Do occupational domains exert an influence on individuals' lifespans?

The examination and statistical analysis of the lifespans, origins, and birth eras of famous people is a continually explored area of research. In this study, we initially gathered data from a rigorously cross-verified dataset encompassing variables such as lifespans, birth years, genders, and other relevant variables. We then employed right censoring to adjust for those still living at the culmination of our investigation, a standard procedure in survival analysis. Ultimately, we utilized Kaplan-Meier plots, hazard ratios, and the Cox Proportional-Hazards Model to scrutinize the influence of predictor variables on survival time. This enabled us to discern which domains exhibit longer or shorter lifespans and ascertain the statistical significance of these disparities.

For the survival analysis, we initially sourced comprehensive data from [1], encompassing a representative cohort of individuals. This dataset incorporated covariates such as birth year, gender, region, domain of influence, lifespan, era, and the individual's living status. Of primary interest in our subsequent analysis is the "domain of influence," categorized into four groups: Culture, Leadership, Discovery/Science, and Sports/Games. We omitted individuals born before 1500 due to the insufficiency of data, which hindered a substantial interpretation of lifespan trends before this period. Recognizing that some individuals within this dataset are still alive as of 2022, we applied right censoring, resulting in a total of 90,488 observations.

Given that our event of interest is death, it’s intuitive to visualize the probability of death against the ages among four groups of domains. A prevalent method for this purpose is the survival function, S(t), representing the probability of survival beyond time t. At the origin, the survival probability is set at 1. Subsequently, the Kaplan-Meier survival estimate—a step function that adjusts at each distinct survival time—is applied. Given the skewed distribution of survival, the median offers a more interpretable measure than the mean. Our findings reveal that individuals within the Sports/Games domain possess the highest median lifespan at 81 years, whereas those in the Leadership domain record the lowest at 76 years. Figure 1 illustrates our Kaplan-Meier survival plot.

Subsequently, our aim was to ascertain whether the lifespan differences among the four groups are statistically significant. Consequently, our null hypothesis posits that there is no significant disparity in lifespans across these groups. To test this hypothesis, we employed the log-rank test. With a chi-square value of 18,061 on 3 degrees of freedom and a p-value less than 2e-16, we have compelling evidence to reject the null hypothesis. This solidifies the notion that significant variations in lifespans exist based on the domains of occupation.

Nonetheless, a significant limitation of the survival function is its inherent bias when comparing survival probabilities between disparate age groups. Specifically, the cohort of individuals alive at age 20 far exceeds that at age 80, rendering direct comparisons inequitable. To address this, we will employ the Cox Proportional-Hazards Model, emphasizing the hazard ratio for a more nuanced analysis of the survival data.

In survival analysis, the hazard function h(t) is employed to assess the instantaneous risk of death at a given time t. The Cox Proportional-Hazards Model postulates the logarithm of h(t) as the response variable, while intentionally leaving the baseline hazard function $\log h_0(t)$ unspecified. In our application of the Cox Proportional-Hazards Model, we incorporated all available covariates. Notably, the coefficients of birth year, gender, and era returned extremely low p-values. The p-value for the domain coefficients is only slightly less than 0.01, but still allows us to reject the null hypothesis confidently, suggesting a significant effect.

Additionally, we evaluated the hazard ratios between pairs of groups differentiated by their domains of occupation. A hazard ratio exceeding 1 indicates an increased risk of death, whereas a hazard ratio below 1 signifies a reduced risk. For instance, when comparing individuals in the Sports/Games domain to those in the Leadership domain, the hazard ratio is 0.81. This implies that, at any given point, there are 0.81 times as many deaths in the Sports/Games domain compared to the Leadership domain. With a p-value below 0.001, it's evident that individuals in the Sports/Games domain exhibit a statistically significant reduced hazard of death relative to their counterparts in the Leadership domain. This finding aligns with our observations from the Kaplan-Meier plot and the log-rank p-value test.

From the analysis above, we can conclusively address the initial question of this memo: occupation domains do influence individuals' lifespans, and the lifespan disparities across the four domains are statistically significant. Individuals from the Sports/Games domain have the highest median lifespan, at 81 years. This is intuitive, as consistent exercise and rigorous physical activity inherently foster cardiovascular health, along with strengthening muscles and bones. Conversely, those in Leadership roles exhibit the lowest median lifespan, at 76 years. Such a trend can be attributed to the nature of leadership roles, often characterized by high stress, extended working hours, and considerable responsibilities.

Moreover, individuals in science-related domains tend to be well-educated and possess an advanced understanding of health-related information. This allows them to follow health guidelines more rigorously. The intellectual engagement pervasive in their careers also augments their mental health and overall well-being, thereby positioning them second in terms of lifespan among the four groups.

Reference

[1] https://data.sciencespo.fr/dataset.xhtml?persistentId=doi:10.21410/7E4/RDAG3O

Figures

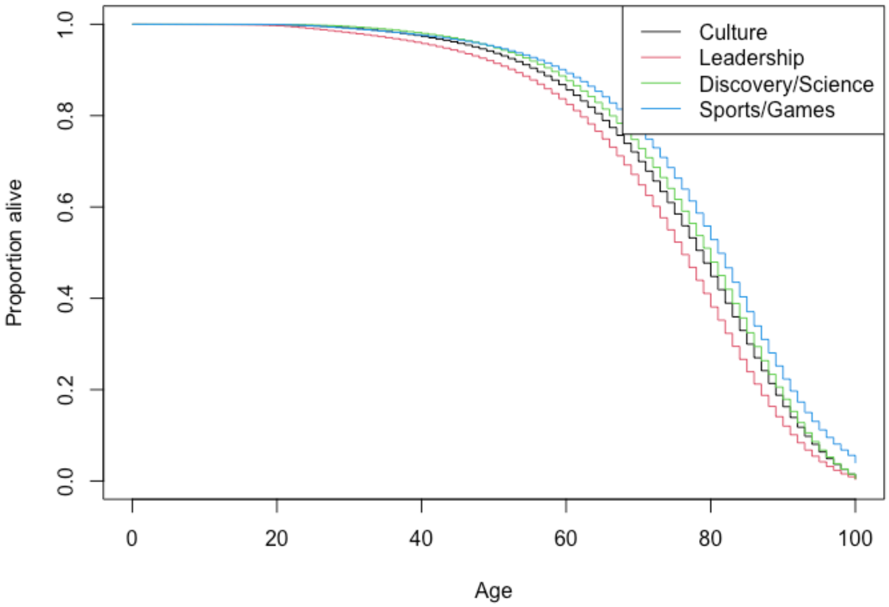

Figure 1: The Kaplan-Meier plot: proportion of alive against ages among four groups of domains of occupations