# C240B / STAT C245B
# Project Presentation I – Proposal

# Sociological statistical inference by survival analysis

Z. Tom Hu
Sijin Wu

Berkeley
UNIVERSITY OF CALIFORNIA

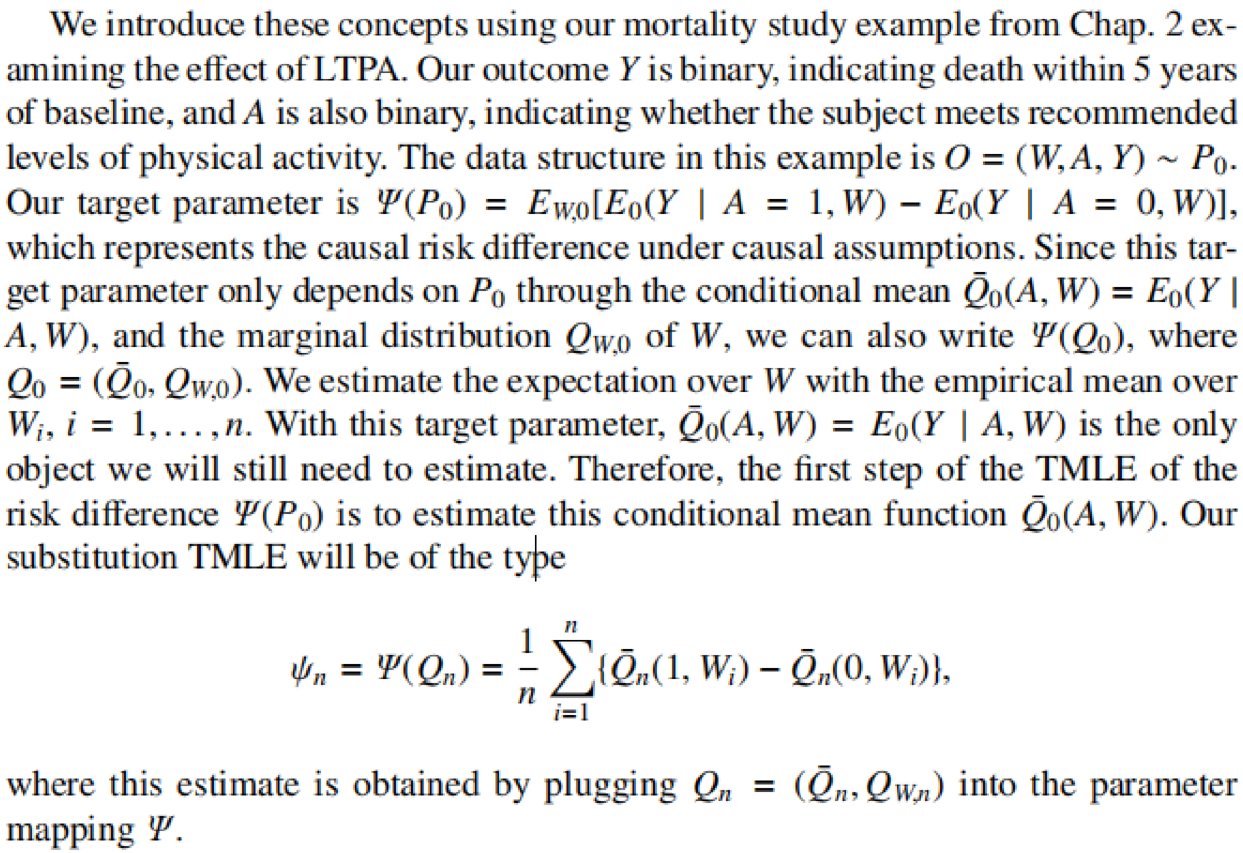
# Introduction

This is the first chapter in our text focused on estimation within the road map for targeted learning. Now that we’ve defifined the research question, including our data, the model, and the target parameter, we are ready to begin. For the estimation of a target parameter of the probability distribution of the data, such as target parameters that can be interpreted as causal effffects, we implement TMLE. The first step in this estimation procedure is an initial estimate of the data-generating distribution P0, or the relevant part Q0 of P0 that is needed to evaluate the target parameter.

# Background

Let’s start our discussion with studies where Y is binary, such as in our mortality study example. When Y is binary, there is no difference between the conditional mean or conditional probability distribution, so this distinction plays no role

# Background



We introduce these concepts using our mortality study example from Chap. 2 examining the effect of LTPA. Our outcome $Y$ is binary, indicating death within 5 years of baseline, and $A$ is also binary, indicating whether the subject meets recommended levels of physical activity. The data structure in this example is $O = (W, A, Y) \sim P_0$. Our target parameter is $\Psi(P_0) = E_{W,0}[E_0(Y \mid A = 1, W) - E_0(Y \mid A = 0, W)]$, which represents the causal risk difference under causal assumptions. Since this target parameter only depends on $P_0$ through the conditional mean $\bar{Q}_0(A, W) = E_0(Y \mid A, W)$, and the marginal distribution $Q_{W,0}$ of $W$, we can also write $\Psi(Q_0)$, where $Q_0 = (\bar{Q}_0, Q_{W,0})$. We estimate the expectation over $W$ with the empirical mean over $W_i$, $i = 1, \ldots, n$. With this target parameter, $\bar{Q}_0(A, W) = E_0(Y \mid A, W)$ is the only object we will still need to estimate. Therefore, the first step of the TMLE of the risk difference $\Psi(P_0)$ is to estimate this conditional mean function $\bar{Q}_0(A, W)$. Our substitution TMLE will be of the type

$$\psi_n = \Psi(Q_n) = \frac{1}{n}\sum_{i=1}^{n}\{\bar{Q}_n(1, W_i) - \bar{Q}_n(0, W_i)\},$$

where this estimate is obtained by plugging $Q_n = (\bar{Q}_n, Q_{W,n})$ into the parameter mapping $\Psi$.

# Data

**China Family Panel Studies（CFPS）**

https://opendata.pku.edu.cn/dataverse/CFPS

China Family Panel Studies (CFPS) is a nationally representative, annual longitudinal survey of Chinese communities, families, and individuals launched in 2010 by the Institute of Social Science Survey (ISSS) of Peking University, China.

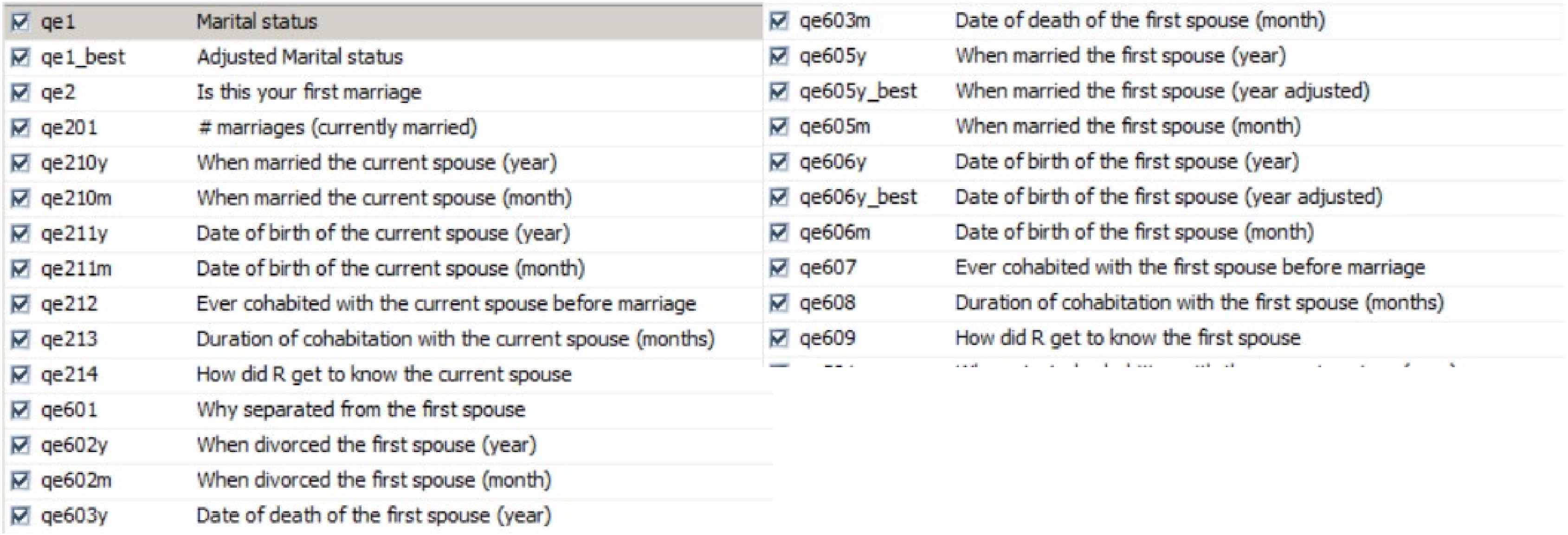
The CFPS is designed to collect individual-, family-, and community-level longitudinal data in contemporary China. The studies focus on the economic, as well as the non-economic, wellbeing of the Chinese population, with a wealth of information covering such topics as economic activities, education outcomes, family dynamics and relationships, migration, and health.

We have the whole CFPS data recorded in 2010, 2012, 2014, 2016, 2018 which can provide rich information to do social statistic inference. Each size of them is about 30000 individuals and 1000 variables.

# Data

The datasets are stored in the .dat format originally,  and for the convenience of our future analysis, we transported them into .csv format.

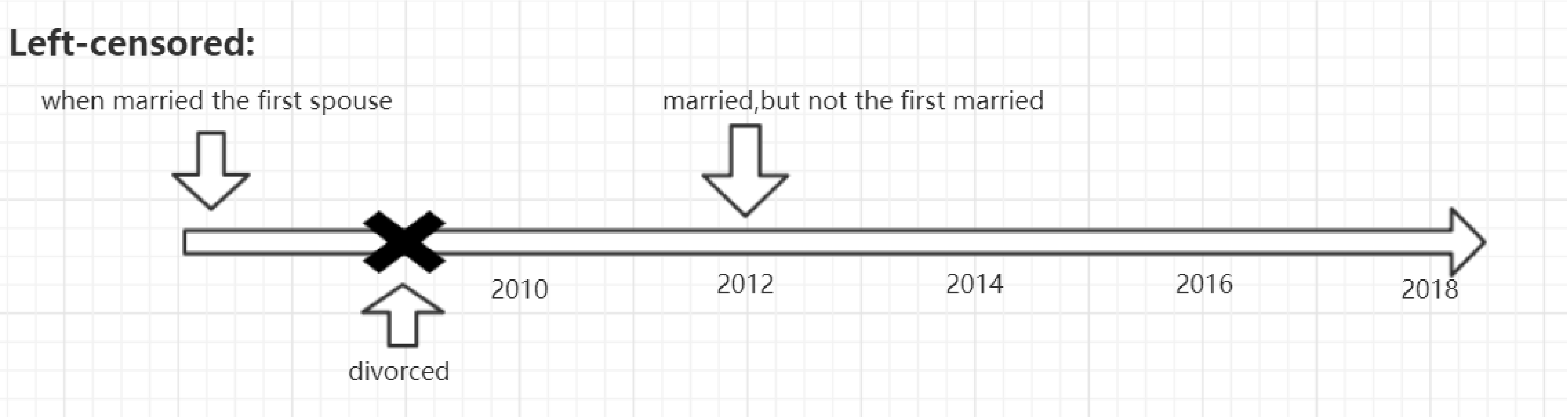
If we want to focus on marriage status, there will be specific information.



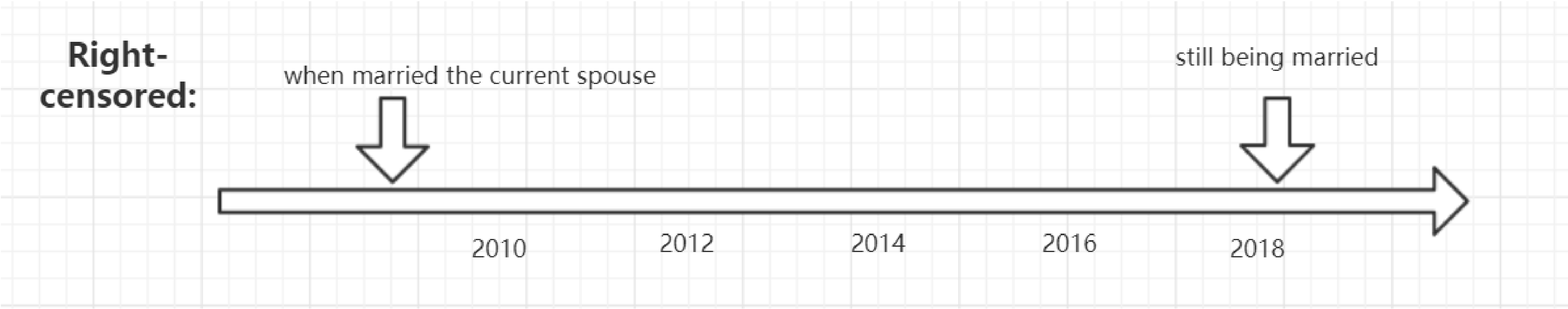
| | | |
|---|---|---|
| ☑ | qe1 | Marital status |
| ☑ | qe1_best | Adjusted Marital status |
| ☑ | qe2 | Is this your first marriage |
| ☑ | qe201 | # marriages (currently married) |
| ☑ | qe210y | When married the current spouse (year) |
| ☑ | qe210m | When married the current spouse (month) |
| ☑ | qe211y | Date of birth of the current spouse (year) |
| ☑ | qe211m | Date of birth of the current spouse (month) |
| ☑ | qe212 | Ever cohabited with the current spouse before marriage |
| ☑ | qe213 | Duration of cohabitation with the current spouse (months) |
| ☑ | qe214 | How did R get to know the current spouse |
| ☑ | qe601 | Why separated from the first spouse |
| ☑ | qe602y | When divorced the first spouse (year) |
| ☑ | qe602m | When divorced the first spouse (month) |
| ☑ | qe603y | Date of death of the first spouse (year) |
| ☑ | qe603m | Date of death of the first spouse (month) |
| ☑ | qe605y | When married the first spouse (year) |
| ☑ | qe605y_best | When married the first spouse (year adjusted) |
| ☑ | qe605m | When married the first spouse (month) |
| ☑ | qe606y | Date of birth of the first spouse (year) |
| ☑ | qe606y_best | Date of birth of the first spouse (year adjusted) |
| ☑ | qe606m | Date of birth of the first spouse (month) |
| ☑ | qe607 | Ever cohabited with the first spouse before marriage |
| ☑ | qe608 | Duration of cohabitation with the first spouse (months) |
| ☑ | qe609 | How did R get to know the first spouse |

# Censored Data

There are 2 kinds of censored data.


Left-censored:
when married the first spouse
married,but not the first married
divorced
2010
2012
2014
2016
2018

# Censored Data



# The target parameter

Take the length of a marriage as the example :

We can estimate the marriage survival function of one group of people who are born in the 1980s in Beijing with the same education conditions.

Also, we can compare the influence on marriages’ duration caused by different conditions.

# Following work

The current challenges will be re-arrange the data that we need and select the targeted variables.

The following work will be focused on how to use the survival analysis methods to inference the social events, like the marriage duration.

Once we complete the survival analysis by the targeted learning method, we want to use some other non-target learning methods to compare the results.