Biostatistics

PROJECT 5

Survival Analysis

Iza Danielewska

Dawid Poławski

Warsaw, 27.01.2025

TABLE OF CONTENTS

Description of the population / sample

POPULATION

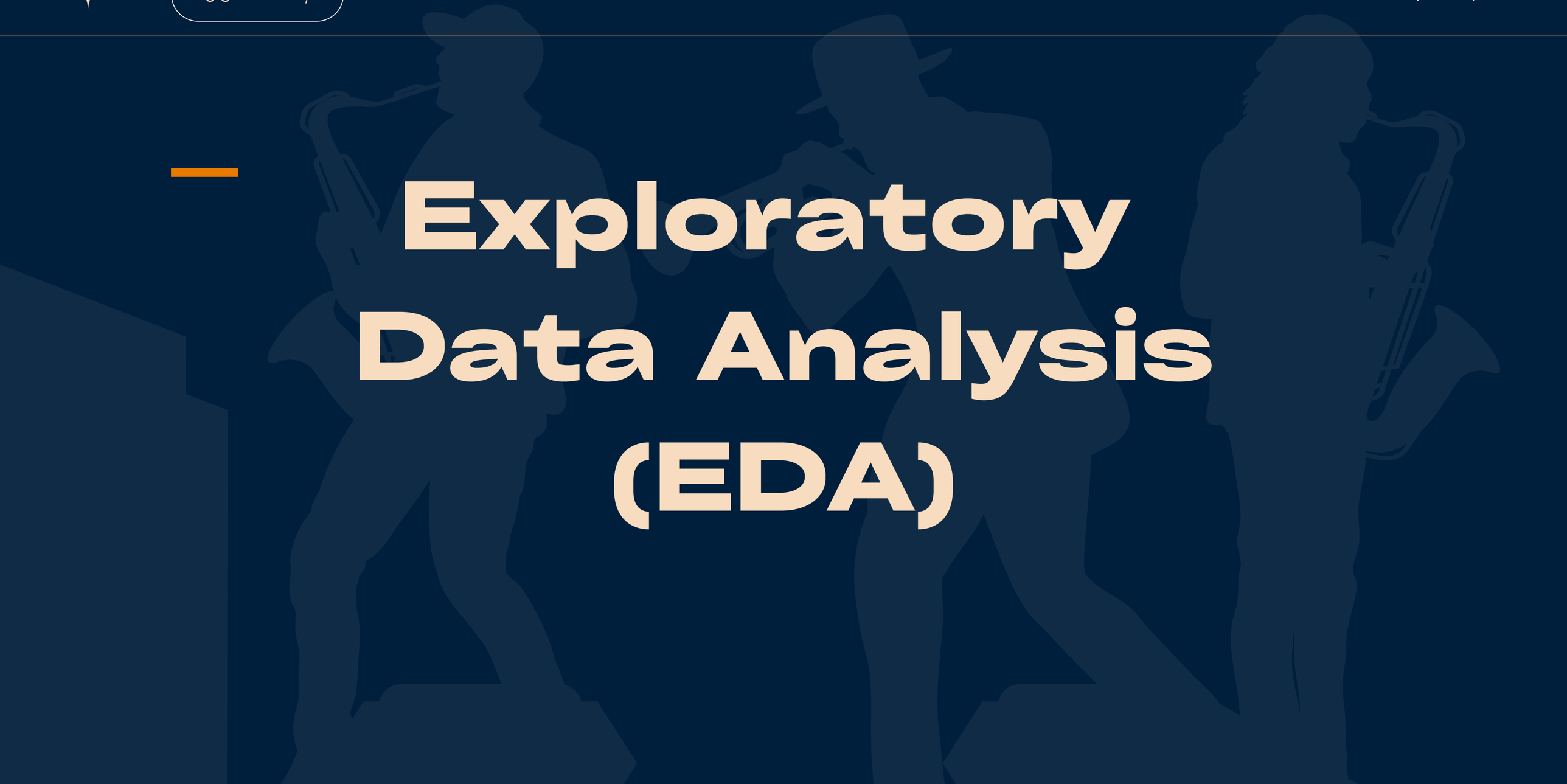
The authors aimed to create a cohort of famous musicians using an unbiased and transparent sampling scheme.

Famous musicians were defined as those who had achieved a number one album on the UK charts, chosen due to their long history and consistent representation as a marker of success.

SAMPLE

The authors collected data spanning from 1956, when the UK charts were first established, to the end of 2007.

The dataset included musicians (solo artists and band members) who achieved a number one album in the UK between 1956 and 2007
(n = 1046 musicians, with 71 deaths, 7%).



Exploratory Data Analysis (EDA)

STATISTICS

- Age at death

Min	1 st. Qu.	Median	Mean	3 st. Qu.	Max	NA's
21.73	37.98	49.94	49.75	59.82	88.98	975

- Age of those alive at the end of the study

Min	1 st. Qu.	Median	Mean	3 st. Qu.	Max	NA's
22.93	38.78	47.98	48.51	58.64	91.23	71

- Age of Musicians - entire dataset (censored)

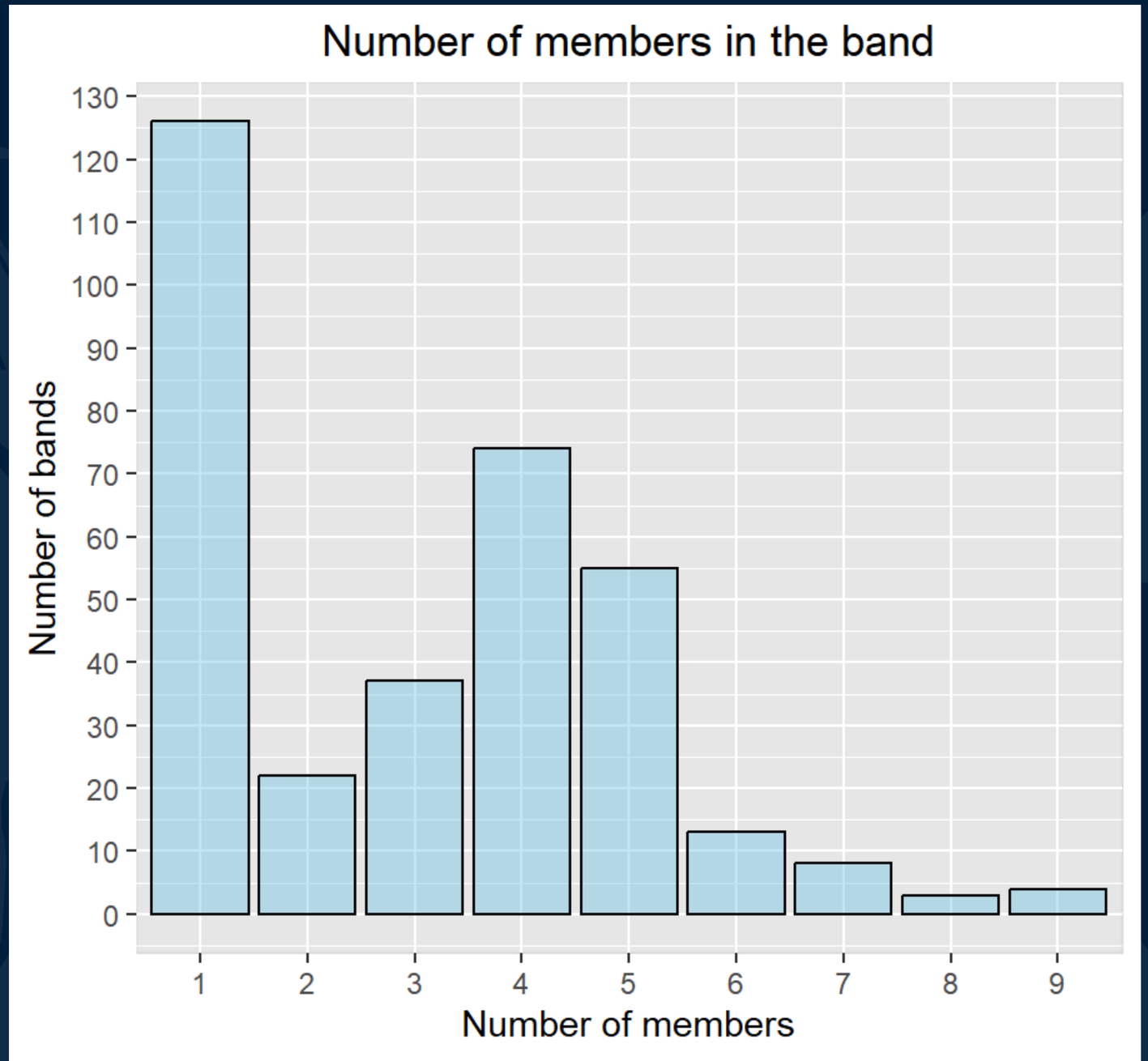
Min	1 st. Qu.	Median	Mean	3 st. Qu.	Max
21.73	38.73	47.99	48.60	58.74	91.23

STATISTICS

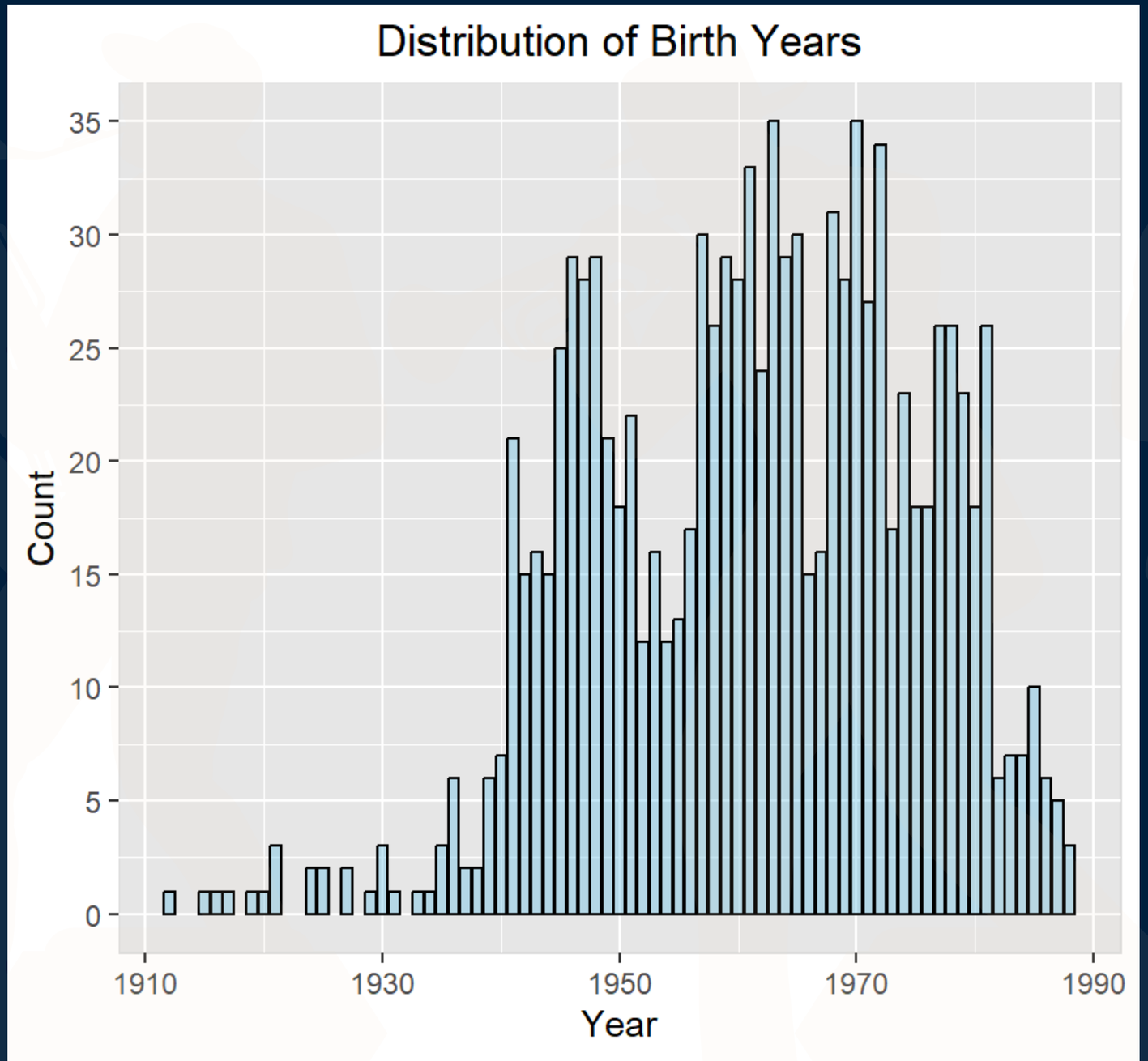
- Time to reach number one on the UK album charts

Min	1 st. Qu.	Median	Mean	3 st. Qu.	Max
11.66	23.49	26.61	27.80	30.75	65.14

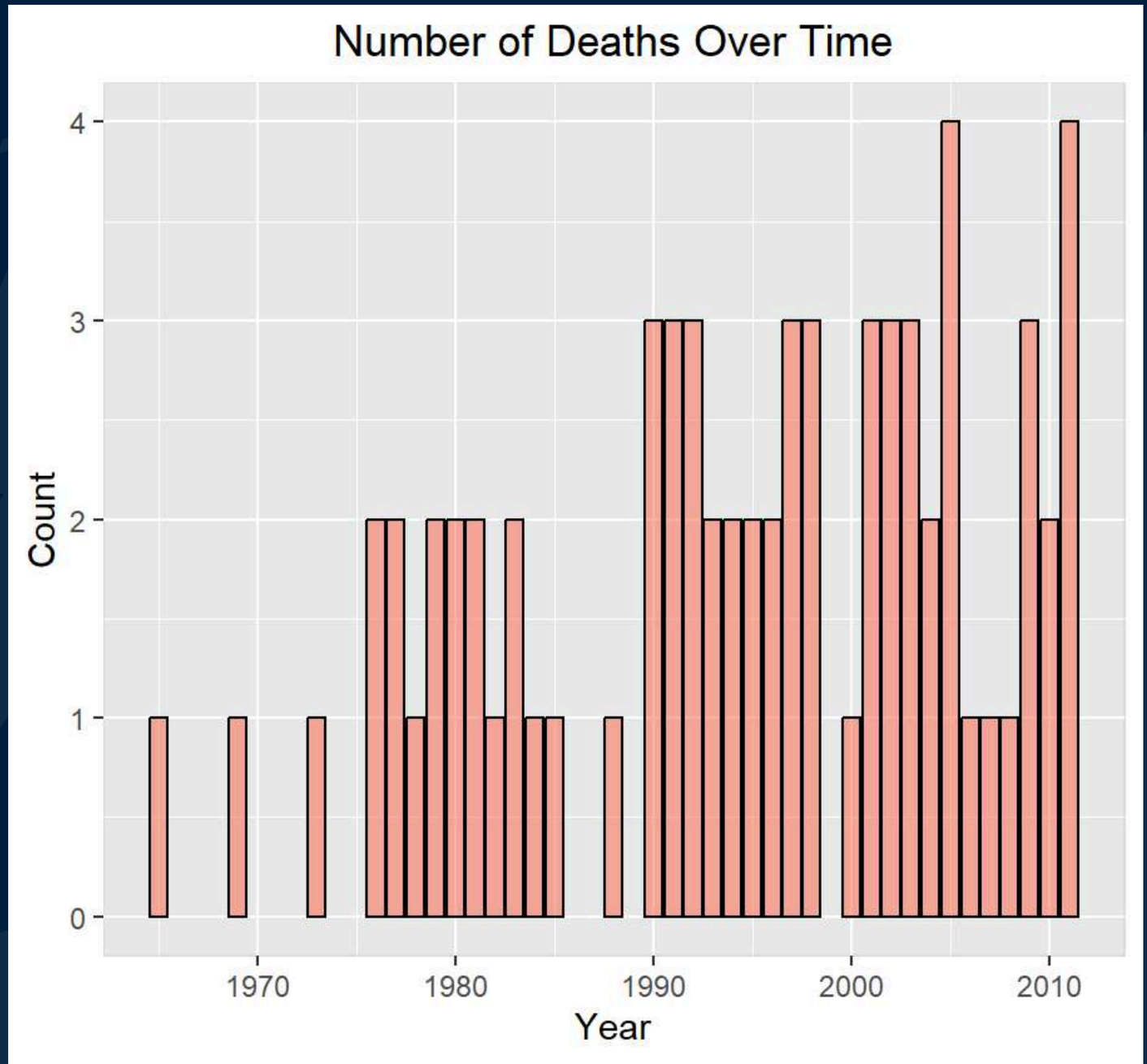
NUMBER OF MEMBERS IN THE BAND



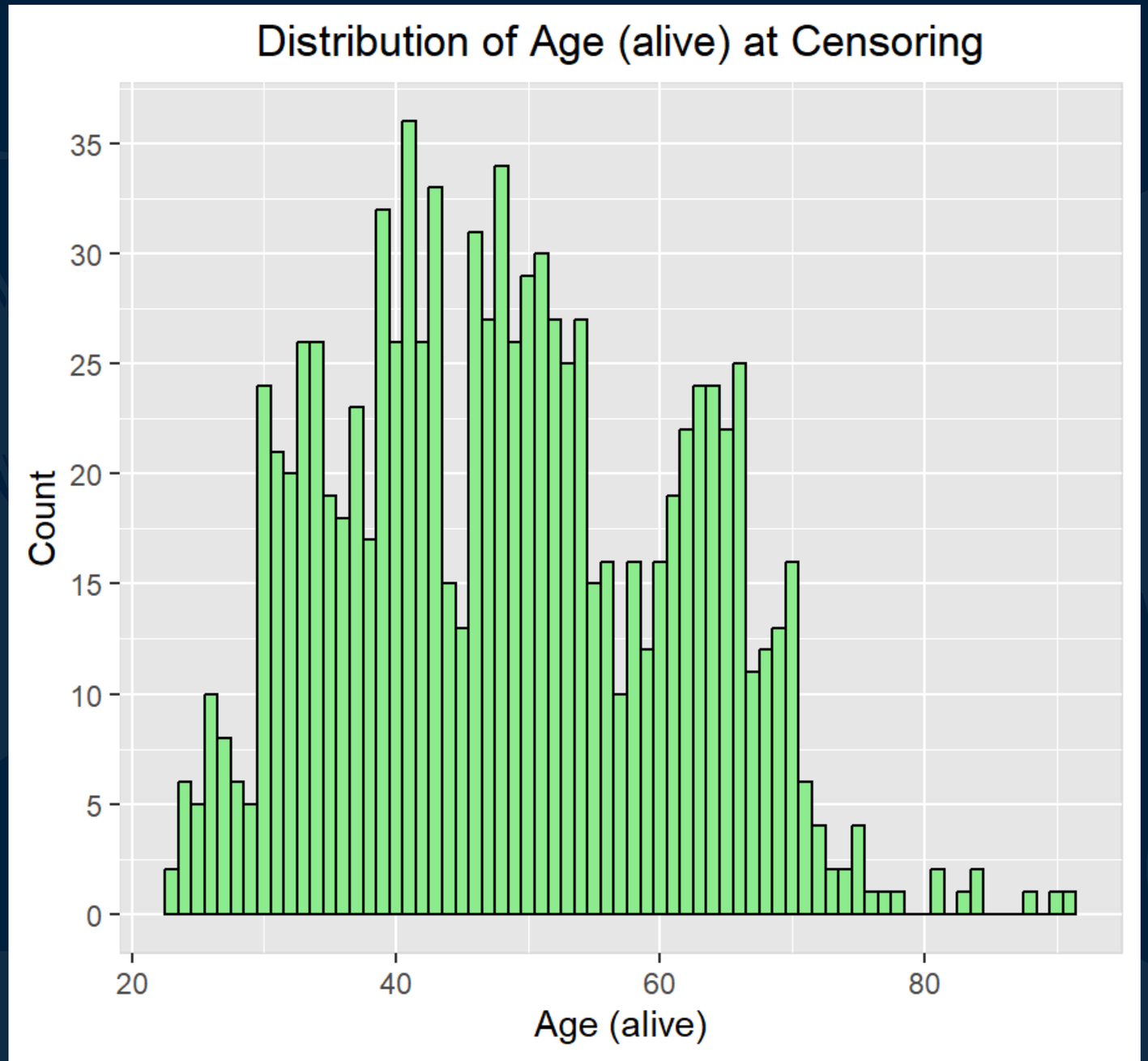
DISTRIBUTION OF BIRTH YEARS



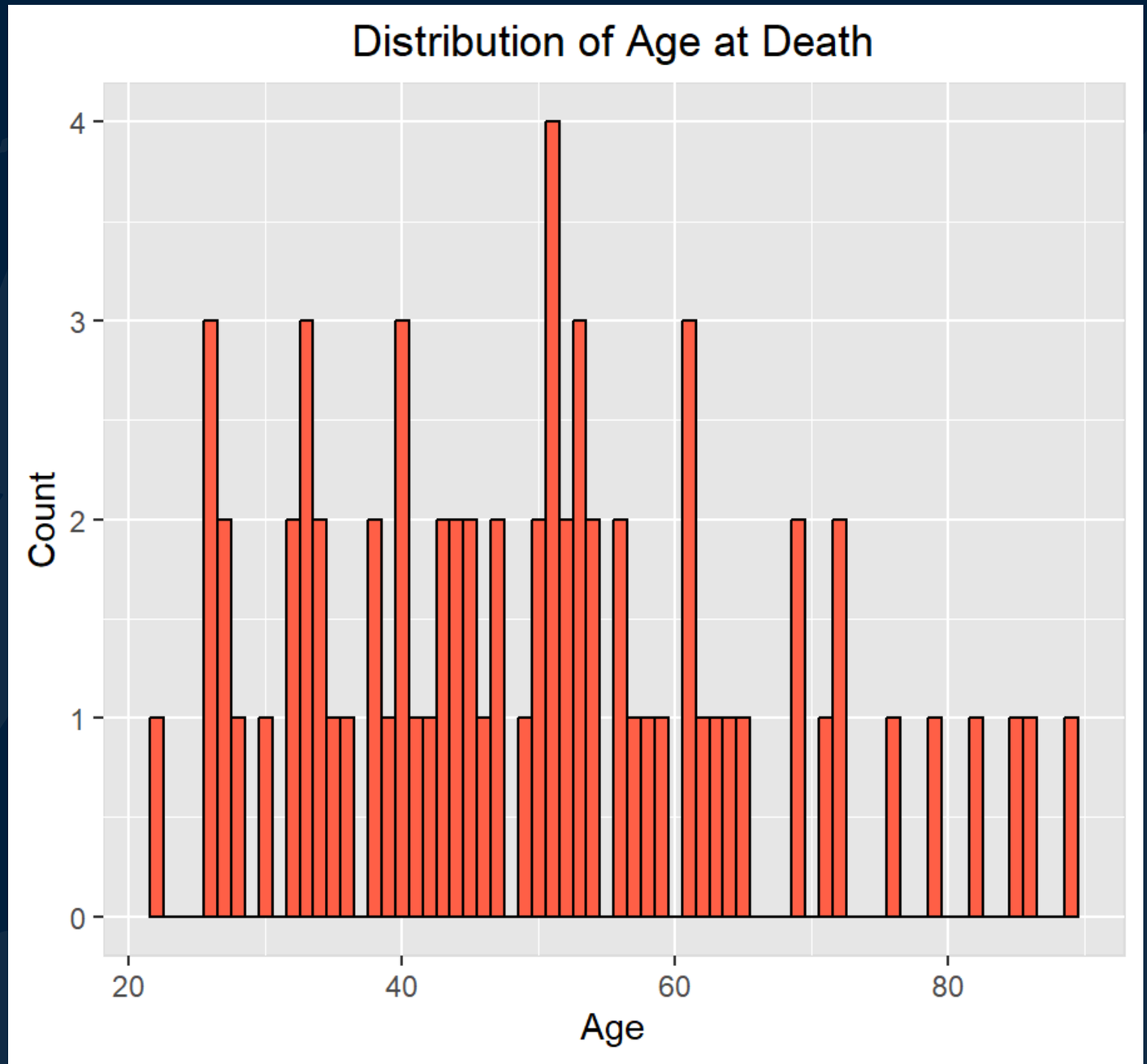
NUMBER OF DEATHS OVER TIME



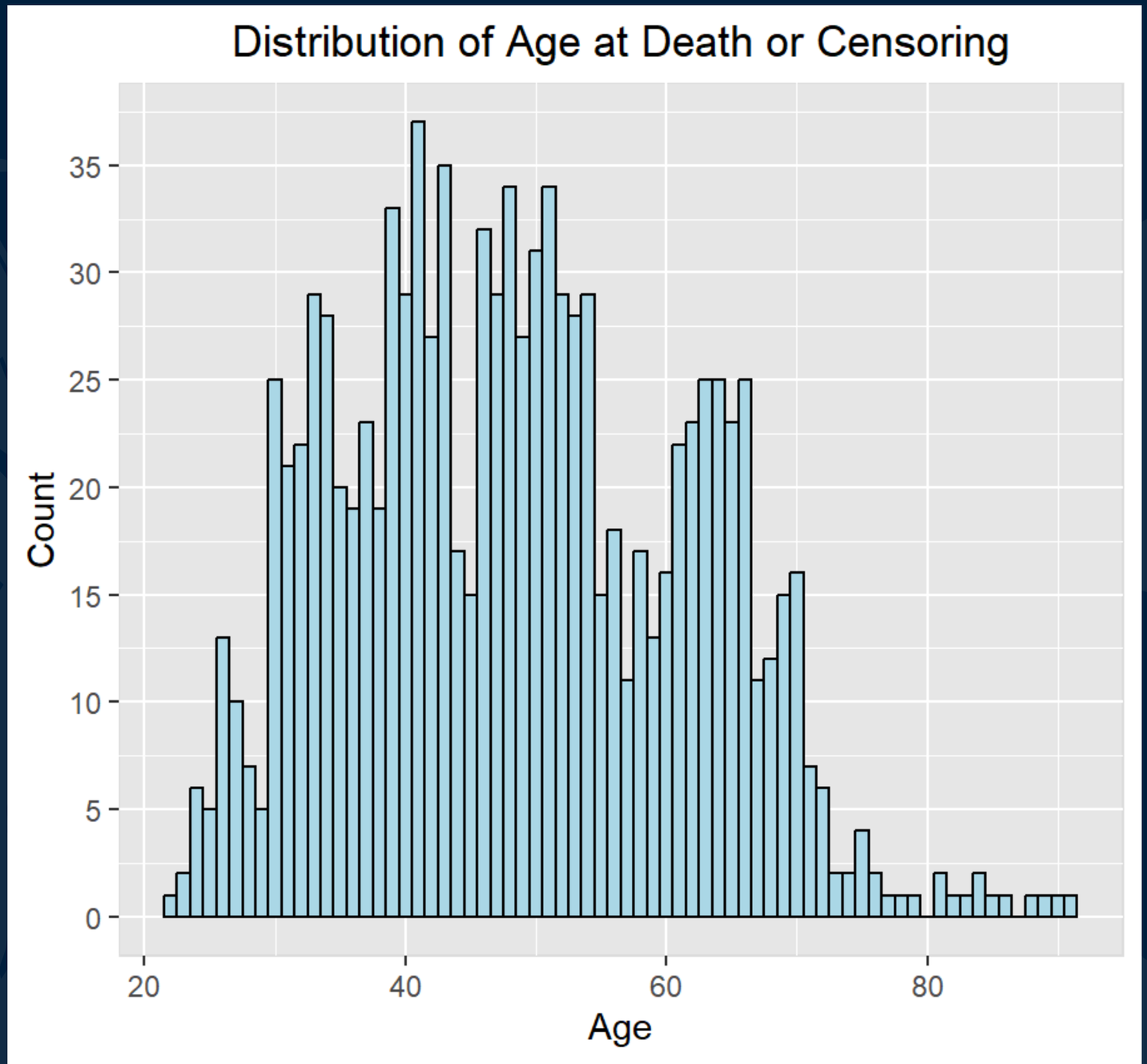
DISTRIBUTION OF AGE (ALIVE) AT CENSORING



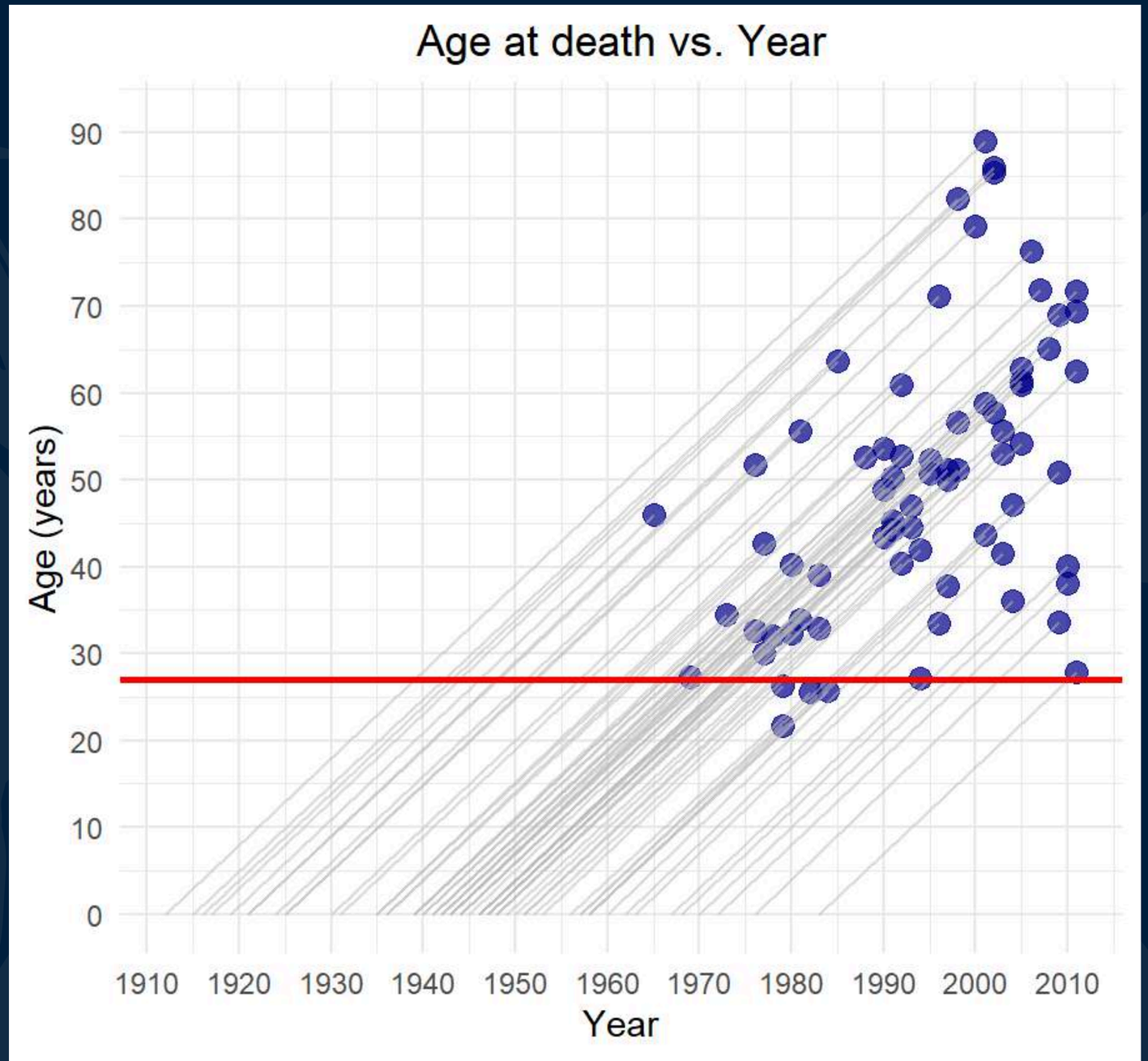
DISTRIBUTION OF AGE AT DEATH



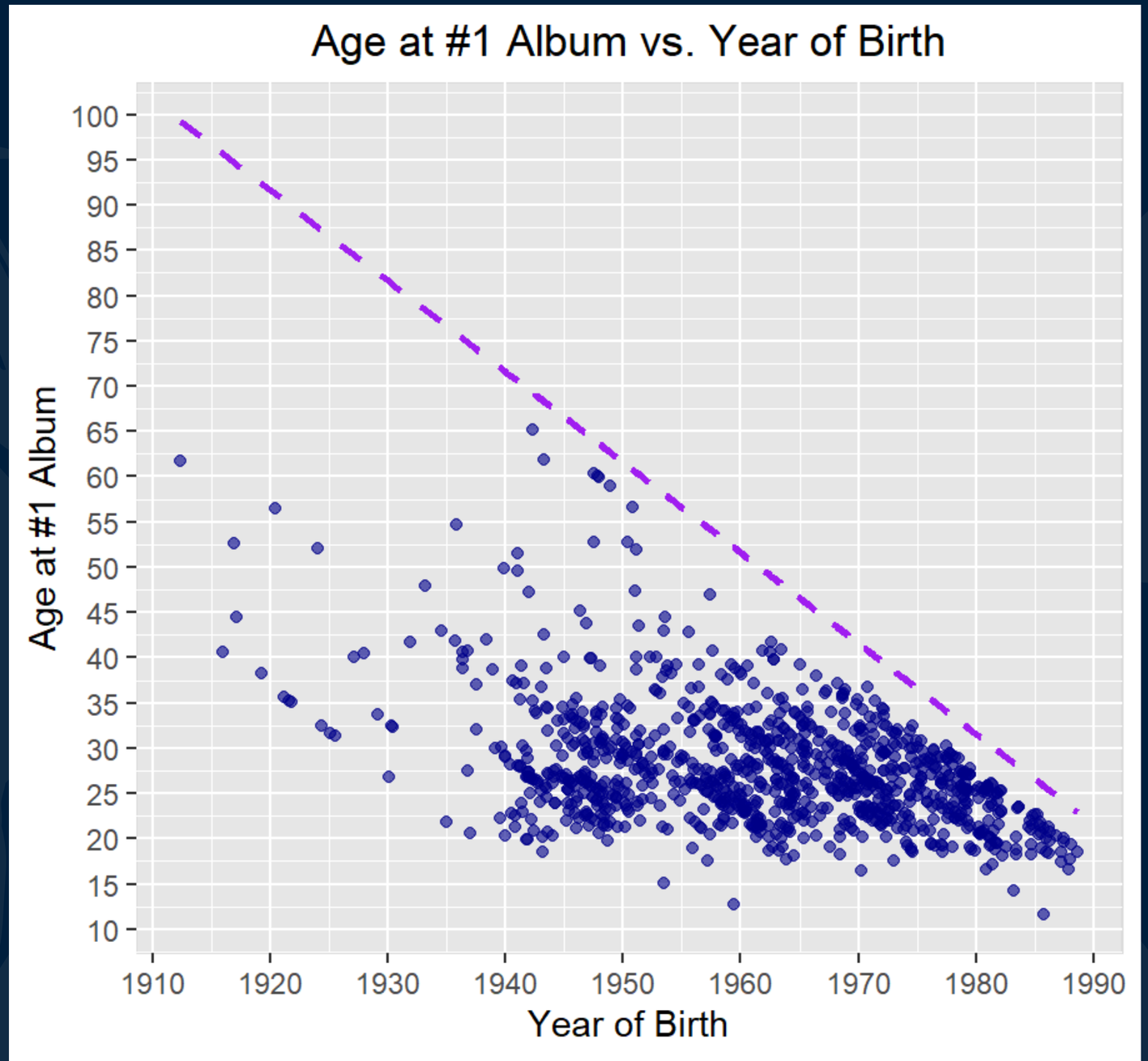
DISTRIBUTION OF AGE AT DEATH OR CENSORING



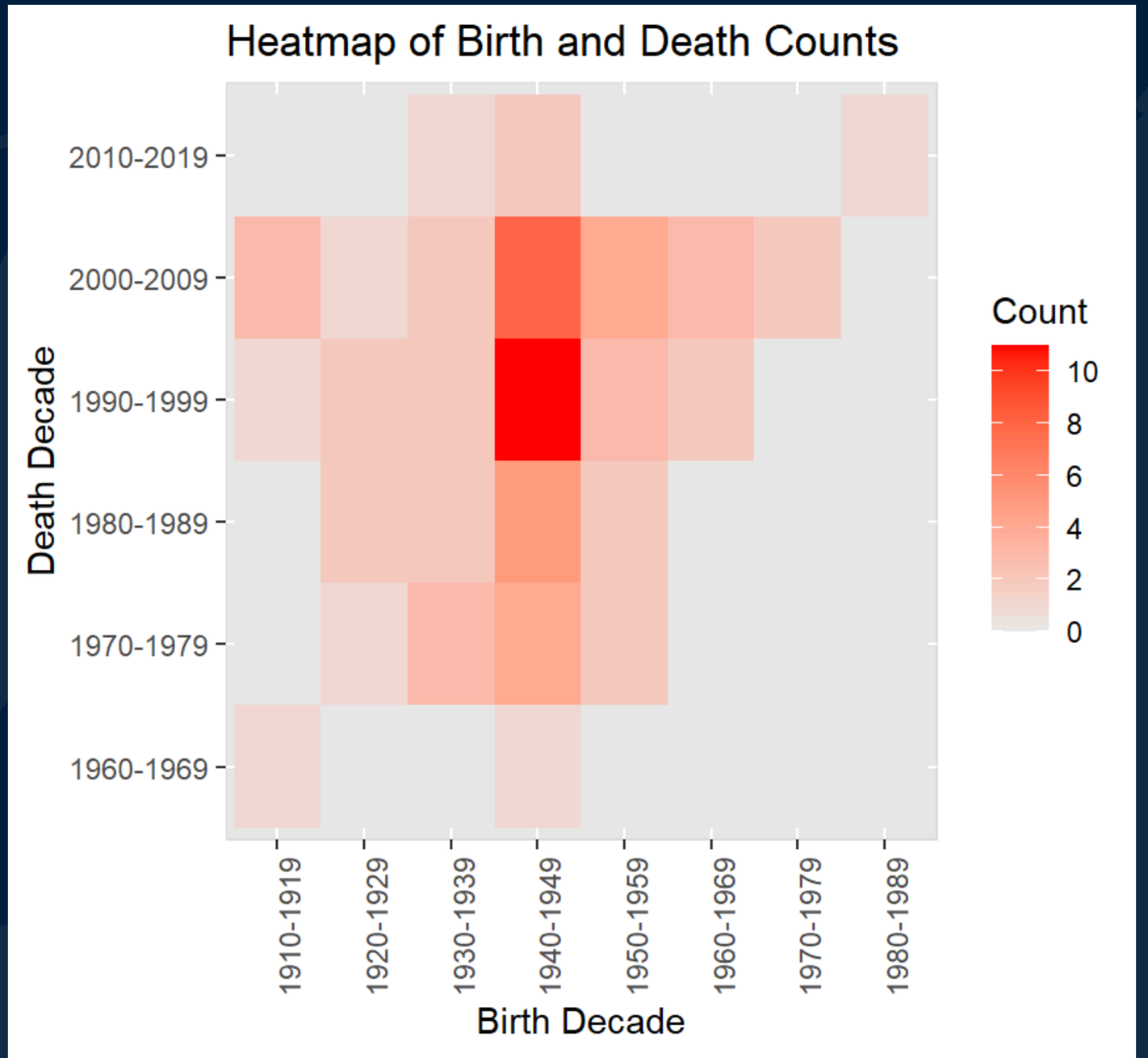
AGE AT DEATH VS. YEAR

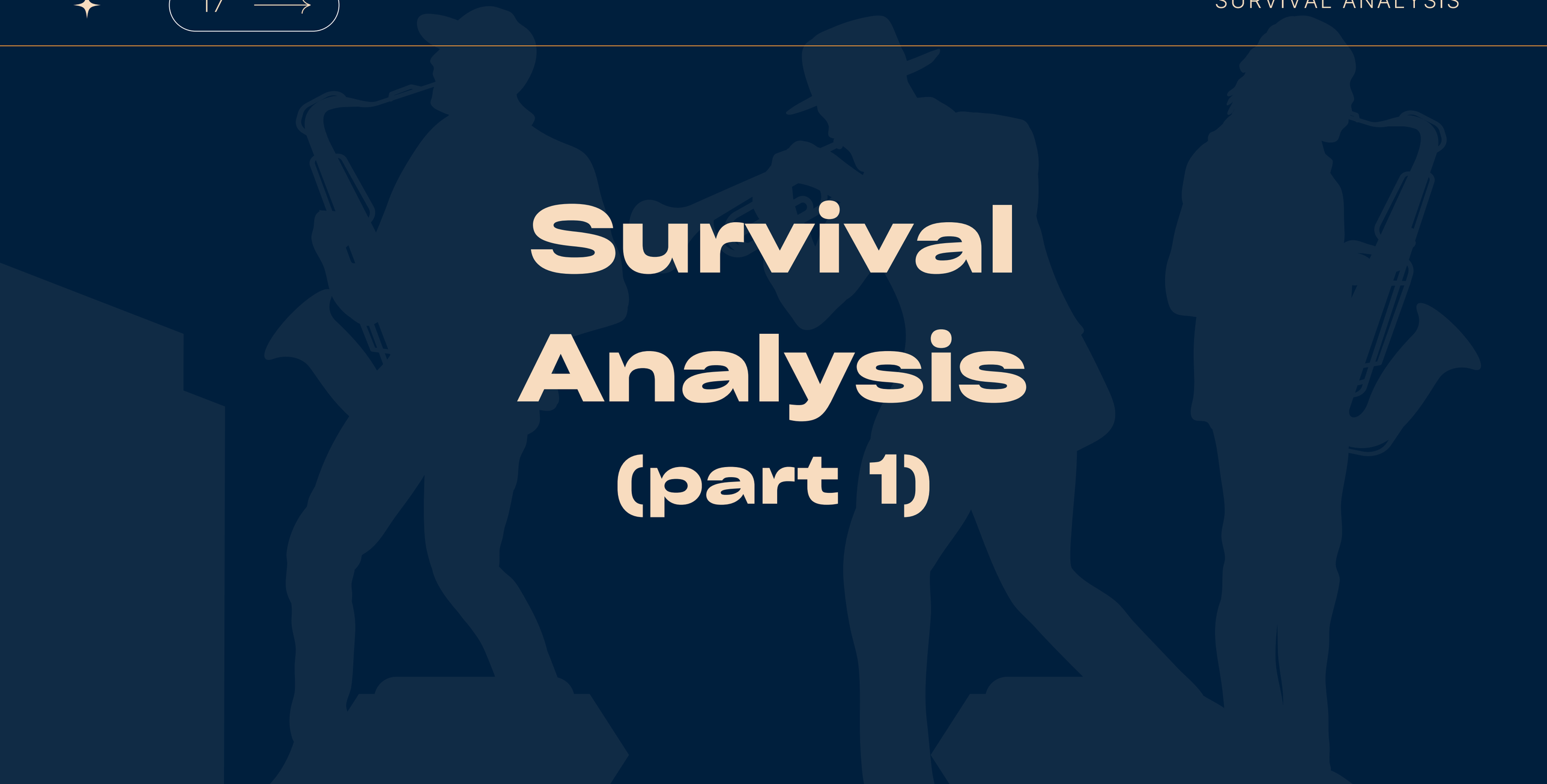


AGE AT #1 ALBUM VS. YEAR OF BIRTH



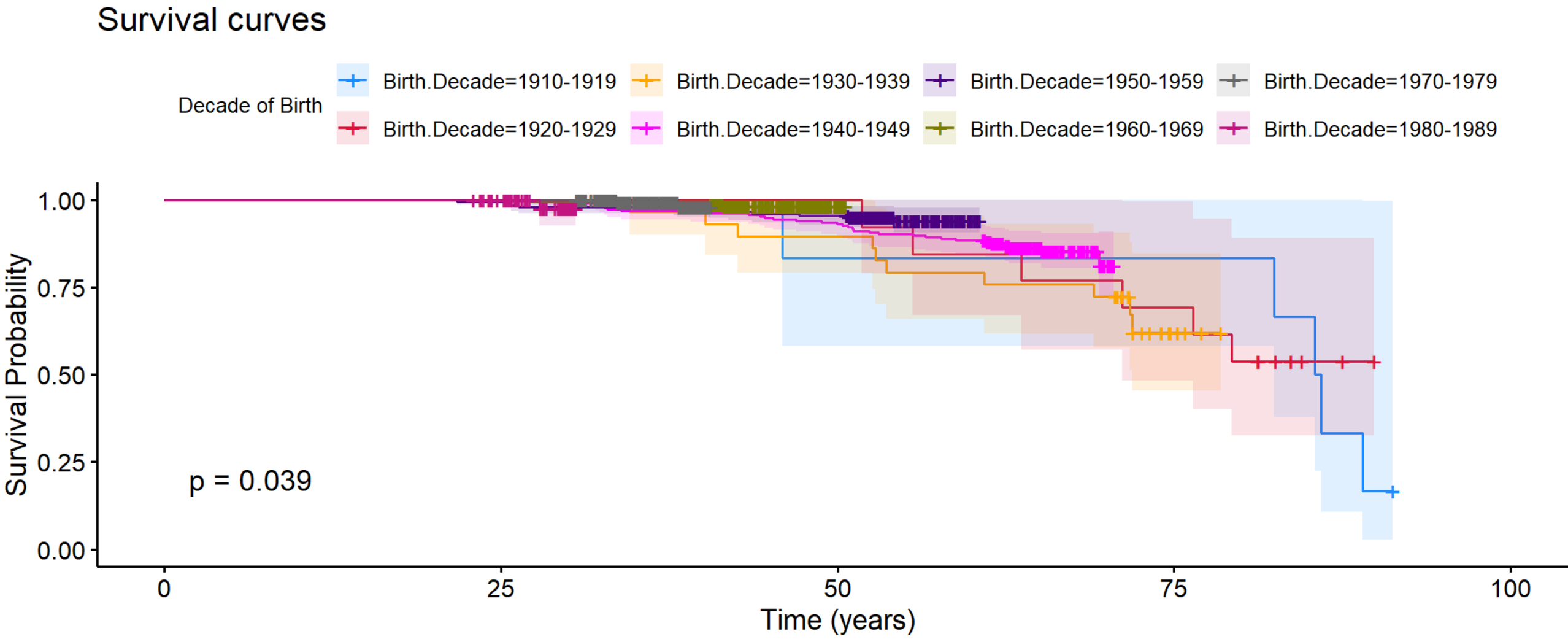
HEATMAP OF BIRTH AND DEATH COUNTS





Survival Analysis (part 1)

SURVIVAL CURVES



- p-value = 0.039, so we reject the null hypothesis.
- So here are significant statistical differences between the survival curves.

COMPARISON SURVIVAL CURVES FOR EACH DECADE

a) Log-rank test

```
survdiff_test_logrank <- survdiff(Surv.object.age.censored ~
Birth.Decade, data = data, rho = 0)
```

p-value = 0.03876264

The musician born in different decades have different probability of long life.

b) Peto-Peto test

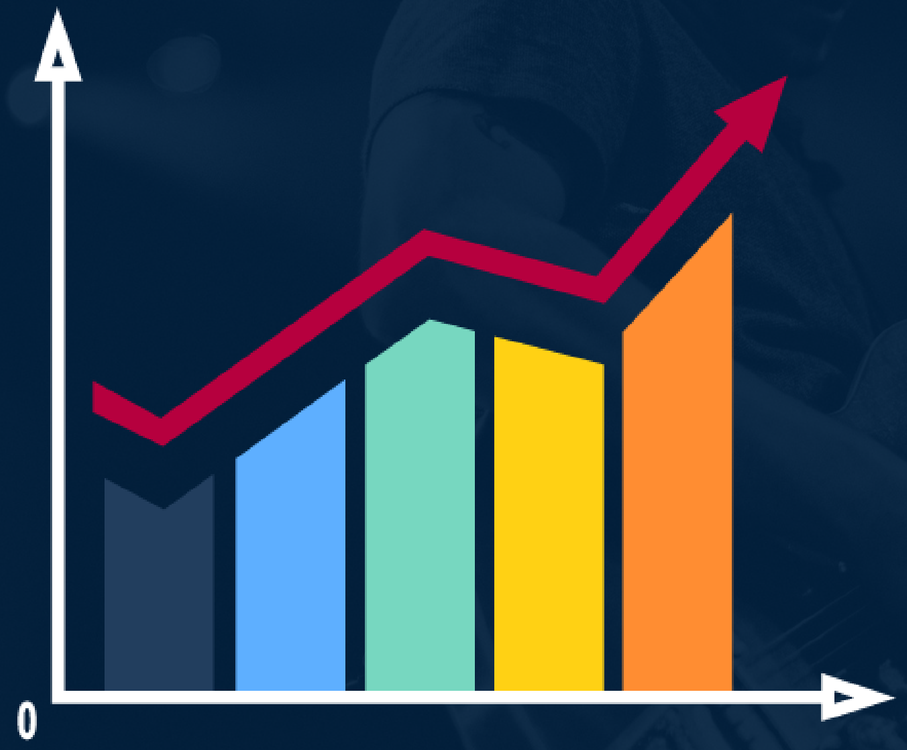
```
survdiff_test_petopeto <- survdiff(Surv.object.age.censored ~
Birth.Decade, data = data, rho = 1)
```

p-value = 0.03963116

The musician born in different decades have different probability of long life.

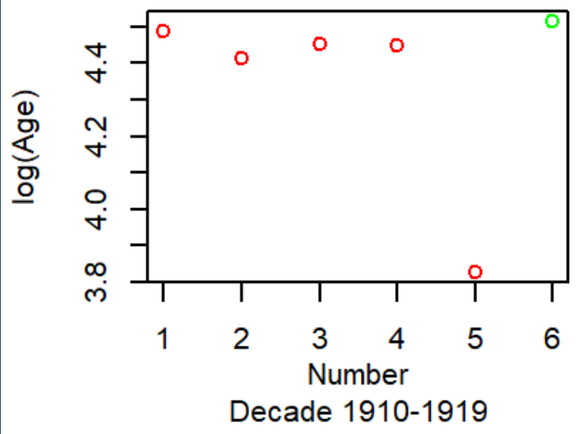
COMPARISON SURVIVAL CURVES FOR EACH DECADE

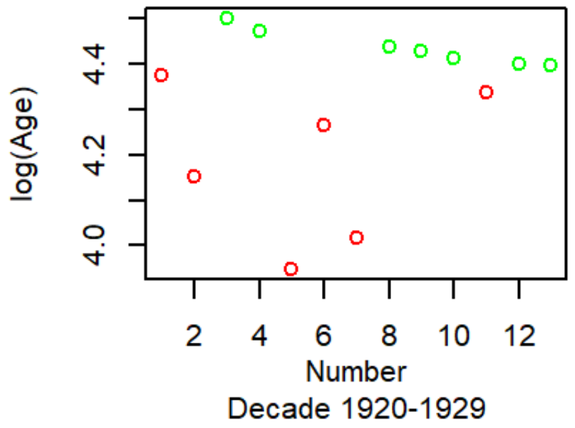
- We use these tests because we have censored data, and we want to compare survival curves using non-parametric tests.
- Both of these tests included censored data in their calculations.

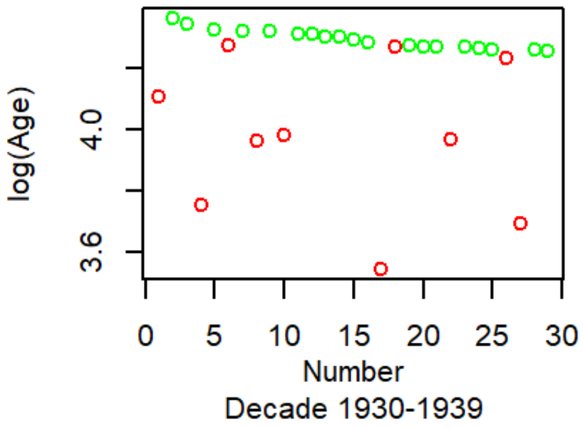


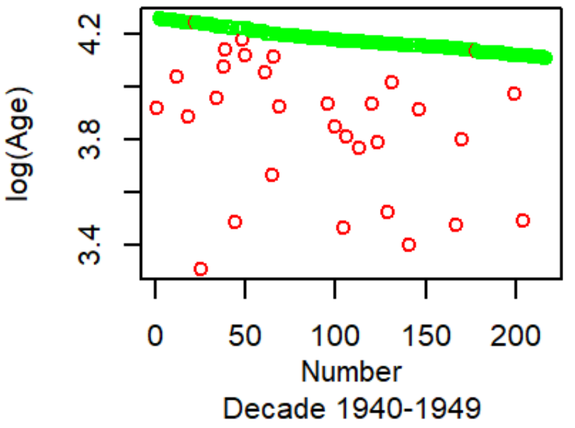
COMPARE SURVIVAL CURVES USING REGRESSION MODEL

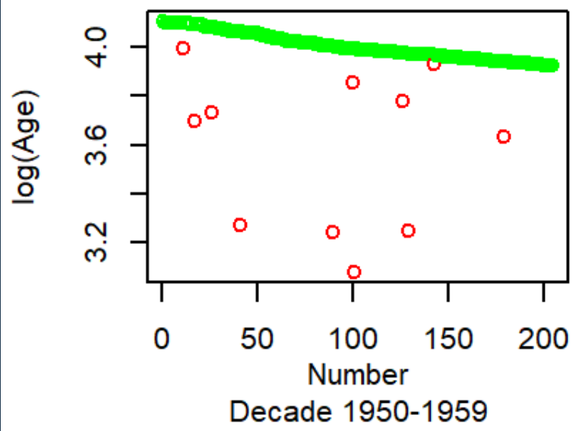
Behavior of log(Age.censored) for each decade.

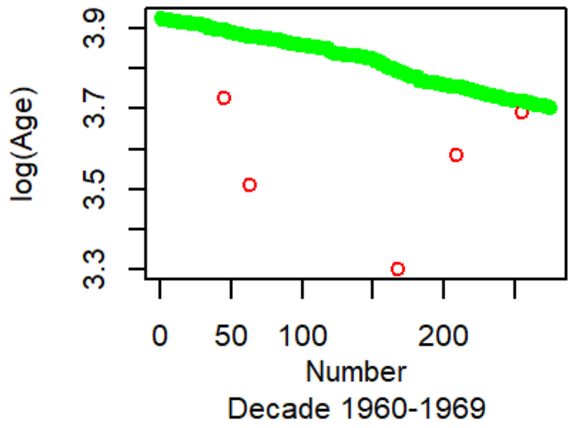


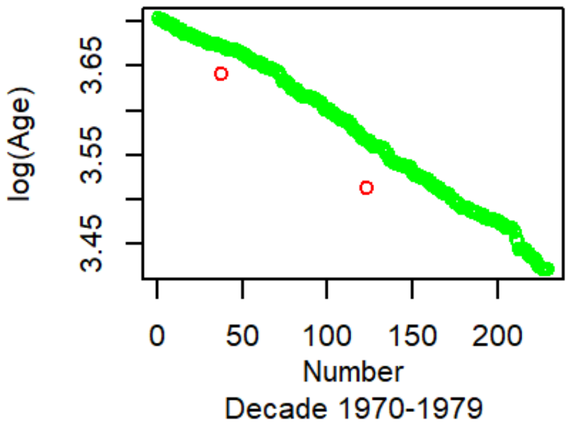


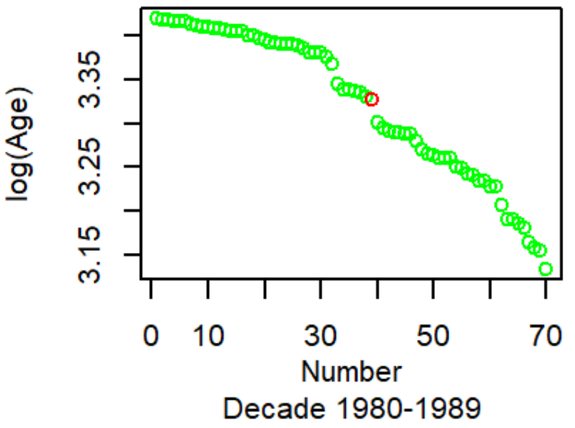












COMPARISION SURVIVAL CURVES USING REGRESSION MODEL

- Censored data mostly align in a straight line, while data on time of death do not exhibit this pattern.
- It’s difficult to determine which model we should choose.

COMPARISON SURVIVAL CURVES USING REGRESSION MODEL

- weimodel <- survreg(Surv.object.age.censored ~ Birth.Decade, data = data, dist = "weibull")
- lognmodel <- survreg(Surv.object.age.censored ~ Birth.Decade, data = data, dist = "lognormal")
- loglmodel <- survreg(Surv.object.age.censored ~ Birth.Decade, data = data, dist = "loglogistic")

AIC	
weimodel	950.0137
lognmodel	953.2168
loglmodel	951.9822

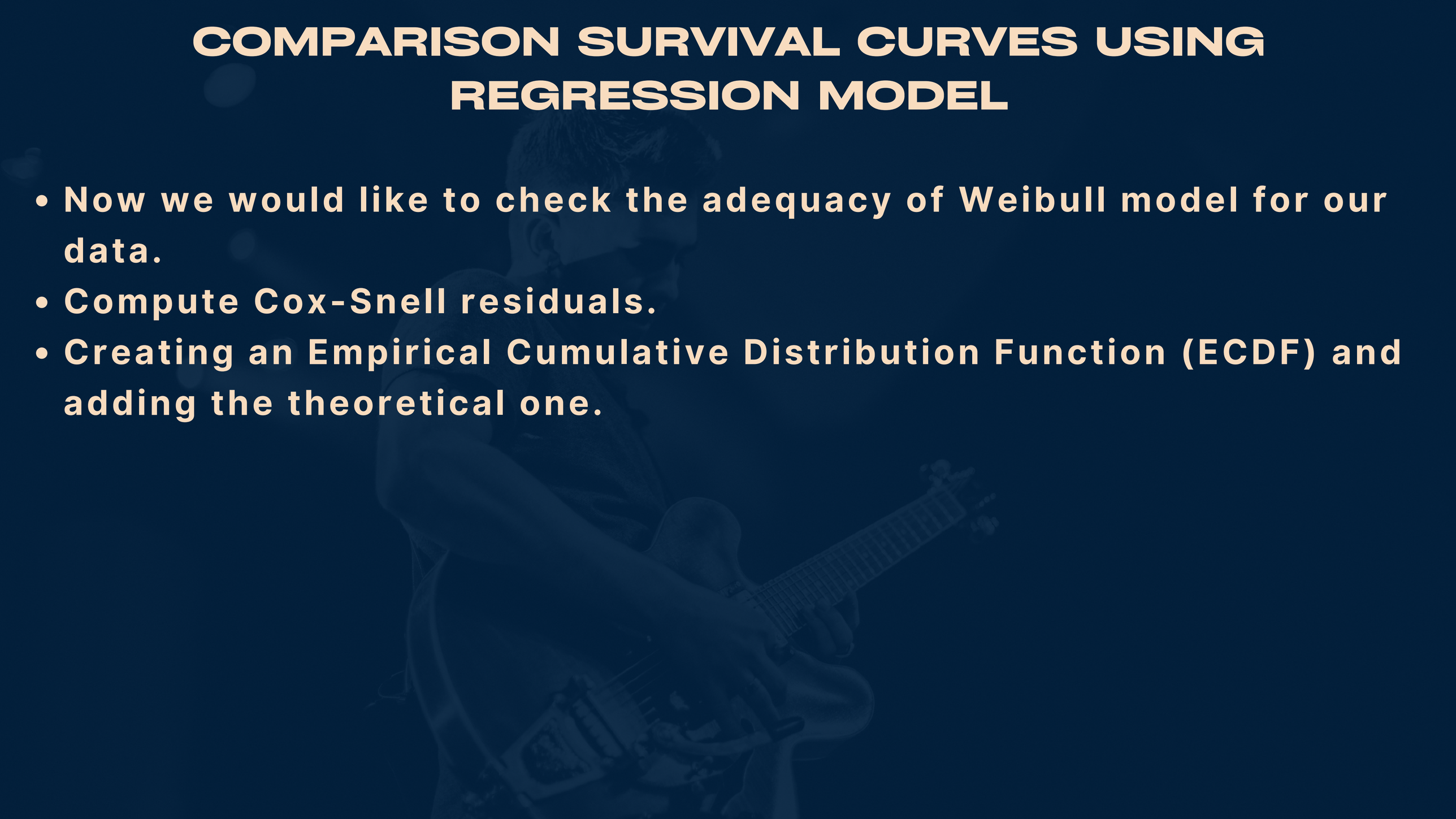
COMPARISON SURVIVAL CURVES USING REGRESSION MODEL

	Value	Std. Error	z	p
(Intercept)	4.4716	0.1304	34.29	<2e-16
Birth.Decade1920-1929	0.1101	0.1777	0.62	0.536
Birth.Decade1930-1939	0.0754	0.1624	0.46	0.643
Birth.Decade1940-1949	0.2504	0.1531	1.64	0.102
Birth.Decade1950-1959	0.3797	0.1831	2.07	0.038
Birth.Decade1960-1969	0.5217	0.2257	2.31	0.021
Birth.Decade1970-1979	0.4984	0.2887	1.73	0.084
Birth.Decade1980-1989	0.1004	0.3479	0.29	0.773
Log(scale)	-1.2325	0.1140	-10.81	<2e-16

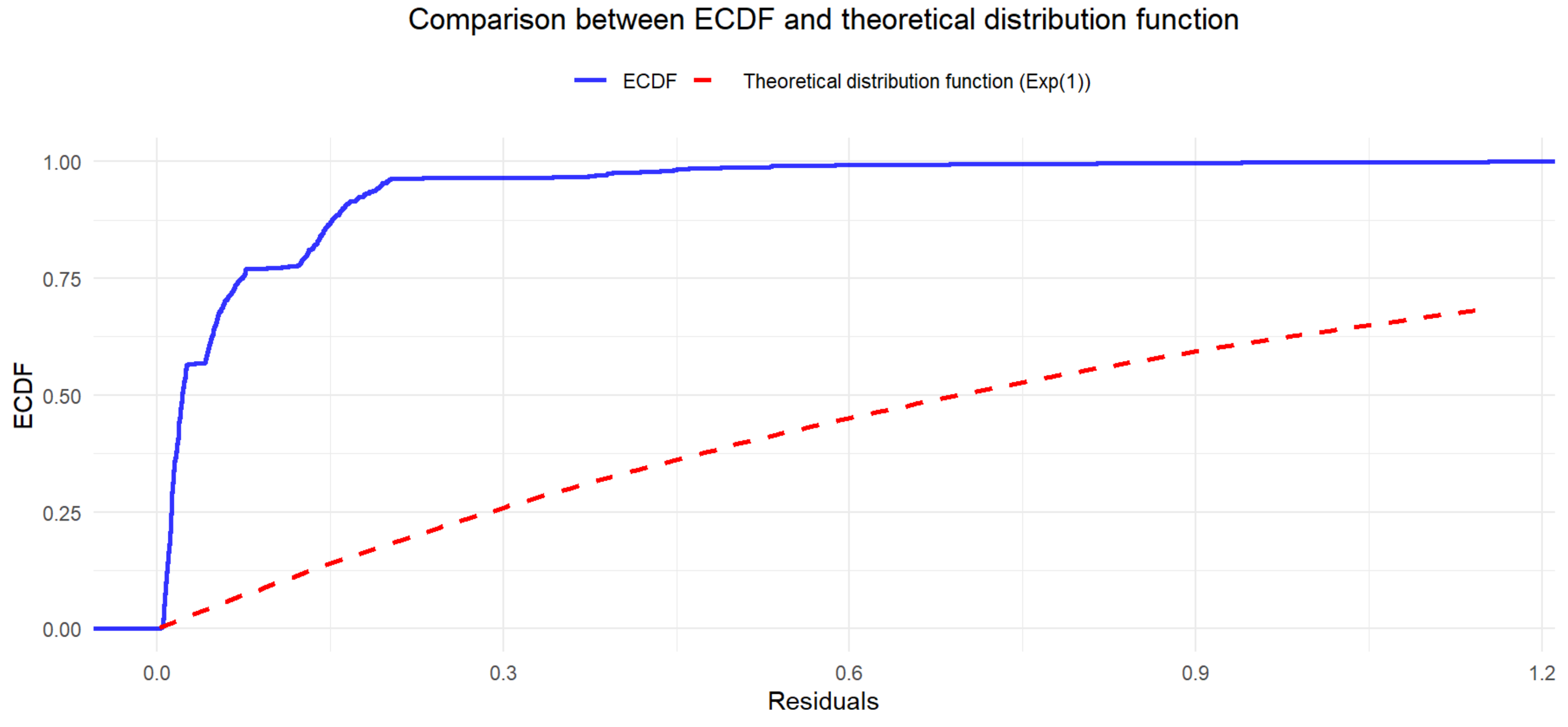
- The birth decade has an impact on survival time, particularly for individuals born in the 1950-1959 and 1960-1969 periods, where the results are statistically significant.
- The other decades do not show significant differences.
- But we need to check the model assumptions to ensure our conclusions are valid.

COMPARISON SURVIVAL CURVES USING REGRESSION MODEL

- Now we would like to check the adequacy of Weibull model for our data.
- Compute Cox-Snell residuals.
- Creating an Empirical Cumulative Distribution Function (ECDF) and adding the theoretical one.



COMPARISON BETWEEN ECDF AND THEORETICAL DISTRIBUTION FUNCTION



COMPARISON BETWEEN ECDF AND THEORETICAL DISTRIBUTION FUNCTION

Cox-Snel residuals

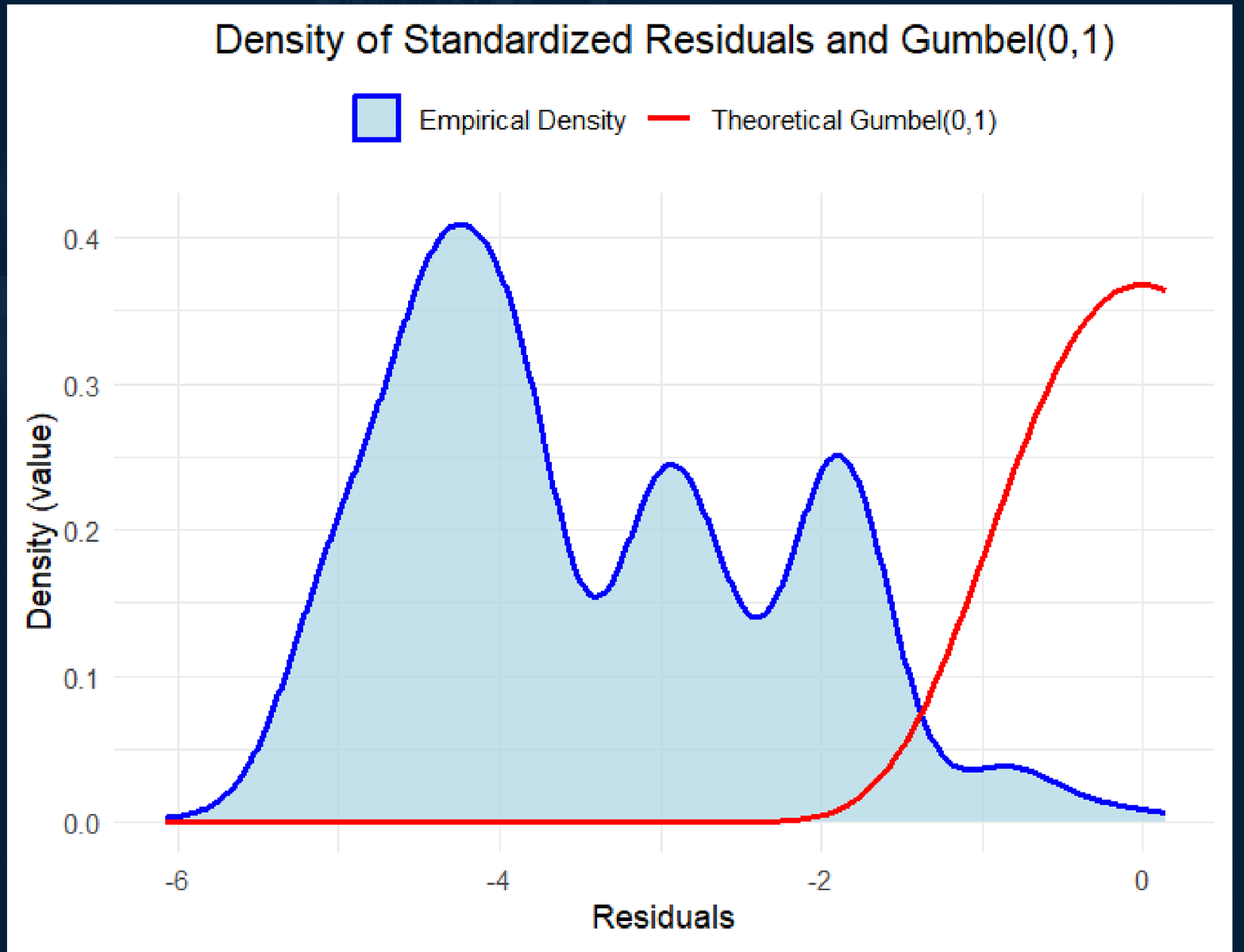
- Comparison between cox snell residuals and exponential theoretical distribution with rate = 1. (ks.test)

p-value < 2.2e-16

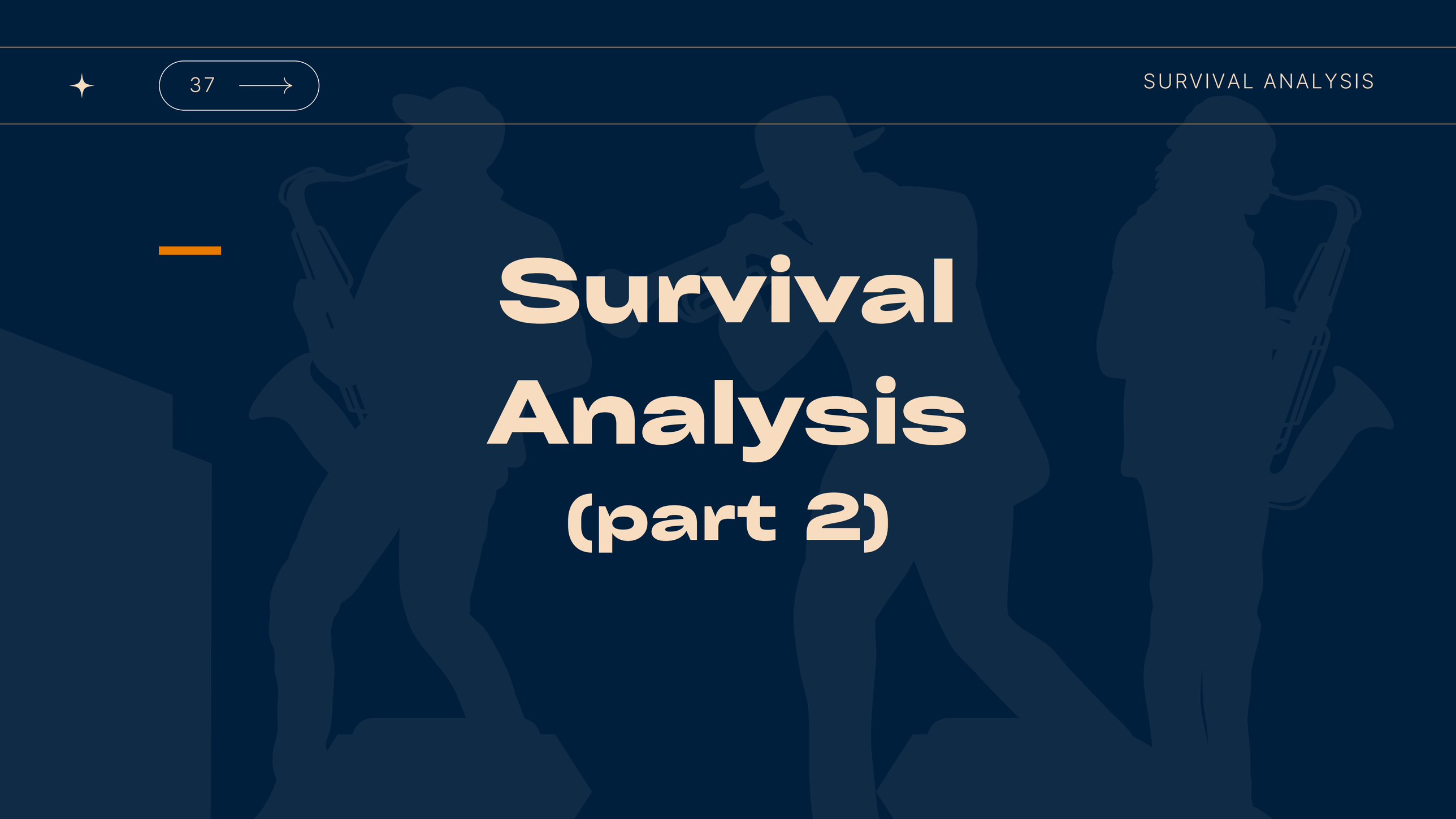
Standardized residuals

- Comparison standardized residuals and density of Gumbel(0,1).

DENSITY OF STANDARIZED RESIDUALS AND GUMBEL(0,1)

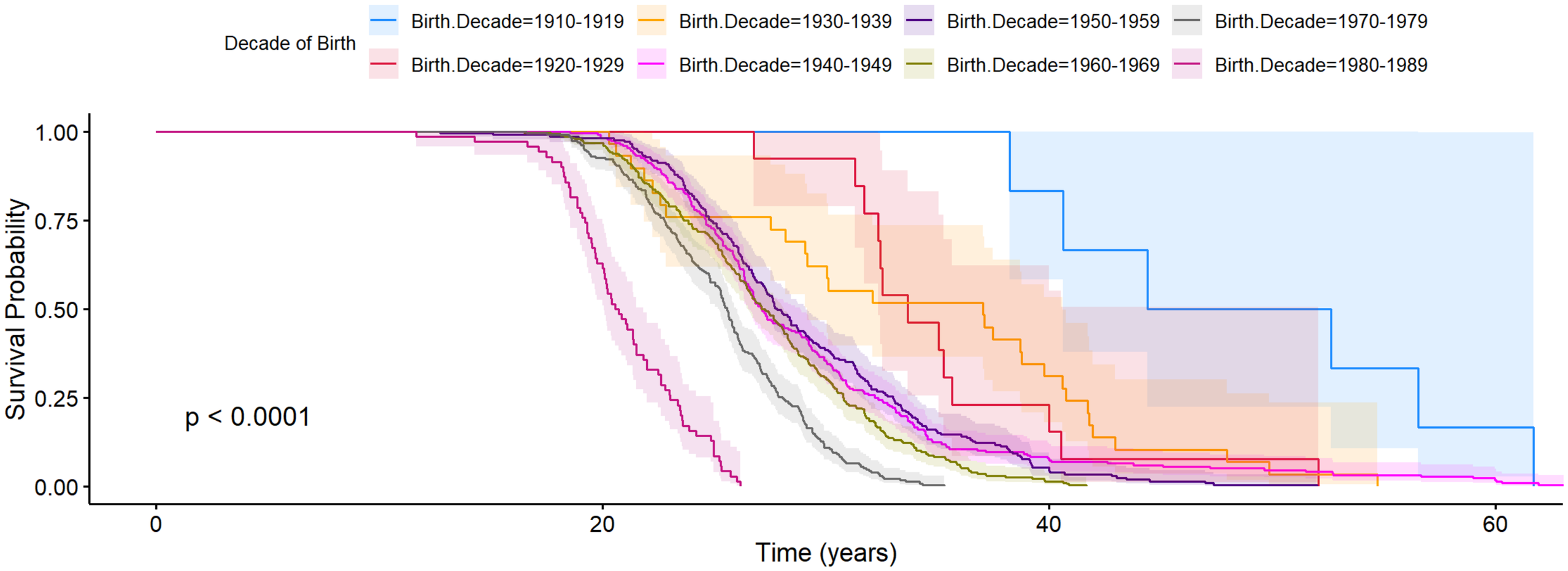


- Based on the plot and the results of the KS test, it appears that the Weibull model is inadequate for this data.
- As the model assumptions are not satisfied, the conclusions drawn from this model may not be entirely accurate or reliable.
- Additionally, the uncensored sample is relatively small and includes only about 70 observations, which is why the obtained results are not fully sufficient.
- Therefore, we need to use another (for example non-parametric) test to draw valid conclusions from this data.



Survival Analysis (part 2)

SURVIVAL CURVES



- **p-value < 0.0001, so we reject the null hypothesis.**
- **We can say that survival curves for each decade are different.**

COMPARISON SURVIVAL CURVES FOR EACH DECADE

a) Log-rank test

```
survdiff_test <- survdiff(Surv.object.time.number.one ~ Birth.Decade,
                          data = data, rho = 0)
```

p-value = 5.155283e-92

b) Peto-Peto test

```
survdiff_test <- survdiff(Surv.object.time.number.one ~ Birth.Decade,
                          data = data, rho = 1)
```

p-value = 1.097146e-81

c) Kruskal-Wallis test

```
kruskal.test(Time.number.one ~ Birth.Decade, data = data)
```

p-value < 2.2e-16

COMPARISON SURVIVAL CURVES FOR EACH DECADE

We can also use KS test and Kruskal-Wallis test because we have complete data at this case.

- We have to remember about correction due to multiple testing (We used Bonferroni Correction).

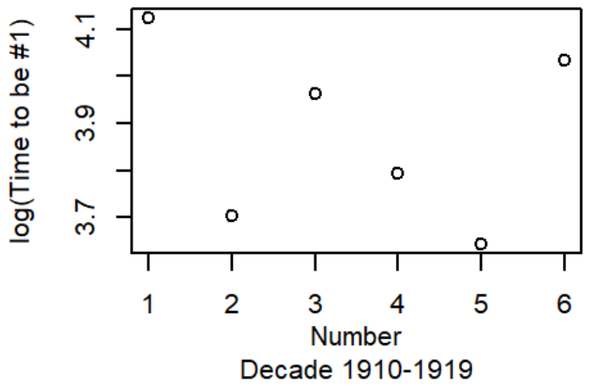
Results

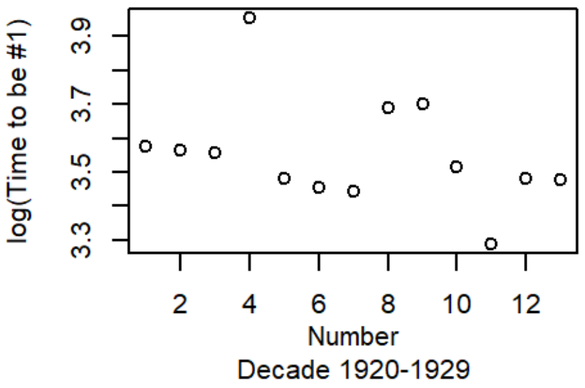
The Log-rank test, Peto-Peto test, and Kruskal-Wallis test indicate that the time to reach number 1 on the UK charts is significantly different for each decades.

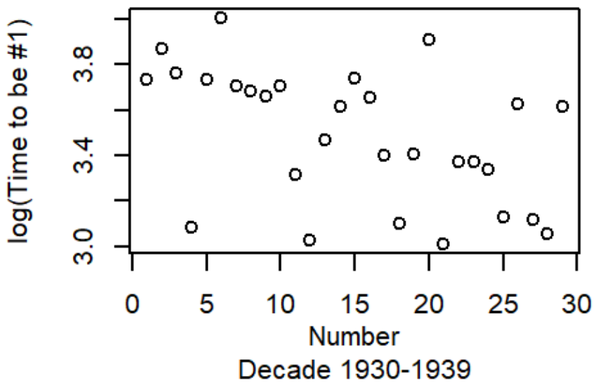
However, the KS test suggests that some decades have similar times to reach number one on the UK charts.

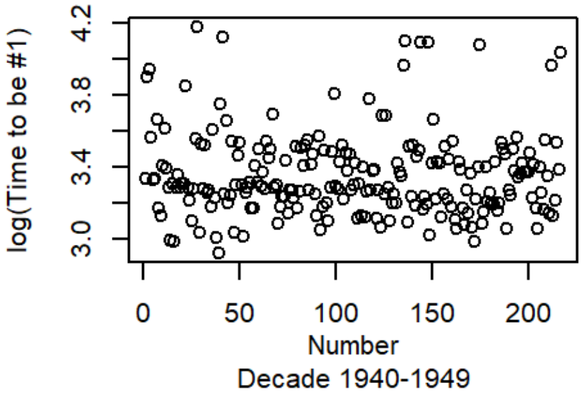
REGRESSION MODELS

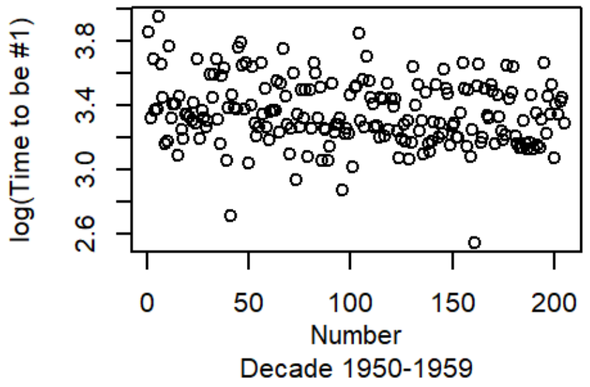
Let's have a look how behave observations (time to be #1) for each decade.

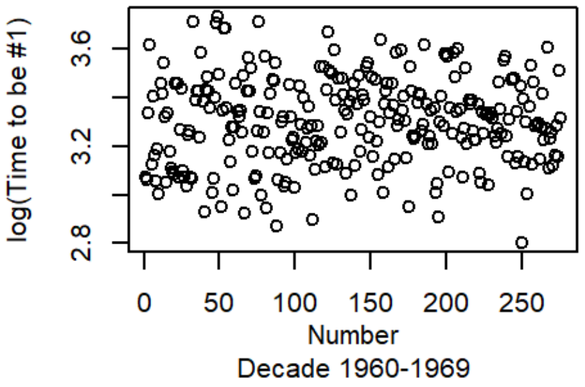


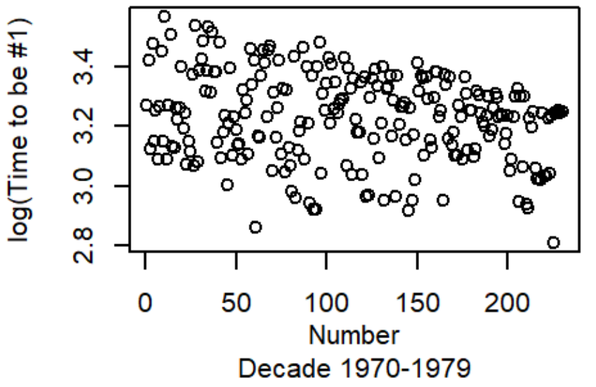


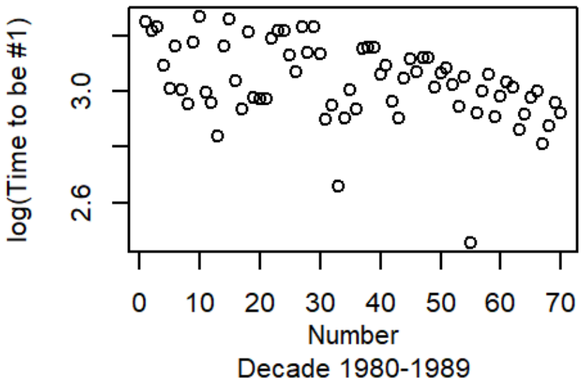








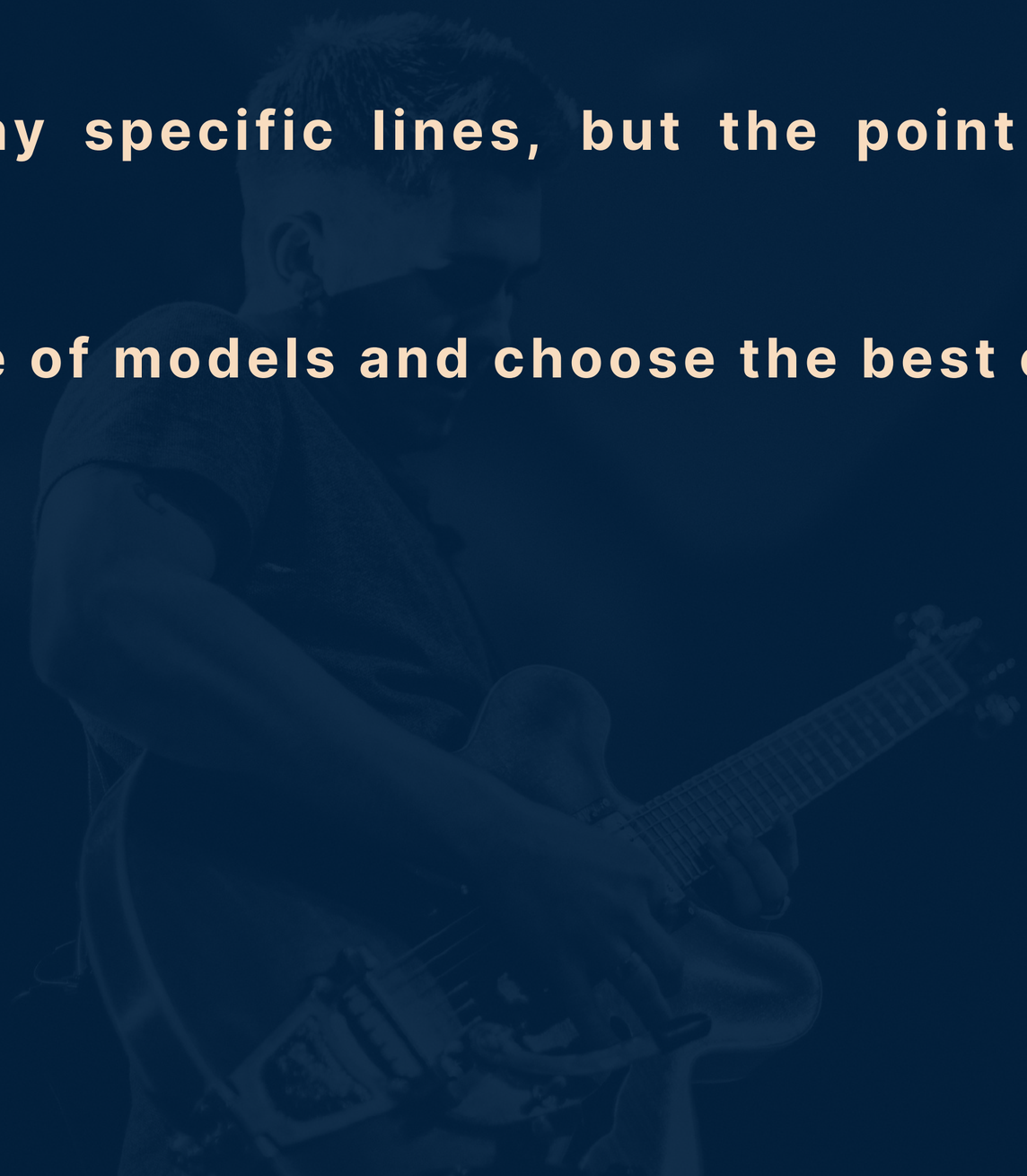




REGRESSION MODELS

We don't see any specific lines, but the point clouds form distinct shapes.

Let's check some of models and choose the best one, using AIC.



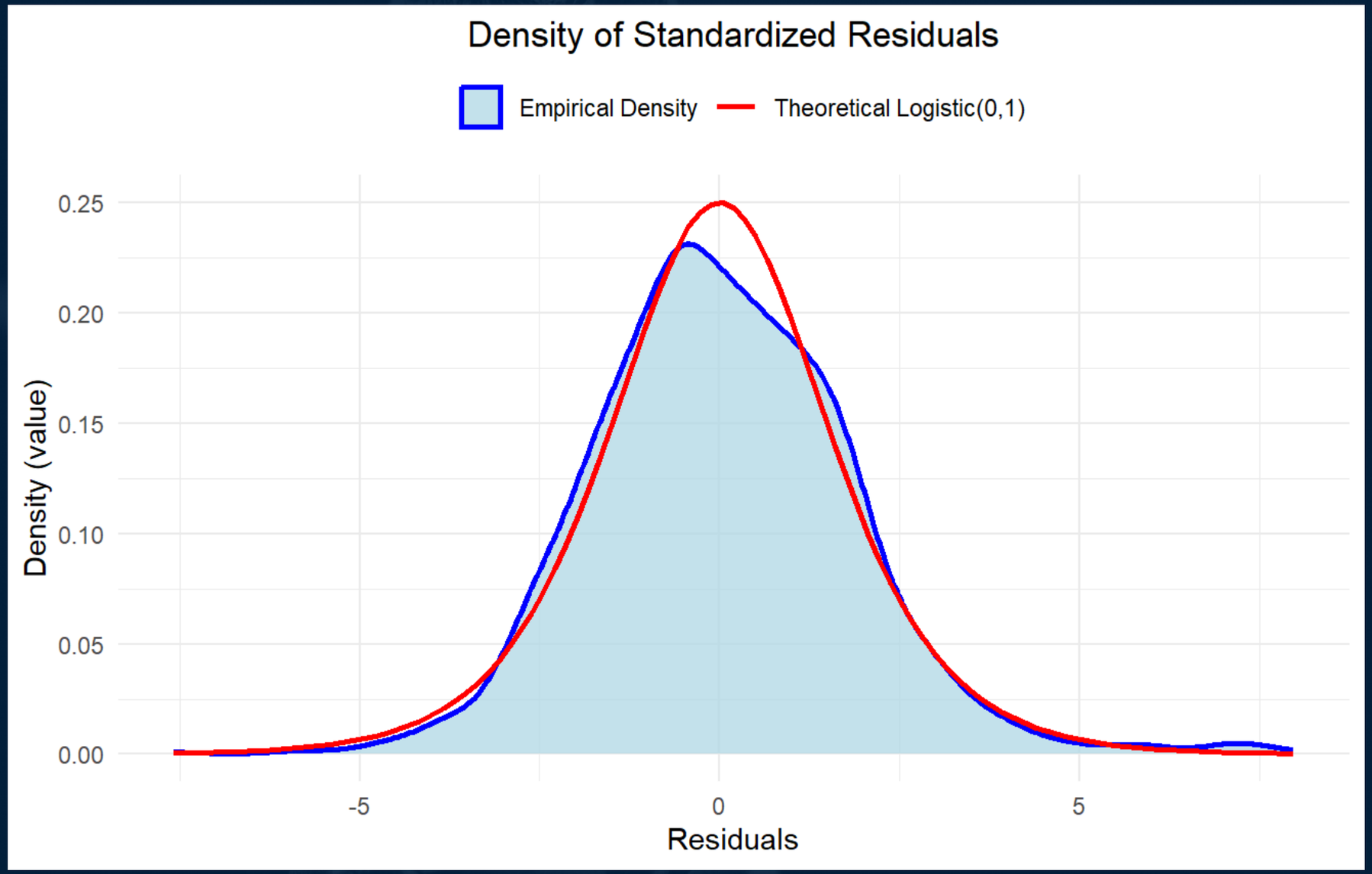
REGRESSION MODELS

```
weimodel <- survreg(Surv.object.time.number.one ~ Birth.Decade,
data = data, dist = "weibull")

lognmodel <- survreg(Surv.object.time.number.one ~ Birth.Decade,
data = data, dist = "lognormal")

loglmodel <- survreg(Surv.object.time.number.one ~ Birth.Decade,
data = data, dist = "loglogistic")
```

AIC	
weimodel	6747.949
lognmodel	6440.870
loglmodel	6408.636

REGRESSION MODELS - “LOGLMODEL”

Let's check the distribution of the residuals in “loglmodel”.



REGRESSION MODELS - “LOGLMODEL”

We conducted a Kolmogorov-Smirnov test (KS test) to check whether the standardized residuals follow a logistic distribution with parameters location = 0 and scale = 1.

p-value = 0.6048

Hence, the standardized residuals follow a Logistic(0,1) distribution, indicating that the log-logistic model is adequate.

REGRESSION MODELS - “LOGLMODEL”

	Value	Std. Error	z	p
(Intercept)	3.8764	0.0797	48.66	< 2e-16
Birth.Decade1920-1929	-0.3370	0.0927	-3.63	0.00028
Birth.Decade1930-1939	-0.3605	0.0904	-3.99	6.7e-05
Birth.Decade1940-1949	-0.5469	0.0807	-6.78	1.2e-11
Birth.Decade1950-1959	-0.5284	0.0808	-6.54	6.0e-11
Birth.Decade1960-1969	-0.5747	0.0804	-7.14	9.1e-13
Birth.Decade1970-1979	-0.6457	0.0805	-8.02	1.1e-15
Birth.Decade1980-1989	-0.8375	0.0822	-10.18	< 2e-16
Log(scale)	-2.2427	0.0257	-87.19	< 2e-16

- All variables in the model are significant according to the Wald test.
- Therefore, we can conclude that the times to reach #1 album are significantly different for each decades.

CONCLUSIONS

- We consider nonparametric methods to be better because they do not rely on distributional assumptions, which in our case turned out to be quite significant.

- In our opinion, parametric tests are better in situations where we have a large amount of uncensored (complete) data because they provide more accurate results and allow us to approximate distributions more effectively.

BIBLIOGRAPHY

- Wolkewitz M, Allignol A, Graves N, Barnett A G. *Is 27 really a dangerous age for famous musicians?* Retrospective cohort study BMJ 2011; 343 :d7799 doi:10.1136/bmj.d7799

- Michalos, A.C., *Encyclopedia of quality of life and well-being research*. Dordrecht: Springer, 2014.

- Bennett S., *Log-Logistic Regression Models for Survival Data*, Journal of the Royal Statistical Society. Series C (Applied Statistics), Vol. 32, No. 2 ,1983 , pp. 165-171.

Biostatistics

PROJECT 5

Survival Analysis

Iza Danielewska

Dawid Poławski

Warsaw, 27.01.2025