Survey Project: Analysis of the Current Population Survey (CPS)

Course: DATA 370 – Survey Sampling and Analysis

Instructor: Liang Wu

Group Members:

- Junbo Zhang
- Lei Wei
- Minghao Lu
- Weizhan Gao

Date: 2025/12/13

1. Survey Documentation & Design Analysis

1.1 Survey Description and Data Source

This project uses data from the Current Population Survey (CPS), a nationally representative survey conducted jointly by the U.S. Census Bureau and the Bureau of Labor Statistics. The CPS is designed to collect information on labor force participation, employment, and demographic characteristics of the civilian noninstitutional population in the United States.

The dataset analyzed in this project comes from the January 2025 CPS Public Use Microdata. The data were obtained from an official public-use release, which provides anonymized individual- and household-level records suitable for academic and classroom analysis. Variable definitions, coding schemes, and record layouts are documented in the official CPS public-use record layout documentation.

1.2 Sampling Design: Stratification and Clustering

The CPS employs a complex, multistage sampling design rather than simple random sampling. The sampling design incorporates both stratification and clustering to improve efficiency and ensure adequate representation across geographic and demographic groups.

Stratification in the CPS is primarily based on geographic characteristics, such as states, metropolitan status, and other regional groupings. Within each stratum, sampling units are selected to ensure broad national coverage.

Clustering is implemented through Primary Sampling Units (PSUs), which are typically defined as geographic areas such as counties or groups of counties. Households are sampled within PSUs, and individuals are then selected within households. As a result, observations within the same PSU or household may be correlated.

1.3 Sampling Stages and Survey Weights

The CPS follows a multistage sampling process. In the first stage, PSUs are selected within strata. In the second stage, households are sampled within selected PSUs. In the final stage, individuals within households are surveyed.

To account for unequal probabilities of selection, nonresponse, and post-stratification adjustments, the CPS provides sampling weights. These weights allow analysts to produce estimates that are representative of the U.S. population. All population-level estimates in this project are therefore intended to be computed using survey weights rather than unweighted sample means.

Because the CPS uses a complex survey design, standard errors and variance estimates must account for stratification, clustering, and weighting. Treating the data as a simple random sample would lead to incorrect inference.

1.4 Confidentiality and Public Use Considerations

The CPS data used in this project are public-use microdata and do not contain personally identifiable information such as names or exact addresses. Geographic identifiers in the public-use files are aggregated or masked to prevent the identification of individual respondents.

These confidentiality protections ensure that the dataset can be used for academic analysis without violating respondent privacy. As a result, the CPS public-use data are appropriate for classroom research and statistical analysis.

2. Preliminary Data Exploration

This section provides an initial descriptive exploration of the CPS January 2025 data.

The purpose of this preliminary analysis is to understand the basic structure and distribution of selected variables before conducting more advanced survey-weighted analyses.

2.1 Variable Selection

Two categorical variables and two continuous variables are selected for descriptive exploration.

The categorical variables are sex (pesex) and employment status (pemlr). Sex is included as a basic demographic characteristic of the sample, while employment status reflects respondents' labor force participation and is a core variable in the CPS.

The continuous variables are age (prtage) and usual weekly hours worked (pehrusl1). Age provides information on the demographic composition of the sample, and usual weekly hours worked captures typical labor supply among employed individuals.

All variable definitions are based on the official CPS public-use documentation.

2.2 Distribution of Age

Figure 1 presents a histogram of respondents' age. Observations with missing values and non-positive ages are excluded from the visualization.

The age distribution spans a wide range, from young adults to older individuals, with a median age of approximately 42 years. The distribution is relatively flat across age groups, reflecting the broad population coverage of the CPS rather than a normally distributed variable. The first and third quartiles are approximately 22 and 62 years, respectively, indicating substantial variation in age across the sample.

Overall, the age distribution appears reasonable and consistent with expectations for a nationally representative household survey.

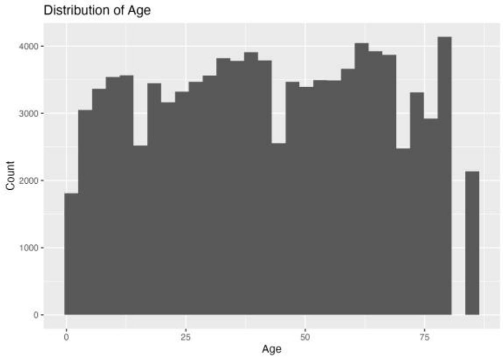

Figure 1. Distribution of Age

2.3 Distribution of Usual Weekly Hours Worked

Figure 2 shows the distribution of usual weekly hours worked among respondents for whom this variable is defined. Only positive values are included, as usual weekly hours worked apply to employed individuals.

The distribution exhibits a pronounced concentration at 40 hours per week, which reflects the standard full-time work schedule reported by many respondents. This is further supported by the summary statistics, where the median and both the 25th and 75th percentiles are equal to 40 hours. The mean usual weekly hours worked is approximately 38 hours, indicating that while full-time work is common, a substantial share of respondents report fewer hours.

The observed pattern is typical of labor force survey data and suggests a mixture of full-time and part-time employment in the sample.

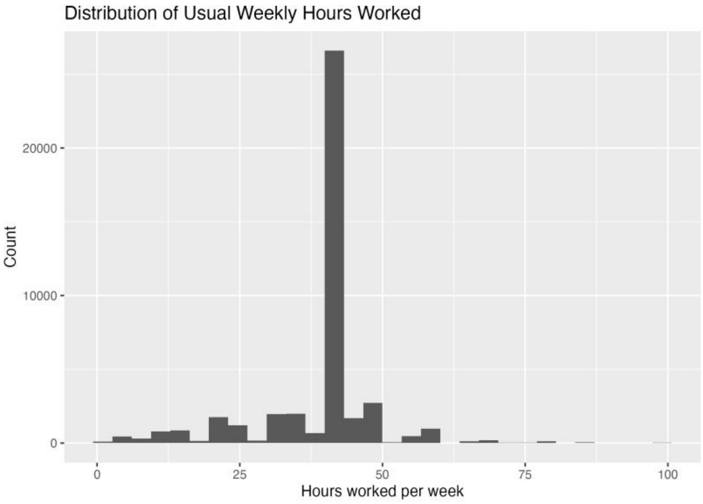

Figure 2. Distribution of Usual Weekly Hours Worked

2.4 Summary Statistics

Table 1 summarizes the mean and quartiles for the two continuous variables. The mean age of respondents is approximately 42 years, with substantial dispersion across the population. Usual weekly hours worked are tightly centered around 40 hours, consistent with the histogram and the standard full-time work norm.

These descriptive statistics provide a useful overview of the data and highlight key features that will be important for subsequent survey-weighted analyses.

Variable	Mean	25th percentile	Median	75th percentile
Age	42.3	22	42	62
Usual weekly hours worked	38.2	40	40	40

Table 1. Summary Statistics for Continuous Variables (Mean, 25th, 50th, and 75th Percentiles)

3. Nonresponse Analysis

This section examines patterns of nonresponse in the CPS January 2025 data, focusing on responses to the usual weekly hours worked variable. Nonresponse is assessed to evaluate whether missing information may introduce potential bias into subsequent analyses.

3.1 Definition of Nonresponse

Nonresponse is defined based on the variable usual weekly hours worked (pehrusl1). Respondents are defined as individuals with positive and non-missing values of pehrusl1. Individuals with zero or missing values of this variable are classified as nonrespondents. This definition reflects the fact that usual weekly hours worked are only applicable to employed individuals.

Using this definition, approximately 43,500 individuals are classified as respondents, while about 82,000 individuals are classified as nonrespondents.

3.2 Age Differences Between Respondents and Nonrespondents

Age distributions are compared between respondents and nonrespondents to assess whether nonresponse is associated with demographic characteristics. After excluding missing or invalid age values, respondents have a slightly higher average age than nonrespondents. The mean age of respondents is approximately 43 years, compared to about 41 years for nonrespondents. Median ages follow a similar pattern.

Although the difference in age is modest, it suggests that nonresponse to the usual weekly hours worked question is not entirely random with respect to age. Younger individuals and those outside the core working-age population are somewhat more likely to be classified as nonrespondents.

3.3 Sex Differences and Missing Information

Sex distributions also differ between respondents and nonrespondents. Among respondents, the distribution of sex is relatively balanced, with males and females accounting for approximately 52 percent and 48 percent of respondents, respectively.

In contrast, a substantial proportion of nonrespondents have missing values for sex. This pattern suggests that nonresponse to the usual weekly hours worked variable is often associated with incomplete interviews or household-level nonresponse, rather than item-level nonresponse alone. Among nonrespondents with observed sex, females slightly outnumber males.

3.4 Implications for Analysis

The observed differences between respondents and nonrespondents indicate that nonresponse in the CPS is not completely random. Respondents tend to be slightly older and have more complete demographic information. As a result, analyses that rely solely on respondents to the usual weekly hours worked question may overrepresent older individuals and underrepresent younger or non-employed populations.

These findings highlight the importance of accounting for nonresponse and survey design features in subsequent analyses, particularly when estimating population-level labor market outcomes.

4. Variance Estimation Methods

4.1 Survey Variance Estimation Approach

The Current Population Survey (CPS) employs a complex, multistage sampling design that incorporates stratification, clustering, and unequal probabilities of selection. As a result, standard variance formulas based on simple random sampling (SRS) assumptions are not appropriate for estimating standard errors.

To account for the complex survey design, variance estimation for CPS estimates is conducted using Taylor series linearization. This method approximates the variance of nonlinear estimators by linearizing them around their expected values and then applying design-based variance formulas that incorporate survey weights, strata, and primary sampling units (PSUs).

Taylor linearization is the default variance estimation method used by the CPS and is implemented in standard survey analysis software, including the survey package in R. This approach allows for valid standard error estimation under the CPS sampling design.

4.2 Design Effects for Key Estimates

The design effect (DEFF) is used to quantify the impact of the complex survey design on variance estimation. It is defined as the ratio of the variance of an estimator under the actual survey design to the variance that would be obtained under simple random sampling with the same sample size.

Formally, the design effect can be expressed as:

DEFF = Var_complex / Var_SRS

or equivalently,

DEFF = $(\text{SE\_complex} / \text{SE\_SRS})^2$

Design effects greater than one indicate that the complex survey design increases variance relative to SRS, often due to clustering or unequal weighting. Design effects less than one may occur when stratification leads to increased efficiency.

In this analysis, design effects are calculated for key estimates, including the mean age and the mean usual weekly hours worked.

4.3 Comparison of Complex Design and SRS Standard Errors

To evaluate the importance of accounting for the CPS survey design, standard errors obtained under the complex design are compared with those calculated under a simple random sampling assumption.

In general, standard errors under the complex design are expected to differ from SRS-based standard errors. Clustering within PSUs tends to increase variance because observations within clusters are correlated, while stratification may reduce variance by ensuring representation across population subgroups. Additionally, unequal sampling weights can further inflate variance.

Comparing complex-design and SRS standard errors highlights the potential bias that can arise if survey design features are ignored. This comparison underscores the importance of using appropriate survey methods when analyzing CPS data.

5. Software Implementation & Analysis

5.1 Software and Survey Design Specification

All survey analyses are conducted in R using the survey package. Person-level survey weights and clustering are explicitly incorporated to account for the complex sampling design of the Current Population Survey (CPS). Due to limitations of the public-use CPS data, explicit strata identifiers are not available; however, variance estimation accounts for clustering and unequal weighting using Taylor series linearization.

5.2 Construction of Survey Design Objects

Survey design objects are constructed after excluding observations with missing or zero survey weights. The primary sampling unit is approximated using the household rotation group identifier, and person-level final weights are applied.

5.3 Comparison of Complex Design and SRS Estimates

Using the complex survey design, the estimated mean age of individuals is approximately 39.7 years, with a standard error of 0.062. Under a simple random sampling (SRS) assumption, the corresponding standard error is larger, at approximately 0.086, resulting in a design effect of about 0.53. This indicates that the CPS sampling design improves efficiency for age estimation, likely due to effective stratification.

5.4 Design Effects and Interpretation

In contrast, the estimated mean usual weekly hours worked is approximately 16.9 hours, with a complex-design standard error of 0.139. The SRS-based standard error for this estimate is substantially smaller, at approximately 0.078, yielding a design effect of about 3.19. This suggests that clustering and unequal weighting substantially increase variance for hours worked, and that ignoring the survey design would lead to an underestimation of uncertainty.

Overall, these results highlight the importance of accounting for complex survey design features when conducting inference using CPS data. Design effects vary across variables, underscoring that the impact of survey design on variance estimation depends on the structure and distribution of the variable being analyzed.