《数据分析基础》课程期末作业

中国人民大學
RENMIN UNIVERSITY OF CHINA

学历水平的综合分析：基于 CFPS 六次数据

作者：孙浩翔

学号：2023202305

学院：信息学院

分数

目录

1 研究背景

随着中国现代化的逐步深入，我国教育事业取得了显著进展，文盲率也持续下降 [2]. 但与此同时，学历贬值也在逐渐成为热门的社会话题.

中国家庭追踪调查（China Family Panel Studies，CFPS [1]）旨在通过跟踪收集个体、家庭、社区三个层次的数据，反映中国社会、经济、人口、教育和健康的变迁. 本文旨在通过分析 CFPS 的全部六次（2010, 2012, 2014, 2016, 2018, 2020）数据，讨论学历水平的分布、影响因素与变化趋势，从而更好地剖析整体学历水平，为解决相关问题提供科学依据与政策支持. 值得注意的是，本文中所研究的学历水平都是指“被调查人所取得的最高学历”，并没有包含肄业、在读等情况，因此不能与受教育水平完全对应. 同时，本数据分析的结论均基于 CFPS 数据集，因此结果可能受采样方法影响. 例如，CFPS 数据集在各省的采样数并没有按照各省人口数量按比例分配，如果偏远地区的样本数较多，会导致计算得到的全国文盲率偏高.

2　学历水平的分布特征

我们将学历水平划分为：

数值	标签	数值	标签
1	文盲/半文盲/幼儿园	5	大专
2	小学	6	大学本科
3	初中	7	硕士
4	高中/中专/职高/技校	8	博士

经过筛选后，CFPS 六次调查所涵盖的群体数均为 30000 左右，数据量足够支撑我们做对整体情况的分析. 下面，我们按省分析了 2020 年的成年人学历水平，具体如下（注：有个别省出现异常条形图的原因是样本数过少，例如海南省只有 3 个样本点）：

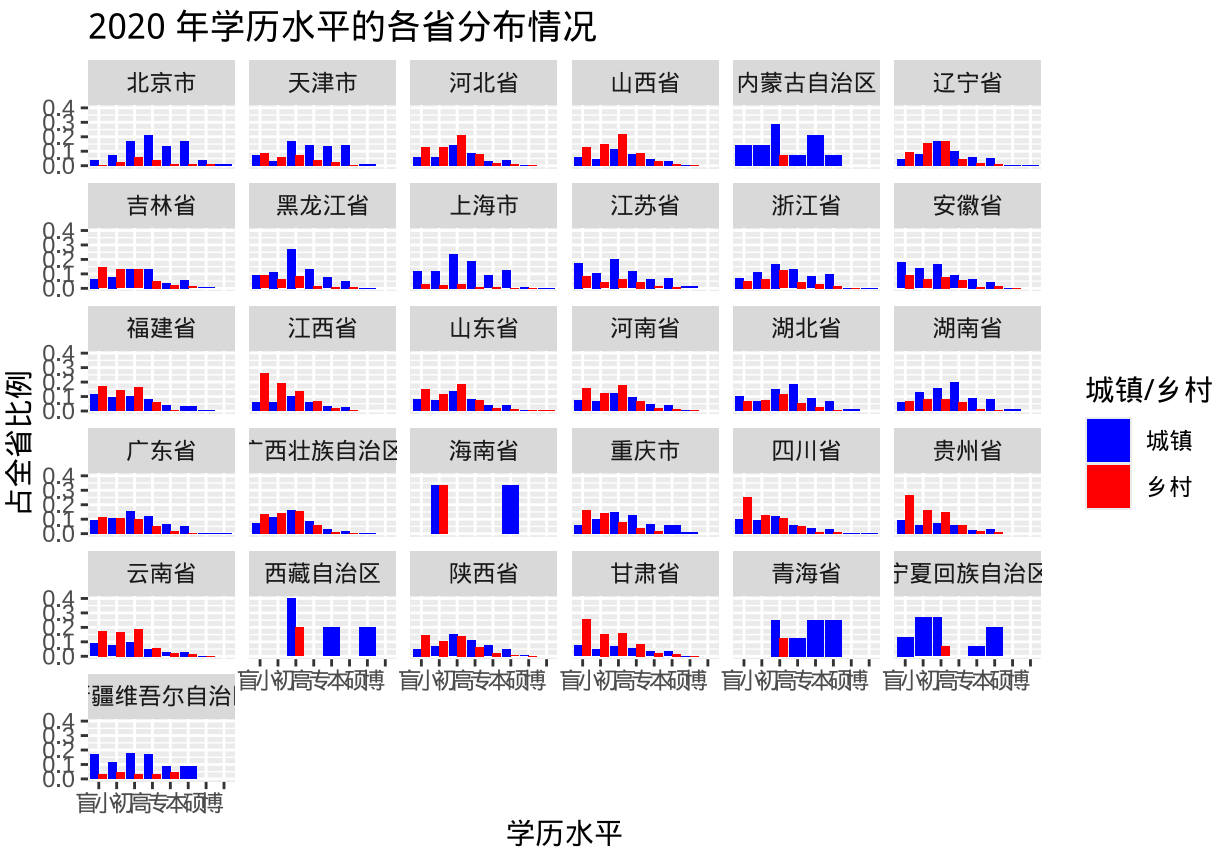

图 1：2020 年成年人学历水平的各省分布情况

从图中可以看出，2020 年成年人学历水平的各省分布呈现如下特征：

(1) 大多数省份的城镇人口学历水平以初中学历最多，向更高学历和更低学历均递减.

(2) 硕博，尤其是博士学历，在总人口中所占比重极低，同时在本硕博占比上，城镇均显著高于乡村，体现出高级人才集中在城镇的特性.

(3) 对同一个省而言，乡村人口的学历水平比城镇人口的学历水平低，具体体现为有若干省的乡村呈现文盲水平分布最多的特点.

(4) 省或直辖市之间的学历水平分布不一致. 以北京市与贵州省为例，可以直观地看到人才向一线城市聚集的趋势.

下图中，我们分析了全国范围内学历水平分布在 10 年之内的变化：

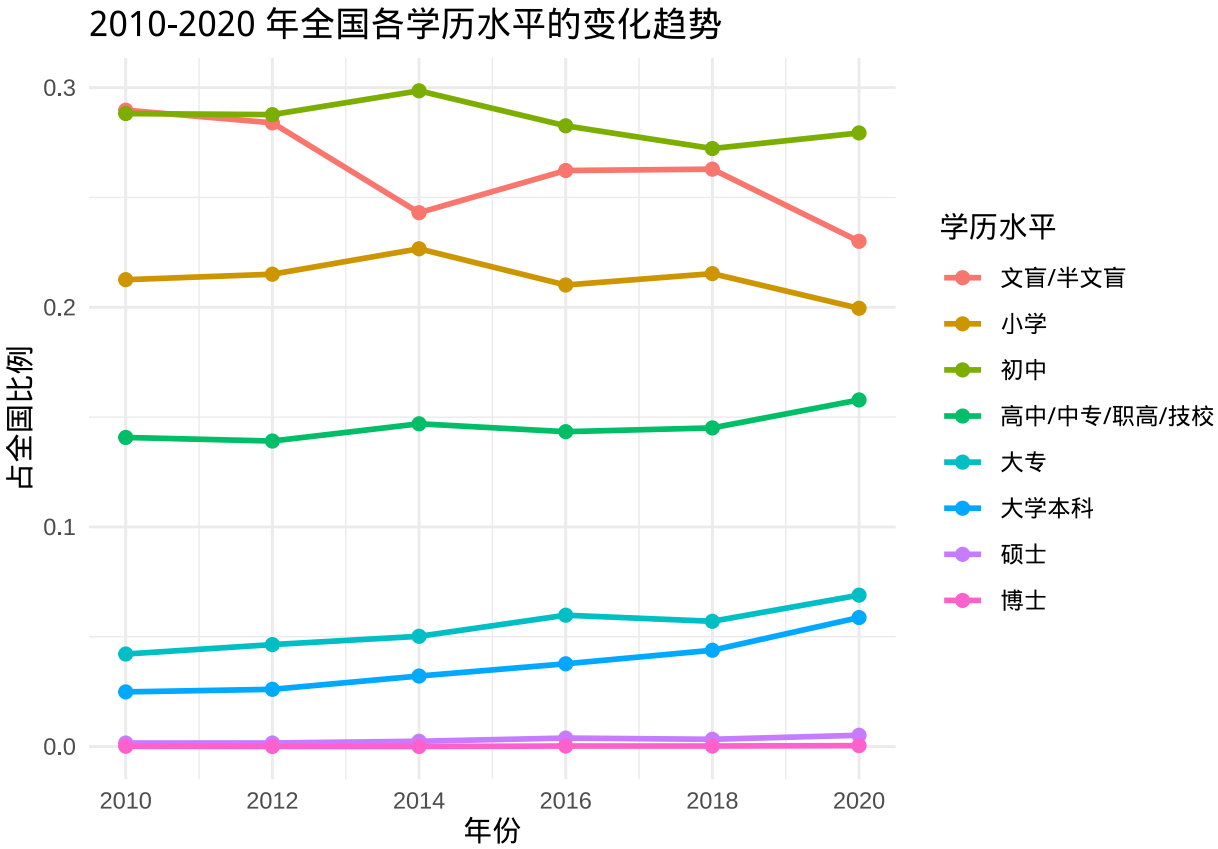

图 2：2010-2020 年成年人学历水平的全国范围变化特征

从图中可以看出，2010-2020 年成年人学历水平的全国分布呈现以下特征：

(1) 文盲/半文盲率大幅下降，小学文化水平所占比重几乎不变.

(2) 硕博学历所占人口比重仍然极低.

(3) 大学本科与大学专科学历大幅上升，印证了大学扩招的影响.

(4) 从全国角度看，初中学历仍比例最高.

3 学历水平的影响因素

接下来，我们以 2020 年数据为基础，来分析学历水平的影响因素.

3.1 因素 1：地域

地域关乎经济发展水平，可以推算出地域应当影响当地的学历水平分布. 我们将通过具体的数据分析来验证这一观点.

首先，为了便于我们的分析，我们依据需要读书的时间为学历水平定义具体的分值（即文化水平）. 例如高中毕业需要 12 年，本科毕业需要 16 年. 具体如下：

学历	分值	学历	分值
文盲/半文盲/幼儿园	0	大专	15
小学	6	大学本科	16
初中	9	硕士	19
高中/中专/职高/技校	12	博士	23

3.1.1 省份对学历水平的影响

我们将通过对若干省或直辖市的数据进行假设检验，来检验省份是否对学历分布带来显著差异. 具体地，我们随机抽取 150 个样本来进行比较，并取显著性水平 $\alpha = 0.05$.

显著性检验的结果如下：

检验双方	P 值	差异是否显著
北京、贵州	$< 2.2 \times 10^{-16}$	是
上海、天津	0.3105	否
上海、吉林	0.0223	是
安徽、河南	0.4047	否
安徽、福建	0.5327	否
安徽、贵州	0.0009812	是

上表中，我们比较了经济发展水平相近的与差异较大的省份或直辖市，假设检验说明省份的经济发展水平对人口学历水平具有显著影响，进一步验证了之前所提出的“一线城市对人才产生虹吸效应”的观点（也可能受教育观念、高考录取率等影响）.

3.1.2 城镇与乡村的学历水平比较

在之前的分析中，我们通过观察图表，提出了“普遍来说，城镇人口比乡村人口具有更高的学历水平”的论点，接下来我们将通过显著性检验来证明这一点.

我们对城镇与乡村的总数据进行随机抽样，抽取 5000 个样本，并取显著性水平 $\alpha = 0.05$ 来验证城镇化是否对学历带来显著差异.

经过计算，得到 $P < 2.2 \times 10^{-16} < 0.05$ ，因此城镇化对学历具有显著影响.

3.2 因素 2：性别

接下来我们将研究性别是否对学历水平产生影响. 与上面一致，我们从男女中分别抽取 5000 个样本，并取显著性水平 $\alpha = 0.05$ 来验证性别是否对学历带来显著差异.

经过计算，得到 $P < 2.2 \times 10^{-16} < 0.05$ ，并且男性平均学历高于女性平均学历，因此性别对学历具有显著影响. 不过值得注意的是，如下图所示，在 2010-2020 年间，女性学历水平也得到长足进步，体现了性别平等的进一步实现.

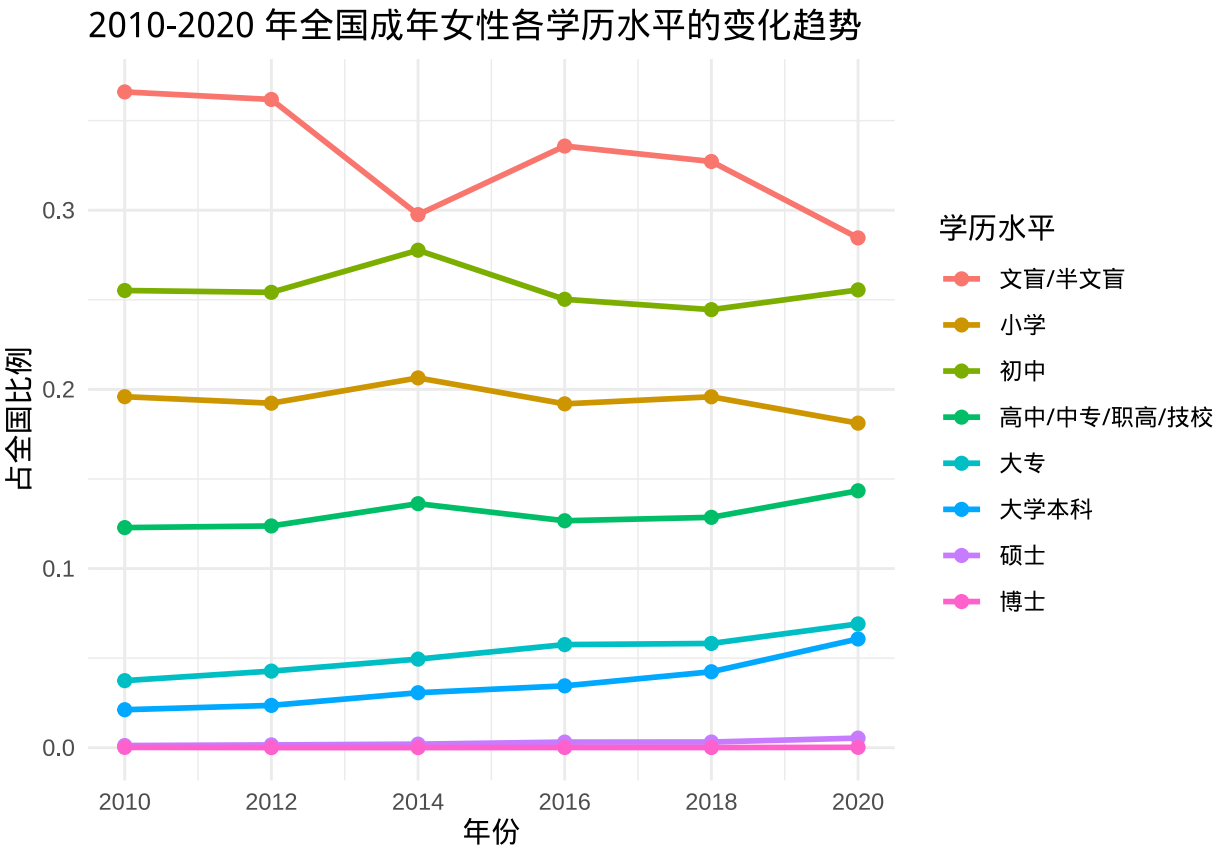

图 3：2010-2020 年女性成年人学历水平的全国范围变化特征

4　学历水平与其他变量的联系

学历作为一个人的重要标签，往往可以指示一些个人能力或者心智水平. 接下来，我们将以薪资水平与智力水平为例，讨论学历水平这一变量与它们的联系.

4.1　同一学历水平的薪资分布

人们常说：“书中自有黄金屋”. 为了证明这一论断，接下来我们使用 2020 年 CFPS 数据来探讨学历水平与薪资水平的联系. 如下图所示：

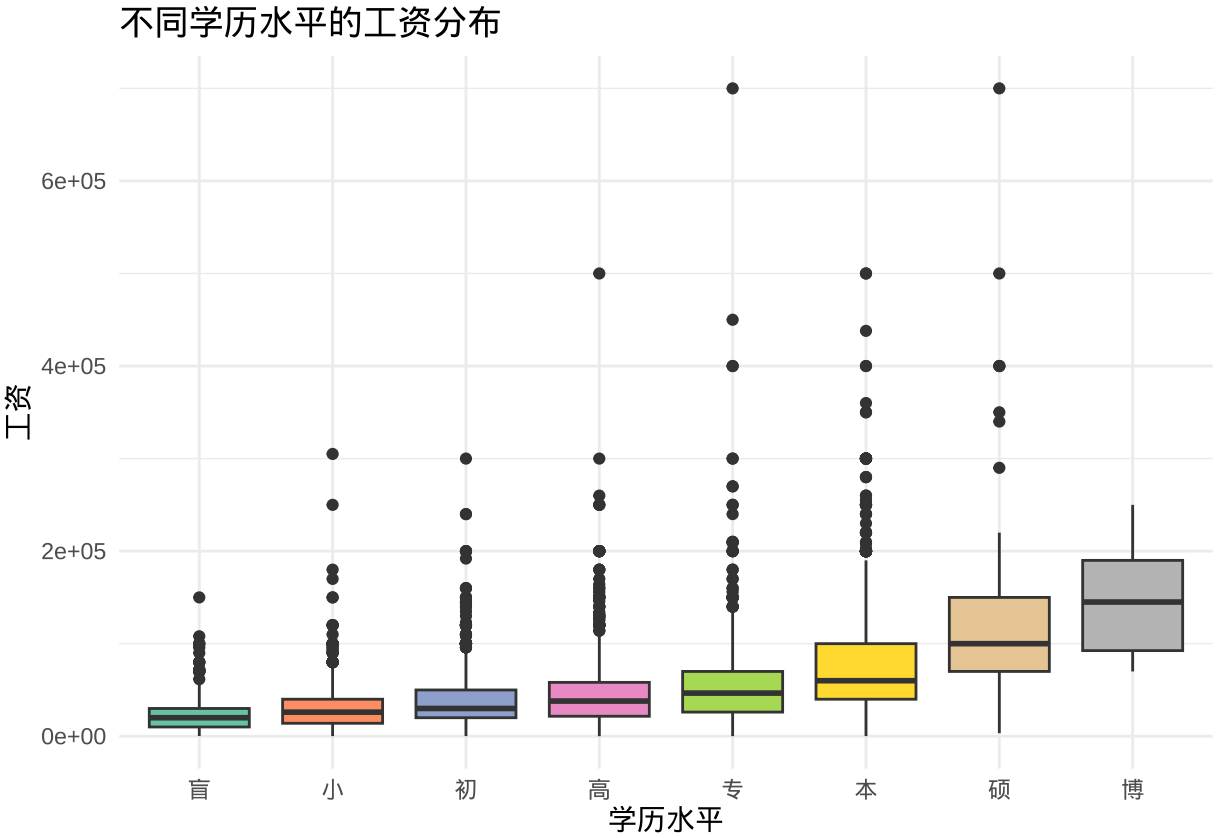

图 4：不同学历水平的薪资水平

从图中，我们可以发现如下特征：

(1) 从统计意义来看，确实存在学历更高时薪资范围与均值会更高的现象.

(2) 即使学历较低，也有获得极高薪资的机会，验证了学历不是影响薪资的唯一因素.

4.2 同一学历水平的智力水平

CFPS 统计数据中包含了智力数据（从 0-7），下图中我们分析了学历水平与智力水平的联系.

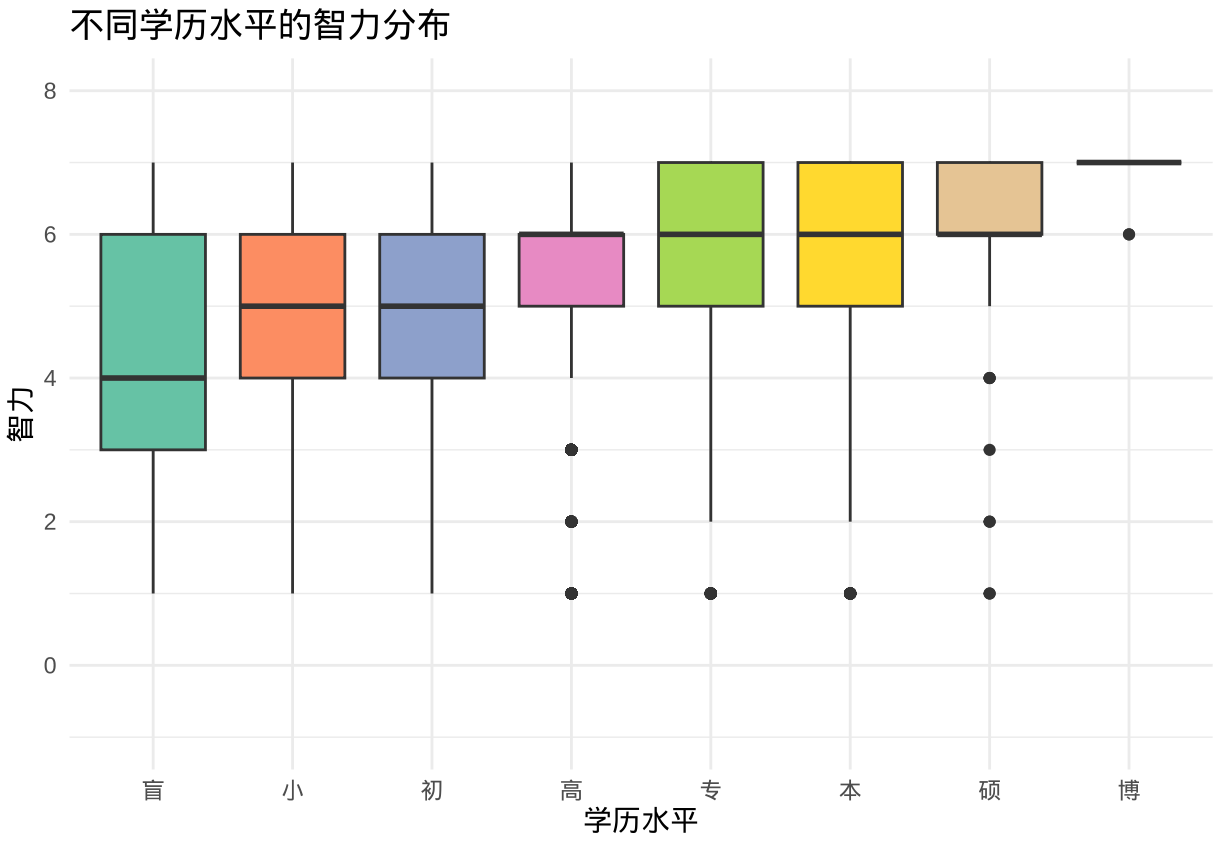

图 5：不同学历水平的智力分布

从图中可以看出，一般来说，学历越高，在智力测试中的得分就会越高. 这可能是由于教育给人带来的思维能力的提高，也有可能揭示了智力与能达到的最高学历的关系.

5 总结与致谢

总的来说，通过这次基于数十万条数据的分析，我们可以初步得到如下的结论：

(1) 根据最新 CFPS 数据，当下中国社会中，初中学历占比最多，同时文盲率不断降低，肯定了基础教育的成果.

(2) 硕博学历占比极低，仍具有一定的含金量. 本科与专科学历受大学扩招影响，占比上升较明显.

(3) 在本硕博占比上，城镇均显著高于乡村，体现出高级人才向城镇集中的特性，这要求乡村出台吸引人才政策，吸引更多高学历人才加入乡村振兴建设.

(4) 普遍而言，乡村人口学历水平低于城镇人口学历水平，乡村教育振兴刻不容缓.

(5) 一线城市具有人才虹吸效应.

(6) 男性平均学历高于女性，提示我们需要进一步完善性别平等.

(7) 从统计意义上，学历更高薪资更高，但是仍然存在许多其他的可能性.

在论文的最后，首先，我想向吴老师表示诚挚的感谢. 作为一名无统计学基础的计算机大一新生，在她的课堂上我学到了许多有关 R 语言、数据分析与统计学的知识. 其次，我想向这门课程的助教闫同学表示诚挚的感谢，她付出了许多时间与精力来批阅同学们的作业. 最后，我想向 CFPS 数据团队表示诚挚的感谢，他们收集了数十万条内容详尽、结构清晰的数据，也为笔者的分析提供了宝贵的原材料.

竹杖芒鞋轻胜马，谁怕？一蓑烟雨任平生。

参考文献

[1] Peking University Institute of Social Science Survey. China Family Panel Studies (CFPS), 2015.

[2] 徐海东 and 周皓. 我国人口受教育状况的发展与启示——基于 1982—2020 年全国人口普查公报数据的思考. 中共福建省委党校（福建行政学院）学报, (5):126–137, 9 2022.

A 全部代码展示

```
# 修复中文不显示的问题

library(showtext)
showtext_auto(enable = TRUE)
font_add('Songti', 'Songti.ttc')
quartz(family='Songti')

# 从 CFPS 数据集中读取 STATA 格式的数据

library(haven)

data2010 <- read_dta("~/Code/data/[CFPS Public Data] CFPS 2010 in Stata (Chinese)/cfps2010adult_202008.dta")
data2012 <- read_dta("~/Code/data/[CFPS Public Data] CFPS2012 in STATA (Chinese)/cfps2012adult_201906.dta")
data2014 <- read_dta("~/Code/data/[CFPS Public Data] CFPS2014 in STATA (Chinese)/cfps2014adult_201906.dta")
data2016 <- read_dta("~/Code/data/[CFPS Public Data] CFPS2016 in STATA (Chinese)/cfps2016adult_201906.dta")
data2018 <- read_dta("~/Code/data/CFPS2018/cfps2018person_202012.dta")
data2020 <- read_dta("~/Code/data/[CFPS+Public+Data]+CFPS+2020_in_STATA_(Chinese)/cfps2020person_202306.dta")

############## 第一阶段分析 ##############

# 第一阶段分析 1: 只保留学历、省份与城乡分类数据

library(dplyr)

data2010_1 <- data2010 %>%
  select(provcd=provcd, urban=urban, edu=cfps2010edu_best)
data2012_1 <- data2012 %>%
  select(provcd=provcd, urban=urban12, edu=edu2012)
data2014_1 <- data2014 %>%
  select(provcd=provcd14, urban=urban14, edu=cfps2014edu)
data2016_1 <- data2016 %>%
  select(provcd=provcd16, urban=urban16, edu=cfps2016edu)
data2018_1 <- data2018 %>%
  select(provcd=provcd18, urban=urban18, edu=cfps2018edu)
data2020_1 <- data2020 %>%
  select(provcd=provcd20, urban=urban20, edu=cfps2020edu)

# 只保留有效数据，即学历从 1-8 等

clean1 <- function(df) { return(df[!is.na(df$edu) & df$edu >= 1 & df$edu <= 8, ]) }
clean2 <- function(df) { return(df[!is.na(df$urban) & df$urban >= 0 & df$urban <= 1, ])}
clean3 <- function(df) { return(df[!is.na(df$provcd) & df$provcd >= 11 & df$provcd <= 65, ])}
clean <- function(df) { return(clean1(clean2(clean3(df)))) }
data2010_1 <- clean(data2010_1)
data2012_1 <- clean(data2012_1)
data2014_1 <- clean(data2014_1)
data2016_1 <- clean(data2016_1)
data2018_1 <- clean(data2018_1)
```

```r
data2020_1 <- clean(data2020_1)

# 绘制条形图，展示 2020 年的分省情况

data2020_1_2 <- data2020_1 %>%
  mutate(urban=factor(urban, levels=c(1, 0), labels=c(" 城镇", " 乡村")),
         edu=factor(edu, levels=1:8, labels=c(" 盲", " 小", " 初",
                                              " 高", " 专", " 本",
                                              " 硕", " 博")),
         provcd=factor(provcd, levels=c(11, 12, 13, 14, 15, 21, 22, 23, 31, 32, 33, 34,
                                        35, 36, 37, 41, 42, 43, 44, 45, 46, 50, 51, 52,
                                        53, 54, 61, 62, 63, 64, 65),
                       labels=c(" 北京市", " 天津市", " 河北省", " 山西省",
                                " 内蒙古自治区", " 辽宁省", " 吉林省", " 黑龙江省",
                                " 上海市", " 江苏省", " 浙江省", " 安徽省",
                                " 福建省", " 江西省", " 山东省", " 河南省",
                                " 湖北省", " 湖南省", " 广东省", " 广西壮族自治区",
                                " 海南省", " 重庆市", " 四川省", " 贵州省",
                                " 云南省", " 西藏自治区", " 陕西省", " 甘肃省",
                                " 青海省", " 宁夏回族自治区", " 新疆维吾尔自治区"))) %>%
  group_by(provcd, edu, urban) %>%
  summarise(count=n(), .groups='drop') %>%
  group_by(provcd) %>%
  mutate(total=sum(count), proportion=count / total) %>%
  ungroup()

library(tidyr)
library(ggplot2)

plot1 <- ggplot(data2020_1_2, aes(x=edu, y=proportion, fill=urban)) +
  geom_bar(stat="identity", position="dodge") +
  facet_wrap(~provcd, ncol=6) +
  scale_fill_manual(values=c(" 城镇"="blue", " 乡村"="red")) +
  labs(title="2020 年学历水平的各省分布情况",
       x=" 学历水平", y=" 占全省比例", fill = " 城镇/乡村") +
  theme(axis.text.x=element_text(angle=0, hjust=1))

plot1

# 第一阶段分析 2: 全国范围的变化

data2010_1 <- data2010_1 %>% mutate(year=2010)
data2012_1 <- data2012_1 %>% mutate(year=2012)
data2014_1 <- data2014_1 %>% mutate(year=2014)
data2016_1 <- data2016_1 %>% mutate(year=2016)
data2018_1 <- data2018_1 %>% mutate(year=2018)
data2020_1 <- data2020_1 %>% mutate(year=2020)
all_data <- bind_rows(data2010_1, data2012_1, data2014_1, data2016_1, data2018_1, data2020_1)

national_trends <- all_data %>%
```

```
  mutate(edu=factor(edu, levels=1:8, labels=c(" 文盲/半文盲", " 小学", " 初中",
         " 高中/中专/职高/技校", " 大专", " 大学本科", " 硕士", " 博士"))) %>%
  group_by(year, edu) %>%
  summarise(count=n(), .groups='drop') %>%
  group_by(year) %>%
  mutate(total=sum(count), proportion=count / total) %>%
  ungroup()

plot2 <- ggplot(national_trends, aes(x=year, y=proportion, color=edu)) +
  geom_line(linewidth=1) +
  geom_point(size=2) +
  scale_x_continuous(breaks=seq(2010, 2020, 2)) +
  labs(title="2010-2020 年全国各学历水平的变化趋势",
       x=" 年份", y=" 占全国比例", color=" 学历水平") +
  theme_minimal()

plot2

############## 第二阶段分析 ##############

# 第二阶段分析 1: 分析省份对学历水平的影响

data2020_1_3 <- data2020_1 %>%
  mutate(urban=factor(urban, levels=c(1, 0), labels=c(" 城镇", " 乡村")),
         edu=ifelse(edu == 1, 0,
             ifelse(edu == 2, 6,
             ifelse(edu == 3, 9,
             ifelse(edu == 4, 12,
             ifelse(edu == 5, 15,
             ifelse(edu == 6, 16,
             ifelse(edu == 7, 19,
             ifelse(edu == 8, 23, NA)))))))),
         provcd=factor(provcd, levels=c(11, 12, 13, 14, 15, 21, 22, 23, 31, 32, 33, 34,
                                        35, 36, 37, 41, 42, 43, 44, 45, 46, 50, 51, 52,
                                        53, 54, 61, 62, 63, 64, 65),
                       labels=c(" 北京市", " 天津市", " 河北省", " 山西省",
                                " 内蒙古自治区", " 辽宁省", " 吉林省", " 黑龙江省",
                                " 上海市", " 江苏省", " 浙江省", " 安徽省",
                                " 福建省", " 江西省", " 山东省", " 河南省",
                                " 湖北省", " 湖南省", " 广东省", " 广西壮族自治区",
                                " 海南省", " 重庆市", " 四川省", " 贵州省",
                                " 云南省", " 西藏自治区", " 陕西省", " 甘肃省",
                                " 青海省", " 宁夏回族自治区", " 新疆维吾尔自治区")))

library(BSDA)

# 进行假设检验的函数
set.seed(202464)

test <- function(city1, city2) {
```

```
  df <- data2020_1_3 %>%
    filter(provcd %in% c(city1, city2))
  x <- (df %>%
    filter(provcd == city1) %>%
    sample_n(150, replace=TRUE))$edu
  y <- (df %>%
    filter(provcd == city2) %>%
    sample_n(150, replace=TRUE))$edu
  z.test(x, y, sigma.x=sd(x), sigma.y=sd(y))
}

test(" 北京市", " 贵州省")
test(" 上海市", " 天津市")
test(" 上海市", " 吉林省")
test(" 安徽省", " 河南省")
test(" 安徽省", " 福建省")
test(" 安徽省", " 贵州省")

# 第二阶段分析 2: 分析城镇化水平对学历水平的影响

urban_data <- (data2020_1_3 %>%
  filter(urban == " 城镇") %>%
  sample_n(5000, replace=TRUE))$edu
rural_data <- (data2020_1_3 %>%
  filter(urban == " 乡村") %>%
  sample_n(5000, replace=TRUE))$edu
z.test(urban_data, rural_data, sigma.x=sd(urban_data), sigma.y=sd(rural_data))

# 第二阶段分析 3: 分析性别对学历水平的影响

data2020_gender <- data2020 %>%
  select(gender=gender, edu=cfps2020edu)
data2020_gender <- clean1(data2020_gender)
data2020_gender_1 <- data2020_gender %>%
  mutate(edu=ifelse(edu == 1, 0,
              ifelse(edu == 2, 6,
              ifelse(edu == 3, 9,
              ifelse(edu == 4, 12,
              ifelse(edu == 5, 15,
              ifelse(edu == 6, 16,
              ifelse(edu == 7, 19,
              ifelse(edu == 8, 23, NA)))))))))
man <- (data2020_gender_1 %>%
  filter(gender == 1) %>%
  sample_n(5000, replace=TRUE))$edu
woman <- (data2020_gender_1 %>%
  filter(gender == 0) %>%
  sample_n(5000, replace=TRUE))$edu
z.test(man, woman, sigma.x=sd(man), sigma.y=sd(woman))
```

```r
data2020_gender <- clean1(data2020 %>%
  select(gender=gender, edu=cfps2020edu)) %>% mutate(year=2020)
data2018_gender <- clean1(data2018 %>%
  select(gender=gender, edu=cfps2018edu)) %>% mutate(year=2018)
data2016_gender <- clean1(data2016 %>%
  select(gender=cfps_gender, edu=cfps2016edu)) %>% mutate(year=2016)
data2014_gender <- clean1(data2014 %>%
  select(gender=cfps_gender, edu=cfps2014edu)) %>% mutate(year=2014)
data2012_gender <- clean1(data2012 %>%
  select(gender=cfps2012_gender_best, edu=edu2012)) %>% mutate(year=2012)
data2010_gender <- clean1(data2010 %>%
  select(gender=gender, edu=cfps2010edu_best)) %>% mutate(year=2010)

all_data <- bind_rows(data2010_gender, data2012_gender, data2014_gender,
                      data2016_gender, data2018_gender, data2020_gender) %>%
  filter(gender == 0)

trends <- all_data %>%
  mutate(edu=factor(edu, levels=1:8, labels=c(" 文盲/半文盲", " 小学", " 初中",
         " 高中/中专/职高/技校", " 大专", " 大学本科", " 硕士", " 博士"))) %>%
  group_by(year, edu) %>%
  summarise(count=n(), .groups='drop') %>%
  group_by(year) %>%
  mutate(total=sum(count), proportion=count / total) %>%
  ungroup()

plot3 <- ggplot(trends, aes(x=year, y=proportion, color=edu)) +
  geom_line(linewidth=1) +
  geom_point(size=2) +
  scale_x_continuous(breaks=seq(2010, 2020, 2)) +
  labs(title="2010-2020 年全国成年女性各学历水平的变化趋势",
       x=" 年份", y=" 占全国比例", color=" 学历水平") +
  theme_minimal()

plot3

############## 第三阶段分析 ##############

# 建立薪资水平与学历水平的一元线性回归模型

data2020_2 <- data2020 %>%
  select(wisdom=qz207, wage=qg12, edu=cfps2020edu) %>%
  filter(wage > 0 & !is.na(wage) & edu >= 1 & edu <= 8 & !is.na(edu))

data2020_2_1 <- data2020_2 %>%
  mutate(edu=factor(edu, levels=1:8, labels=c(" 盲", " 小", " 初",
                                              " 高", " 专", " 本",
                                              " 硕", " 博")))

data2020_2_2 <- data2020_2 %>%
```

```
  mutate(edu=factor(edu, levels=1:8, labels=c(" 盲", " 小", " 初",
                                              " 高", " 专", " 本",
                                              " 硕", " 博")),
         wisdom=as.numeric(wisdom))

palette <- RColorBrewer::brewer.pal(8, "Set2")
ggplot(data=data2020_2_1, aes(x=edu, y=wage)) +
  geom_boxplot(fill=palette) +
  theme_minimal() +
  labs(x=" 学历水平", y=" 工资", title=" 不同学历水平的工资分布") +
  theme(axis.test.x=element_text(angle=0, hjust=1))

ggplot(data=data2020_2_2, aes(x=edu, y=wisdom)) +
  geom_boxplot(fill=palette) +
  theme_minimal() +
  labs(x=" 学历水平", y=" 智力", title=" 不同学历水平的智力分布") +
  theme(axis.test.x=element_text(angle=0, hjust=1)) +
  scale_y_continuous(limits = c(-1, 8))
```