \times 0.005031) + \dots + (92 \times 0.072506) \approx 70.934$$

$$VAR(X) = ((10 - 70.934)^2 \times 0) + ((22 - 70.934)^2 \times 0.005031) + \dots + ((92 - 70.934)^2 \times 0.072506) \approx 208.161$$

- d) The required probabilities are obtained by summing the heights of the bars that are less than or equal to the bar represented by the age of interest. According to Table 2, for age 22, the probability is 0.005031 which is equal to 0.5031%. For age 32, the result is

$$0.005031 + 0.006498 + 0.009472 = 0.021001$$

which corresponds to 2.1001%.

The probabilities for the remaining ages are calculated in the same manner. Accordingly, the resulting values for ages 22 through 92, respectively, are as follows:

0.5031%, 2.1001%, 5.6077%, 12.9367%, 28.9076%, 54.0666%, 82.3747%, 100%

- e) Using the relationships provided in this part, we obtain:

$$\begin{aligned} \text{Var}(X) &= \lambda^2 \Gamma\left(1 + \frac{2}{k}\right) - \lambda^2 \Gamma^2\left(1 + \frac{1}{k}\right) \\ &= \lambda^2 \Gamma\left(1 + \frac{2}{k}\right) - E^2(X) \\ &= \left(\frac{E(X)}{\Gamma\left(1 + \frac{1}{k}\right)}\right)^2 \Gamma\left(1 + \frac{2}{k}\right) - E^2(X) \end{aligned}$$

Therefore:

$$\frac{\Gamma\left(1 + \frac{2}{k}\right)}{\Gamma^2\left(1 + \frac{1}{k}\right)} = \frac{\text{Var}(X) + E^2(X)}{E^2(X)} = \frac{208.161 + (70.934)^2}{(70.934)^2} \approx 1.04$$



Hence:

$$\Gamma\left(1 + \frac{2}{k}\right) \approx 1.04\Gamma^2\left(1 + \frac{1}{k}\right)$$

Using numerical methods, the above equation yields $k = 5.69$. Then, by evaluating $\Gamma\left(1 + \frac{1}{k}\right)$ and $\Gamma\left(1 + \frac{2}{k}\right)$ for $k = 5.69$ and substituting these values into the variance formula, we obtain $\lambda = 76.69$.

- f) We determine $P(X \leq a)$ for $a = 22, 32, 42, 52, 62, 72, 82,$ and 92 using

$$P(X \leq a) = \int_0^a \frac{5.69}{76.69} \left(\frac{x}{76.69}\right)^{4.69} \exp\left(-\left(\frac{x}{76.69}\right)^{5.69}\right) dx$$

The results of these calculations (expressed as percentages), for comparison with those obtained in part (d), are presented in Table 3.

Table 3. Comparison of empirical and Weibull cumulative probabilities

Value of a	22	32	42	52	62	72	82	92
From Bar Chart	0.5031	2.1001	5.6077	12.9367	28.9076	54.0666	82.3747	100
Weibull distribution	0.0818	0.6886	3.1963	10.3772	25.7828	50.2637	76.8759	94.0333

- g) The approximate nature of the population values for the age groups, the grouping of the data, the limited amount of available data, the open-ended age intervals at the lower and upper ends, and the fact that the fitting is based only on the mean and variance are among the factors that reduce the accuracy of the Weibull function in matching the observed results.

Nevertheless, having a probability density function allows scattered data to be aggregated and summarized, and provides the possibility of estimation, prediction, and planning for the future. The results obtained from using the Weibull distribution are therefore reasonably acceptable.

The Weibull distribution captures well the overall trend of increasing stroke probability with age, and from a behavioral perspective it performs better than some simpler alternatives.

The Corridor of Comprehensive Knowledge (CCK) is a not-for-profit educational organization that develops problems, activities, and learning materials connecting mathematics, science, technology, art, and other fields within shared contexts.

Its work includes interdisciplinary workshops, thematic learning scenarios, original educational resources, and the interdisciplinary educational journal *The CCK Clerestory*, all designed to make connections visible, deepen conceptual understanding, and foster curiosity and critical thinking.

cck@theccck.com

www.theccck.com