Youngjun Woo
UMID: yjwoo
January 27th, 2022

Comparative analysis of survival function by region

We aim to analyze whether there is a statistically significant difference in survival functions between regions by using BHHT data. BHHT data consists of 2.2 million notable individuals from Wikipedia editions and Wikidata. After data preprocessing, we used 1,982,476 individuals out of 2.2 million individuals. BHHT data is right-censored data because there are individuals whose death date is unknown or who are still alive. We only used individuals' region of origin, gender, and century of birth in our analysis.

We mostly used the Kaplan-Meier model and log-rank test because we wanted to estimate and compare the survival function from each region. In addition, by comparing the results from the Cox PH model and the Kaplan-Meier model, we wanted to reinforce our analysis result from the Kaplan-Meier model.

First, we estimated and compared the survival function from each region by Kaplan-Meier estimator. Looking at Figure 1, the survival function from Europe has lower a survival probability than those of the other 4 regions across all age ranges. In contrast, Africa and Asia show similar survival functions and higher survival probabilities across all ages than the other three regions. These results are reinforced by statistical testing. Table 1 shows the p-value of the log-rank test between survival functions in each region. As shown in Figure 1, there is no statistically significant difference between the survival functions of Asia and Africa, whereas survival functions in other regions show statistically significant differences. The results of the Cox PH regression model in Table 2 also show the same interpretation. There is a 2% decrease in the expected hazard in Asia as compared to Africa, but this is not a statistically significant difference because the t-test statistic is small. Also, as the survival functions of Oceania and America are similar in the Kaplan-Meier estimator, the increases in expected hazards compared to Africa are similar. There are 12% and 19% increases in expected hazards in America and Oceania compared to Africa respectively. In contrast, the expected hazard in Europe increases by 35% compared to Africa, which is the largest increase compared to other regions. This result is consistent with the fact that the survival function of the Europe region is lower than that of other regions in the Kaplan-Meier estimator.

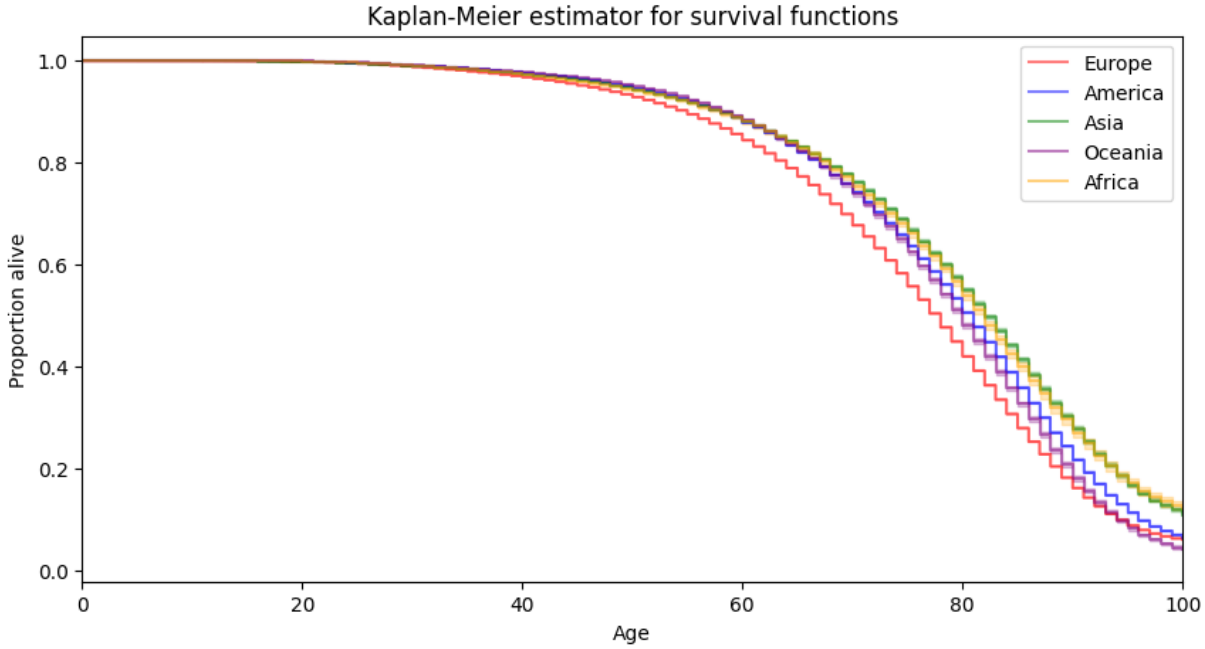

Figure 1

	Europe	America	Asia	Oceania	Africa
Europe	1.0	0.0	0.0000	0.0	0.0000
America	0.0	1.0	0.0000	0.0	0.0000
Asia	0.0	0.0	1.0000	0.0	0.1816
Oceania	0.0	0.0	0.0000	1.0	0.0000
Africa	0.0	0.0	0.1816	0.0	1.0000

Table 1

Model:	PH Reg	Sample size:	1982476
Dependent variable:	clifespan	Num. events:	917232
Ties:	Breslow		

	log HR	log HR SE	HR	t	P>|t|	[0.025	0.975]
reg[T.America]	0.1144	0.0093	1.1212	12.2945	0.0000	1.1009	1.1418
reg[T.Asia]	-0.0159	0.0102	0.9842	-1.5652	0.1175	0.9648	1.0040
reg[T.Europe]	0.3050	0.0092	1.3566	33.2701	0.0000	1.3325	1.3812
reg[T.Oceania]	0.1817	0.0107	1.1993	16.9600	0.0000	1.1744	1.2247

Table 2

To understand why the survival ratio in Europe is generally lower than other regions, we compared the estimated survival functions from each region for each gender. Looking at Figure 2, in the case of the male, the overall survival ratio in the Europe region is lower than other regions as in Figure 1. On the other hand, in the case of females, the difference in survival function by region does not seem to be large compared to that of males. This can also be confirmed through the results of the log-rank test in Table 3. In Table 3, the left side is the t-test statistic for men, and the right side is the t-test statistic for women. It can be seen that the t-test statistic between Europe and other regions on the left is much larger than that on the right.

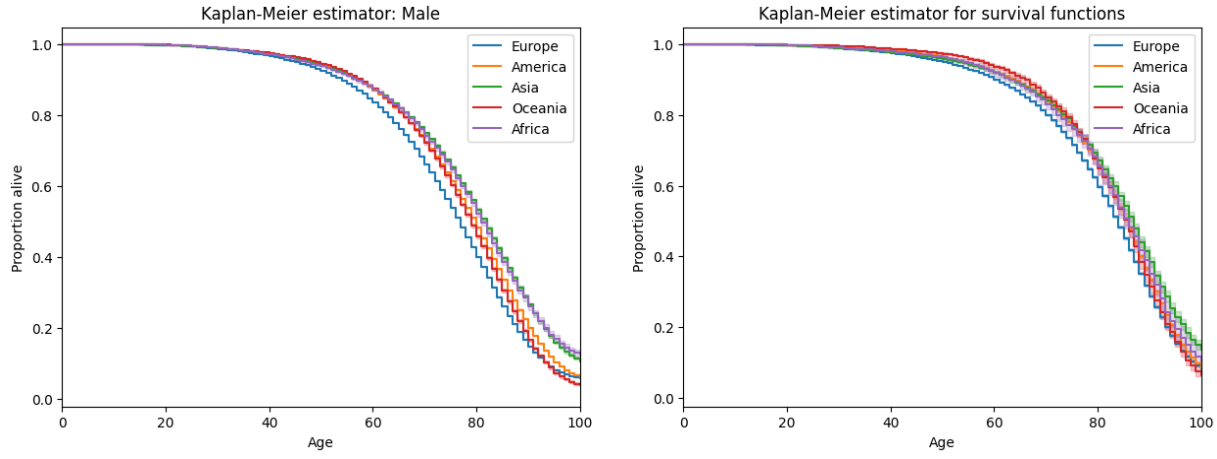

Figure 2

	Europe	America	Asia	Oceania	Africa
Europe	0.00	5357.89	4534.92	412.15	1160.03
America	5357.89	0.00	802.98	130.13	176.28
Asia	4534.92	802.98	0.00	840.70	0.66
Oceania	412.15	130.13	840.70	0.00	310.42
Africa	1160.03	176.28	0.66	310.42	0.00

	Europe	America	Asia	Oceania	Africa
Europe	0.00	384.67	241.60	61.89	40.04
America	384.67	0.00	22.91	0.03	0.03
Asia	241.60	22.91	0.00	10.68	2.59
Oceania	61.89	0.03	10.68	0.00	0.08
Africa	40.04	0.03	2.59	0.08	0.00

Table 3

Thinking that there might be an influence of the century of birth on these results, we compared the survival functions of males by region and birth of century. As demonstrated by its illustration in Figure 3, we exclude the 21st century from the analysis because the sample size is small compared to those of other centuries given that many people born then are still alive. Looking at the trend over time, there are clear differences in survival functions between regions through the 16th, 17th, and 18th centuries. However, from the 19th century onwards, it can be confirmed that the difference is remarkably reduced. It should be noted that, unlike the previous analysis results, the overall survival function of European men did not show a tendency to be lower than other regions' men. In the 16th, 17th, and 18th centuries, the overall survival ratio of Asian males tended to be lower than that of European males.

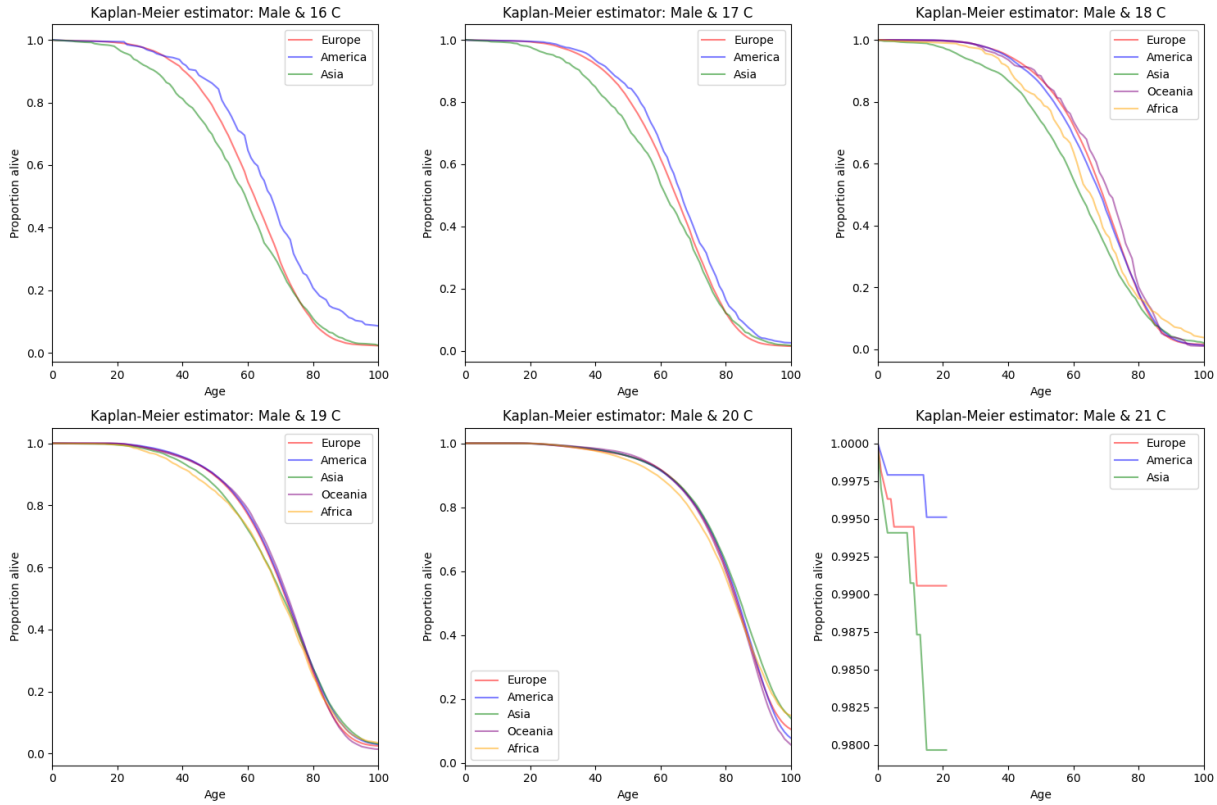

Figure 3