NIEHS Training

Survey Analysis

With NHANES Data

Qinlu (Claire) Wang
Statistician

Bioinformatics and Computational Biosciences Branch (BCBB)
Office of Cyber Infrastructure and Computational Biology (OCICB)
National Institute of Allergy and Infectious Diseases (NIAID)

NIH National Institute of Allergy and Infectious Diseases

NIAID

Outline

1. Survey Introduction
2. Statistical Inference
3. Sampling
4. Graphics
5. Ratios and Linear Regression
6. Categorical Data Regression
7. Tests in Contingency Tables
8. Missing Data

1. Survey Introduction

The National Health and Nutrition Examination Survey (NHANES)

- Target Population
- Survey Objectives
- Data Collection Procedures

Target Population

The NHANES target population is the noninstitutionalized civilian resident population of the United States. The NHANES design has changed periodically to sample larger numbers of certain groups to increase the reliability and precision of estimates of health status indicators for these population subgroups.

The oversampled subgroups in the 2019-2020 survey cycle were:

- Hispanic persons;
- Non-Hispanic black persons;
- Non-Hispanic Asian persons;
- Non-Hispanic white and other persons at or below 185 percent of the Department of Health and Human Services (HHS) poverty guidelines; and
- Non-Hispanic white and other persons aged 80 years and older.

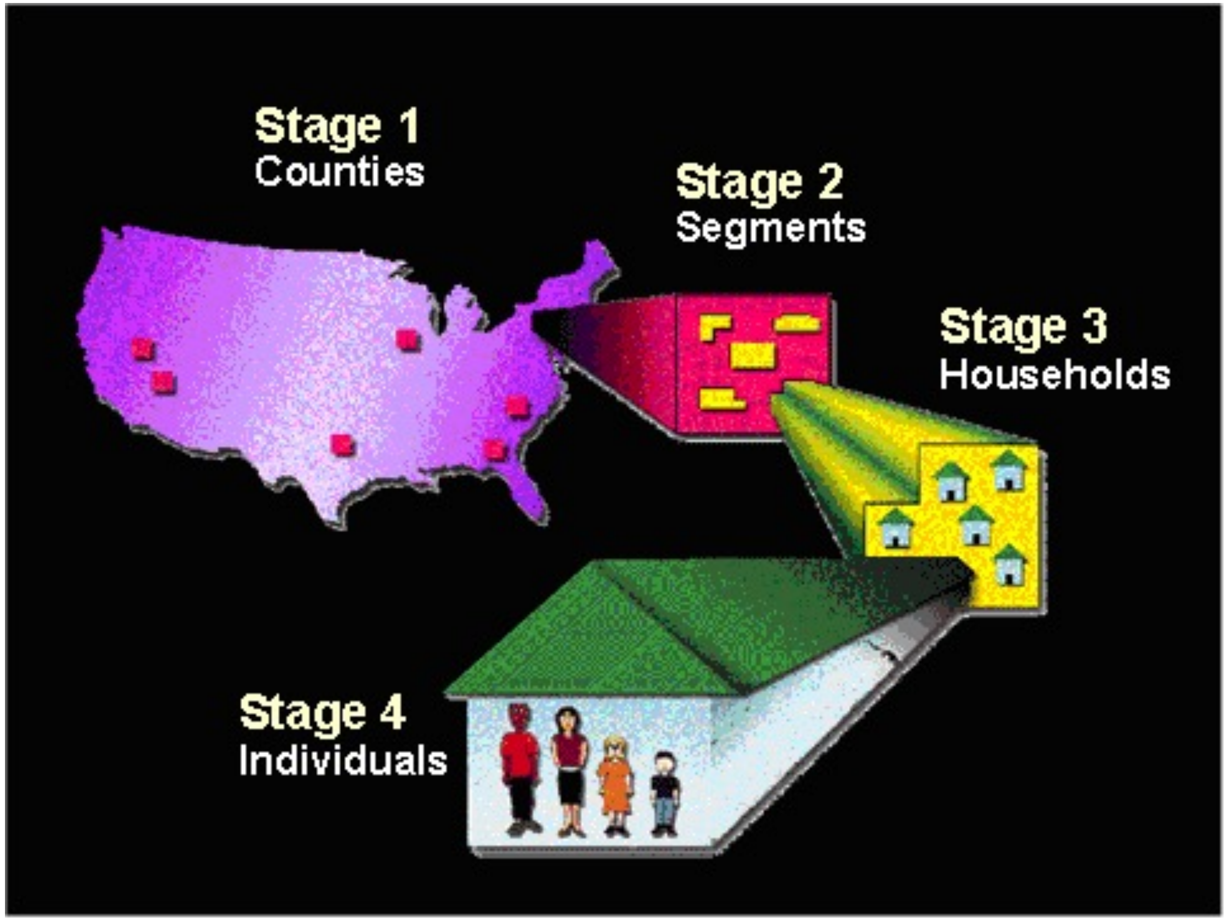
Survey Objectives

The major objectives of NHANES are to:

- Estimate the number and percentage of persons in the U.S. population and in designated subgroups with selected diseases and risk factors;
- Monitor trends in the prevalence, awareness, treatment, and control of selected diseases;
- Monitor trends in risk behaviors and environmental exposures;
- Study the relationship between diet, nutrition, and health;
- Explore emerging public health issues and new technologies;
- Provide baseline health characteristics that can be linked to mortality data from the National Death Index or other administrative records (e.g., enrollment and claims data from the Centers for Medicare & Medicaid Services); and
- Collect and maintain a national probability sample of serum, plasma, and urine samples for potential future public health emergency use and surveillance; and maintain a national probability sample of DNA samples for potential future public health emergency use and surveillance.

Data Collection Procedures

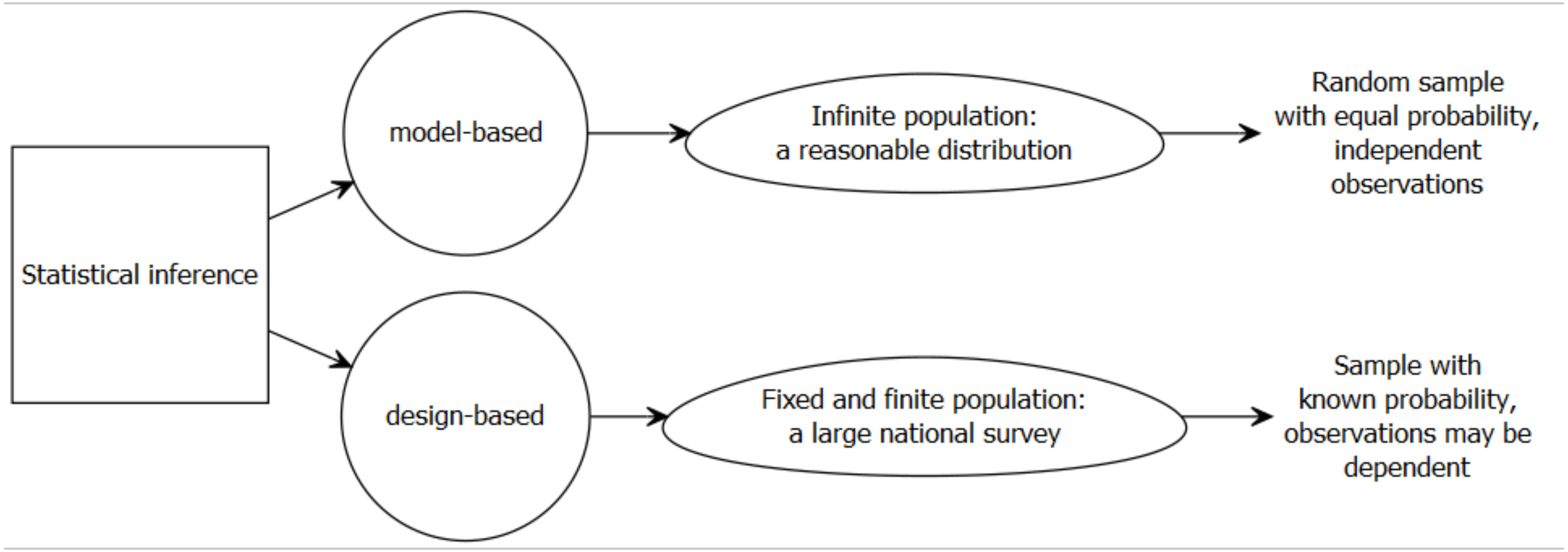
1. Selection of primary sampling units (PSUs), which are counties or small groups of contiguous counties.
2. Selection of segments within PSUs that constitute a block or group of blocks (neighborhoods) containing a cluster of households.
3. Selection of specific households within segments.
4. Selection of individuals within a household.



More references and tools for survey design:

- NHANES Survey Design: https://wwwn.cdc.gov/nchs/nhanes/tutorials/module2.aspx
- The Power of Survey Design by GIUSEPPE IAROSSI: https://www.amazon.com/Power-Survey-Design-Interpreting-Influencing/dp/0821363921
- How to write survey questions: https://www.supersurvey.com/SurveyQuestions
- Close-ended and open-ended questions: https://www.surveymonkey.com/mp/comparing-closed-ended-and-open-ended-questions/
- Sample size calculator: https://www.surveysystem.com/sscalc.htm
- Basics of designing a survey: https://www.youtube.com/watch?v=36s6wBSJW8U&t=1s

2. Statistical Inference



Statistical inference

model-based

Infinite population:
a reasonable distribution

Random sample
with equal probability,
independent
observations

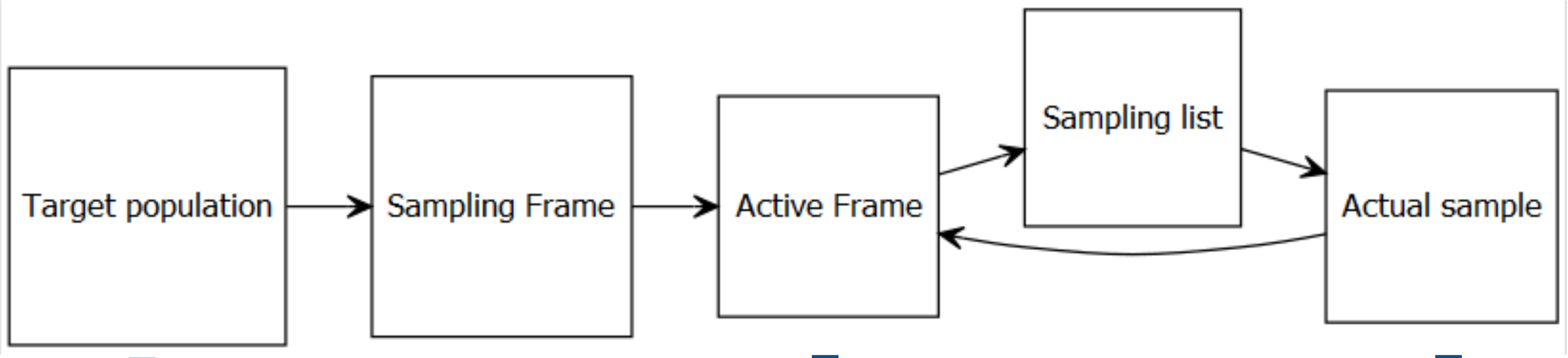
design-based

Fixed and finite population:
a large national survey

Sample with
known probability,
observations may be
dependent

3. Sampling

3.1 Steps of Generalization



Canadian population 12 years of age and over

exclude about 3% population (e.g., aboriginal settlements, Canadian Forces, institutionalized, foster care, 2 selected Quebec health regions)

E.g., not dead or have not moved

some don’t respond

3.2 Sampling Weights

For example:

- A random sample of 3500 people from California (with total population 35 million)
- Any person in California has 1/10000 chance of being sampled

$$\pi_i = \frac{3500}{3500000} = \frac{1}{10000} \; for\ every\ i$$

- Each of the people we sample represents 10000 Californians.
- An individual sampled with a sampling probability of π_i represents $1/\pi_i$ individuals in the population. The value $1/\pi_i$ is called the ***sampling weight***.
- If it turns out that 400 of our sample have high blood pressure, we would expect 400 x 10000 = 4 million people with high blood pressure in California.

Horvitz-Thompson Estimator

Because of the importance of sampling weights and the inconvenience of writing fractions it is useful to have a notation for the weighted observations. If X_i is a measurement of variable X on person i, we write

$$\check{X}_i = \frac{1}{\pi_i} X_i$$

Given a sample of size n the Horvitz-Thompson estimator $\hat{T}_X$for the population total T_X of X is

$$\hat{T}_X = \sum_{i=1}^{n} \frac{1}{\pi_i} X_i = \sum_{i=1}^{n} \check{X}_i$$

The variance estimate is

$$\widehat{\text{var}}\left[\hat{T}_X\right] = \sum_{i,j} \left(\frac{X_i X_j}{\pi_{ij}} - \frac{X_i}{\pi_i} \frac{X_j}{\pi_j} \right)$$

3.3 Design Effect

Compared to a SRS (Sampling random sampling), all of the design features of a complex survey, such as, stratification, cluster sampling, and weighting generally influence the standard errors (SEs) of the estimates. Survey researchers use a ratio called design effect, to account for the difference in SEs between a complex survey versus a SRS:

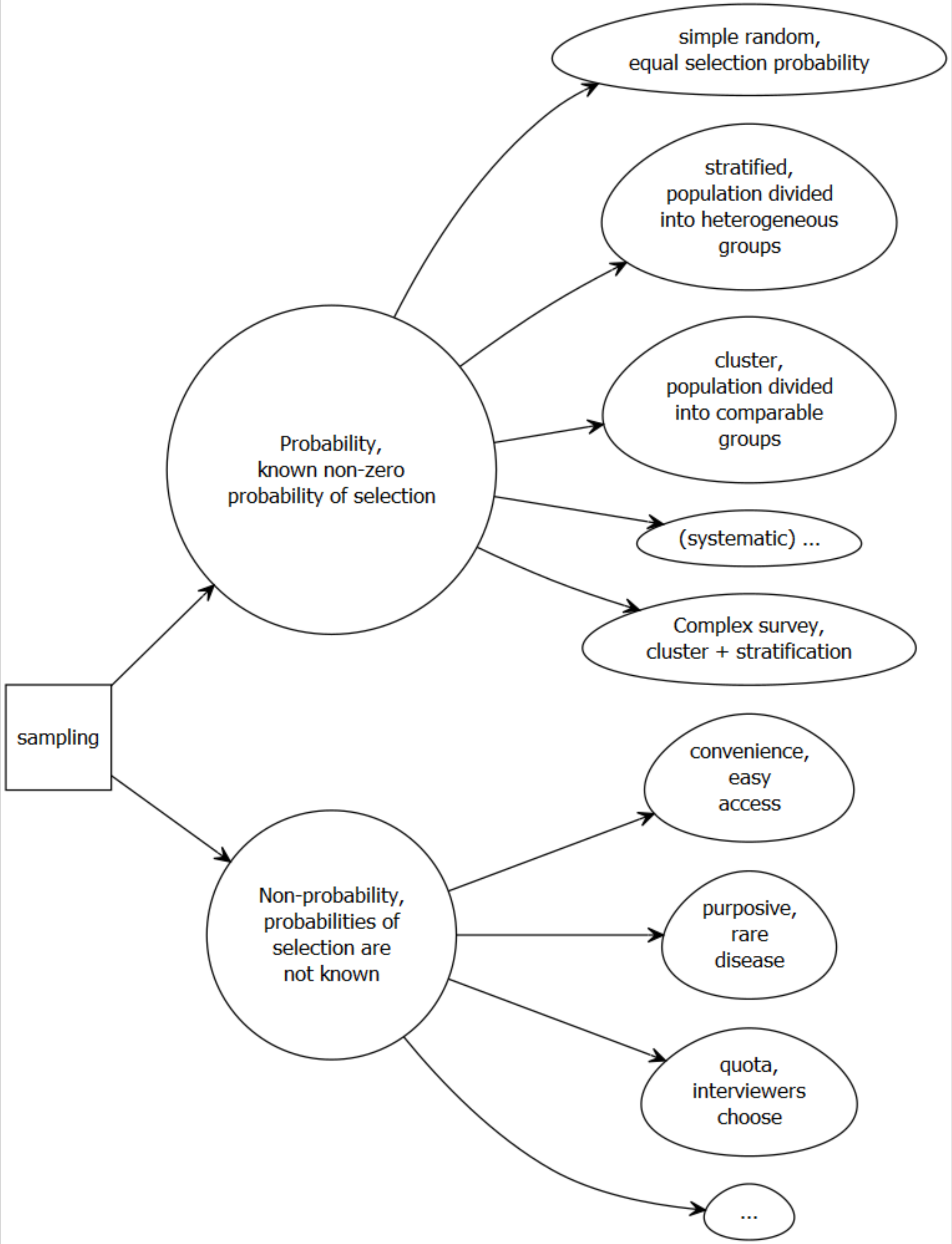
$$DE^2 = \frac{SE^2_{Complex.Survey}}{SE^2_{SRS}}.$$

3.4 Types of Sampling Techniques

3.4.1 Simple Sampling

3.4.2 Stratified Sampling

3.4.3 Cluster Sampling



National Institute of Allergy and Infectious Diseases

NIAID

3.4.1 Simple Sampling

With a simple random sample of size n from a population of size N all the sampling weights are equal to N/n.

The Horvitz-Thompson estimator of the population total of a variable X is

$$\hat{T}_X = \sum_{i=1}^{n} \check{X}_i = \frac{N}{n} \sum_{i=1}^{n} X_i$$

The variance of the Horvitz-Thompson estimator

$$\operatorname{var}\left[\hat{T}_X\right] = \boxed{\frac{N-n}{N}} \times N^2 \times \frac{\operatorname{var}[X]}{n}$$

Finite population correction

The population mean of X can be estimated by dividing the estimated total by the population size N:

$$\hat{\mu}_X = \frac{1}{N}\sum_{i=1}^{n} \check{X}_i = \frac{1}{n}\sum_{i=1}^{n} X_i$$

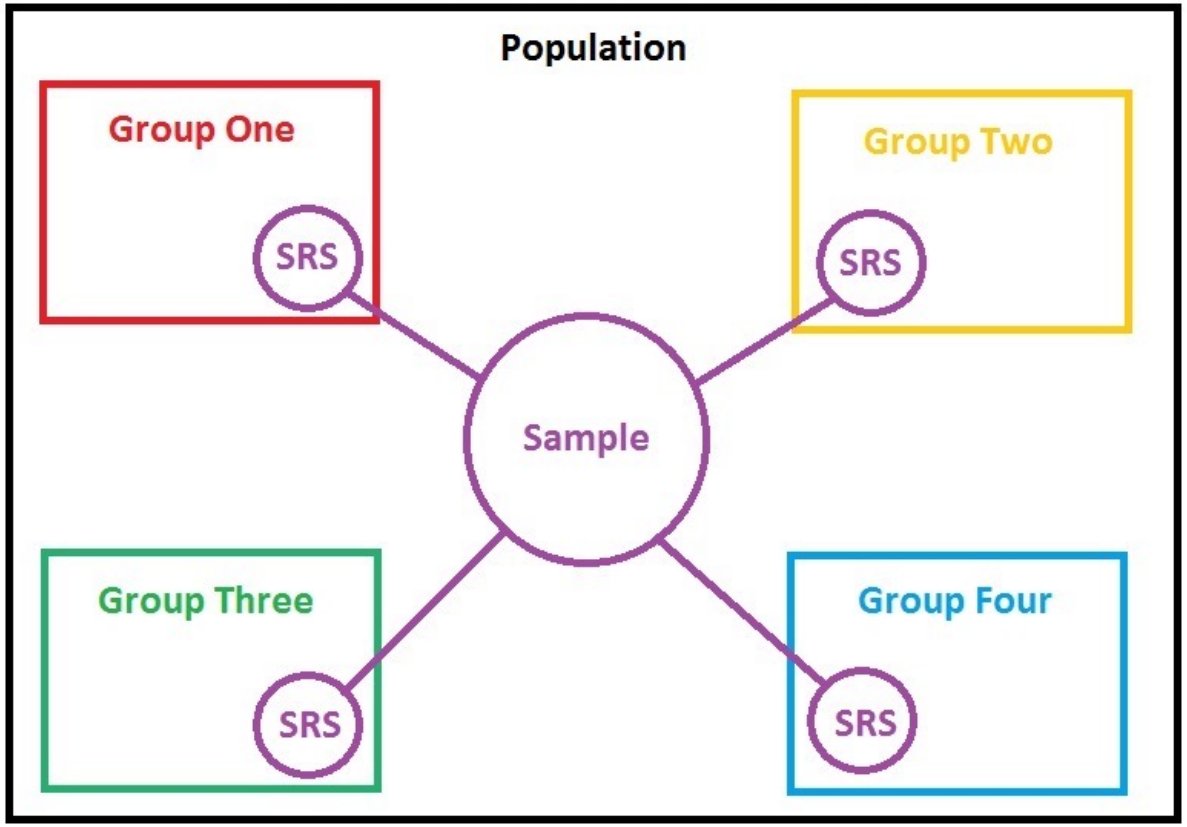
The variance estimate:

$$\widehat{\mathrm{var}}\left[\hat{\mu}_X\right] = \frac{N-n}{N} \times \frac{\widehat{\mathrm{var}}\left[X\right]}{n}$$

3.4.2 Stratified Sampling

- Stratified sampling involves dividing the population up into groups called *strata* and drawing a separate probability sample from each one
- The sample is less variable, and so gives more precise estimates
- Limitation: in order to sample within a stratum, the stratum membership must be known for every individual in the population.

Stratified Random Sampling

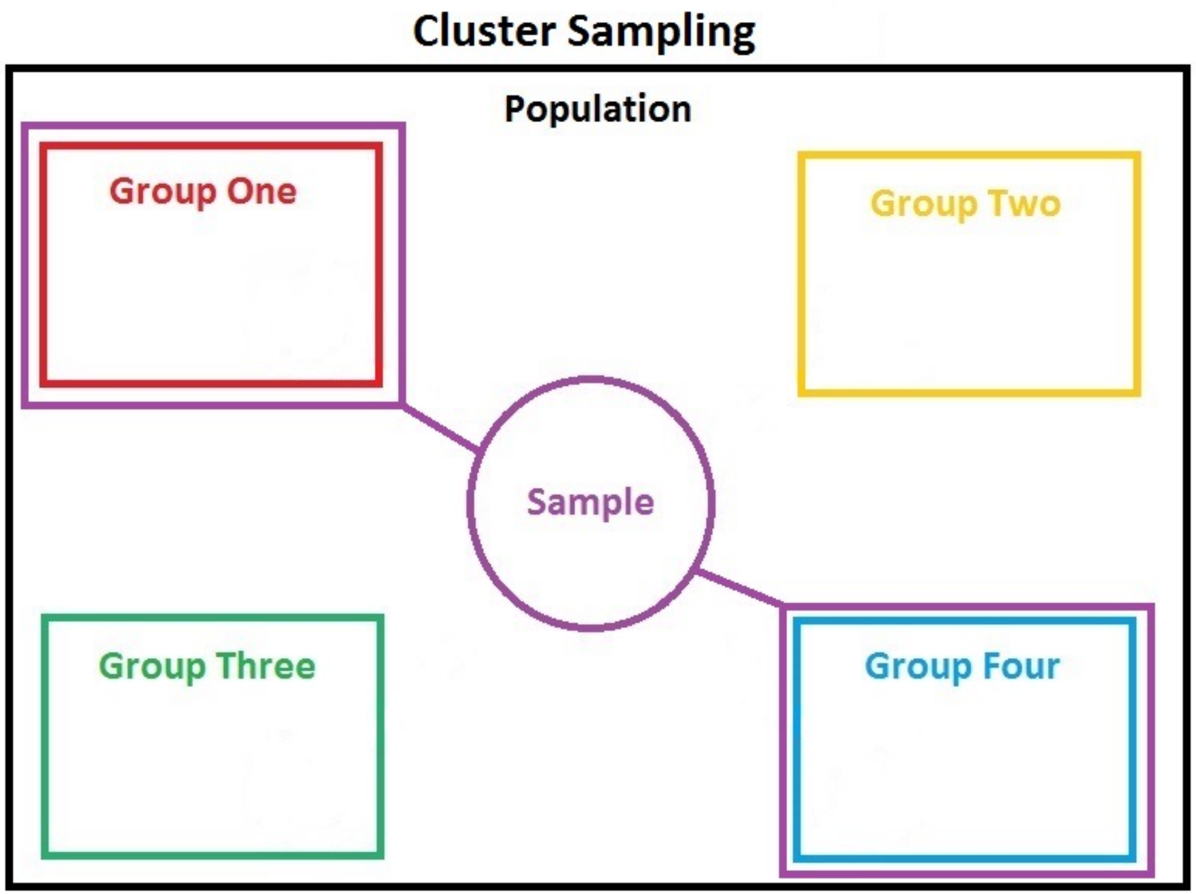


Usual examples:

- different geographical location: Manitoba vs. Nunavut
- high income vs. low income
- gender

3.4.3 Cluster Sampling

- Convenient
- Low cost
- Neighboring subjects may be more correlated with each other.
- Decreased precision for a specified sample size, but can increase sample size and precision for a specified cost.
- Single-stage and multi-stage designs



4. Graphics

The principal difficulty in designing graphics for complex survey data is representing the sampling weights. Three strategies:

- Base the graph on an estimated population distribution.
- Explicitly indicate weights on the graph.
- Draw a simple random sample from the estimated population distribution and graph this sample instead.

4.1 Categorical variables

4.2 One continuous variable

4.3 Two continuous variables

4.4 Conditioning plots

4.1 Categorical Variables

Bar Charts

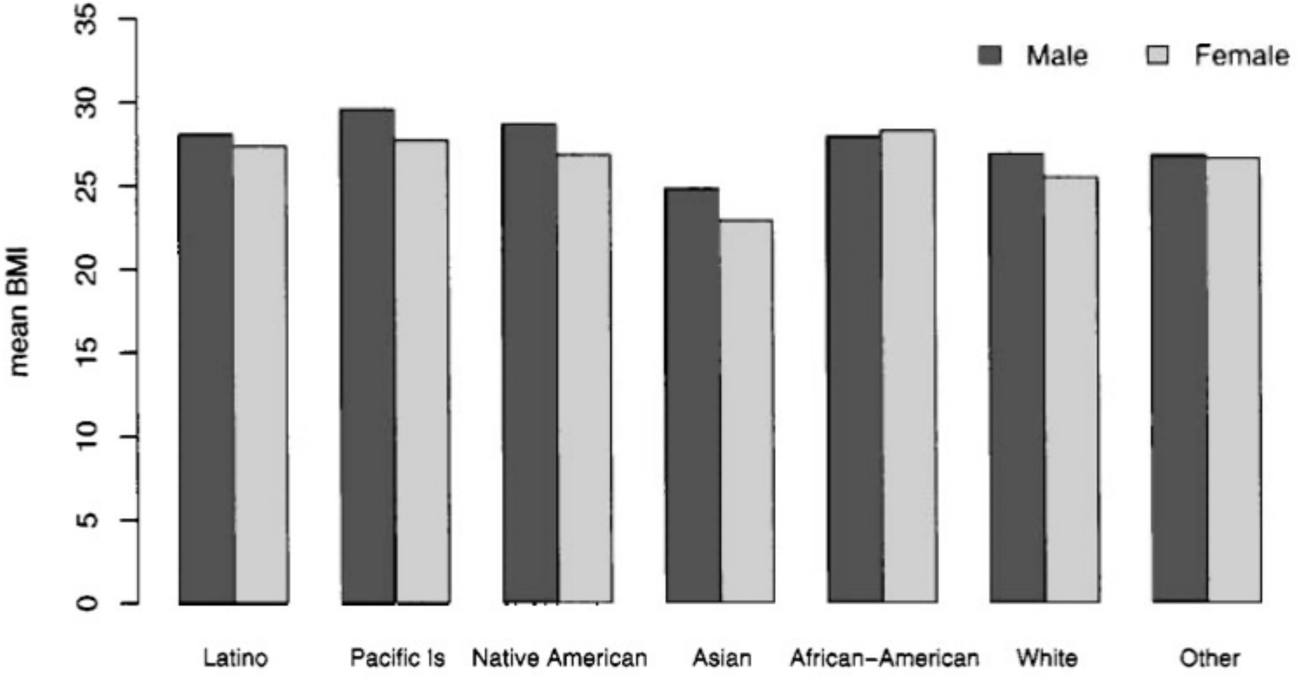


Figure 4.1 Mean BMI by race/ethnicity and gender, from CHIS

Forest Plots

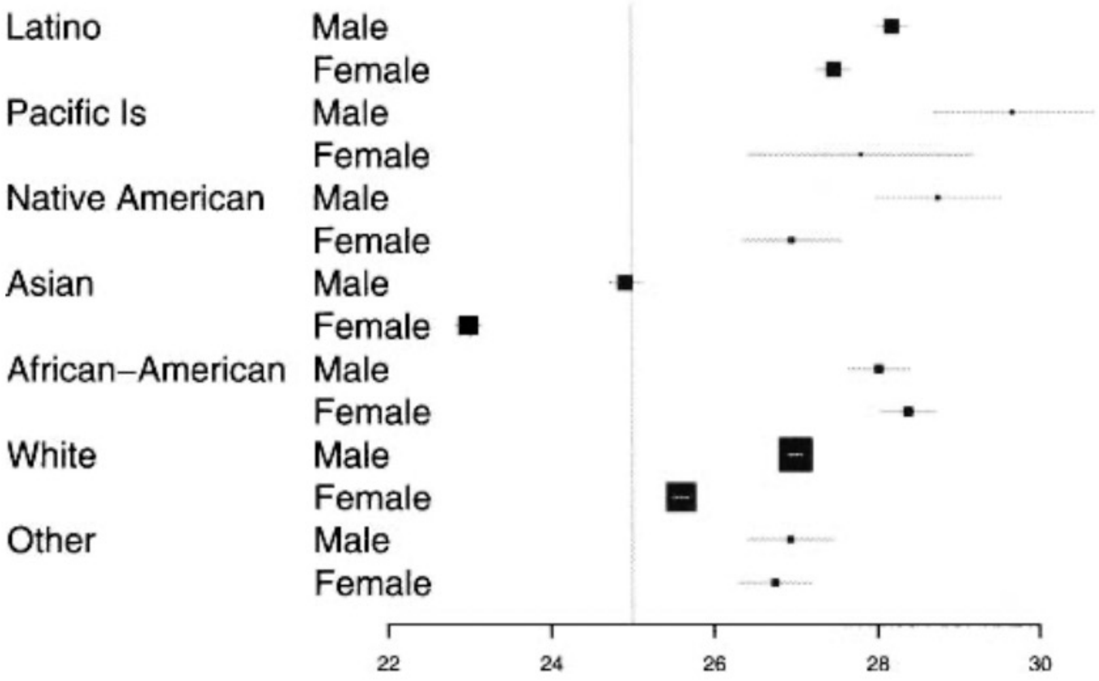


Figure 4.4 Mean BMI± $\sqrt{2}$ standard errors, by race/ethnicity and gender, from CHIS. The vertical line indicate the division between normal and overweight.

Fourfold Plots

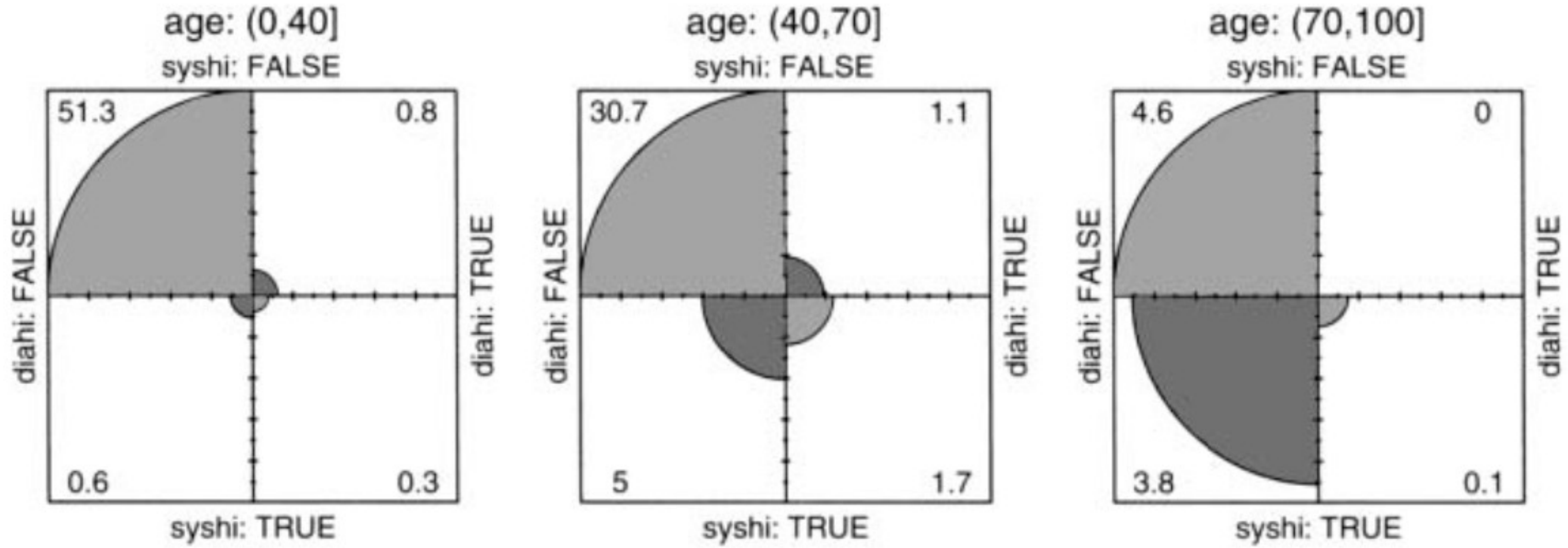


Figure 4.5 Systolic and diastolic hypertension by age group, from NHANES 2003–2004

4.2 One Continuous Variable

Graphs based on the distribution function

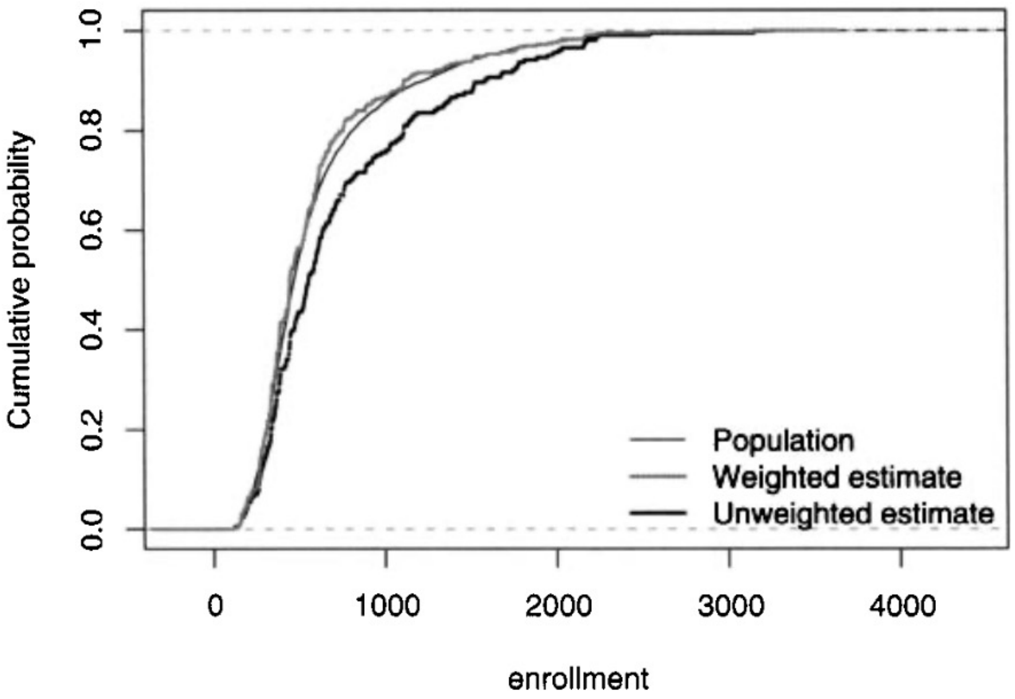


Figure 4.6 Cumulative distribution of California school size: population, weighted estimate, unweighted estimate.

Boxplots

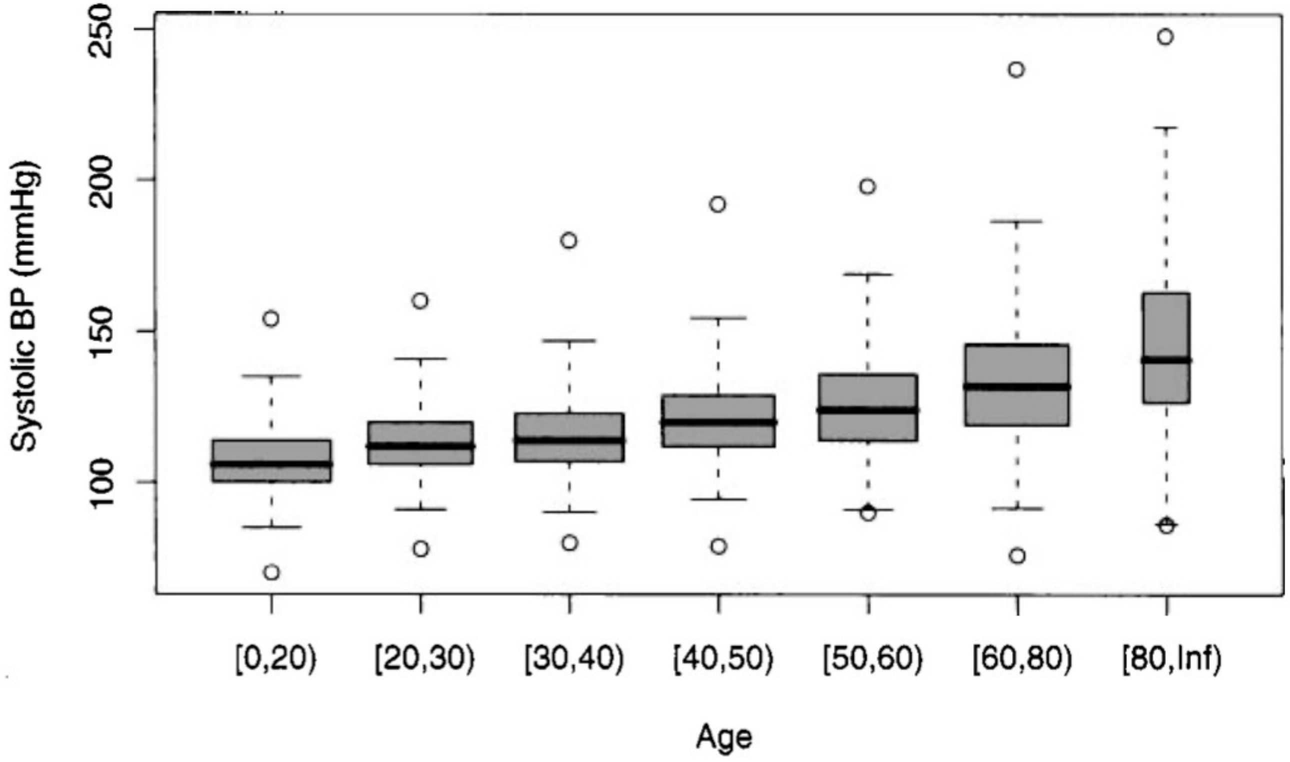


Figure 4.10 Systolic blood pressure by age in NHANES 2003–2004

Graphs based on density

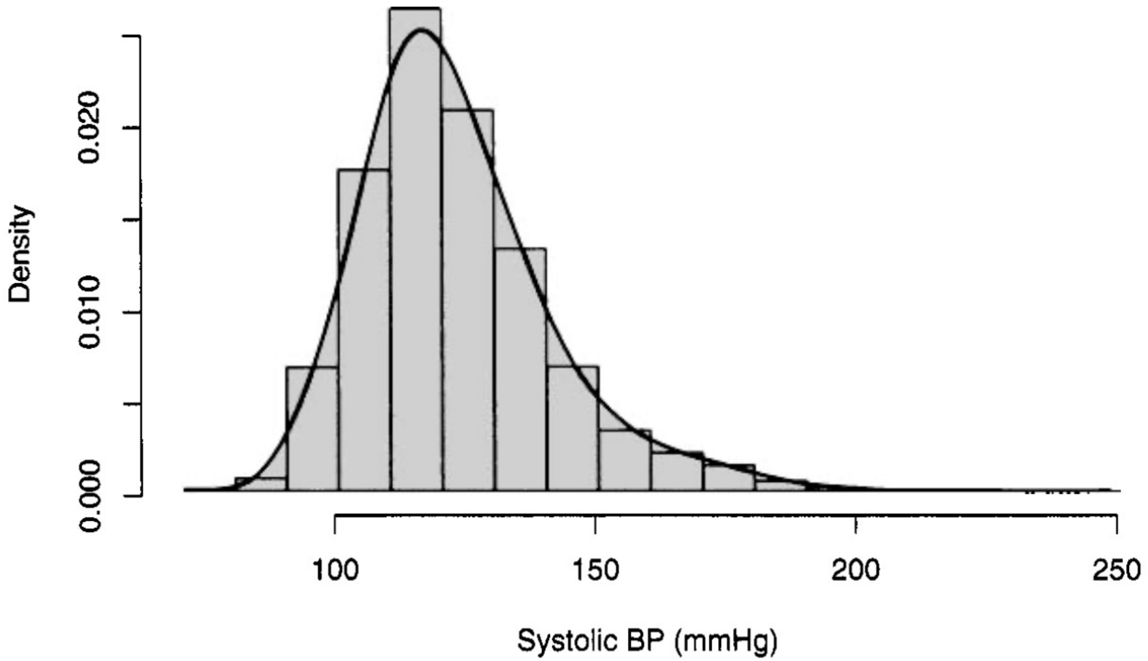


Figure 4.12 Distribution of systolic blood pressure in adults (NHANES 2003–2004)

4.3 Two Continuous Variables

Scatterplots

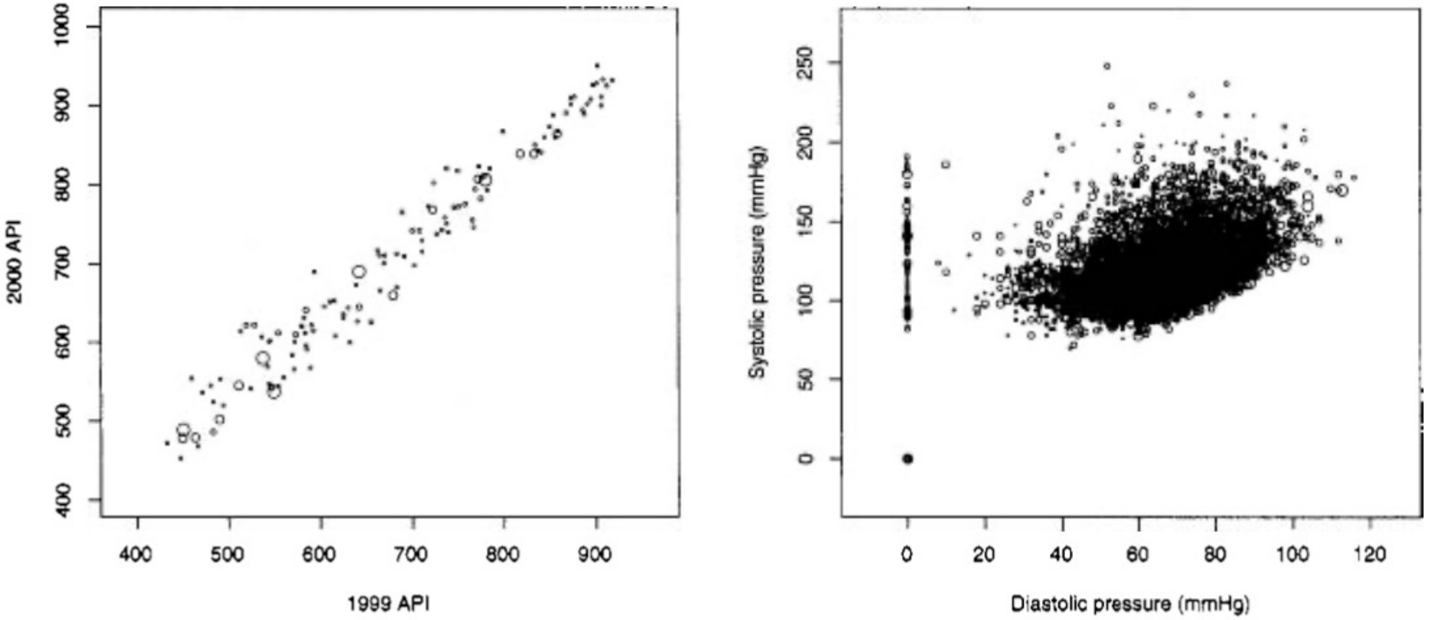


Figure 4.13 Representing sampling weights by glyph size: change in Academic Performance Index from 1999 to 2000, and relationship between systolic and diastolic blood pressure in NHANES 2003–2004.

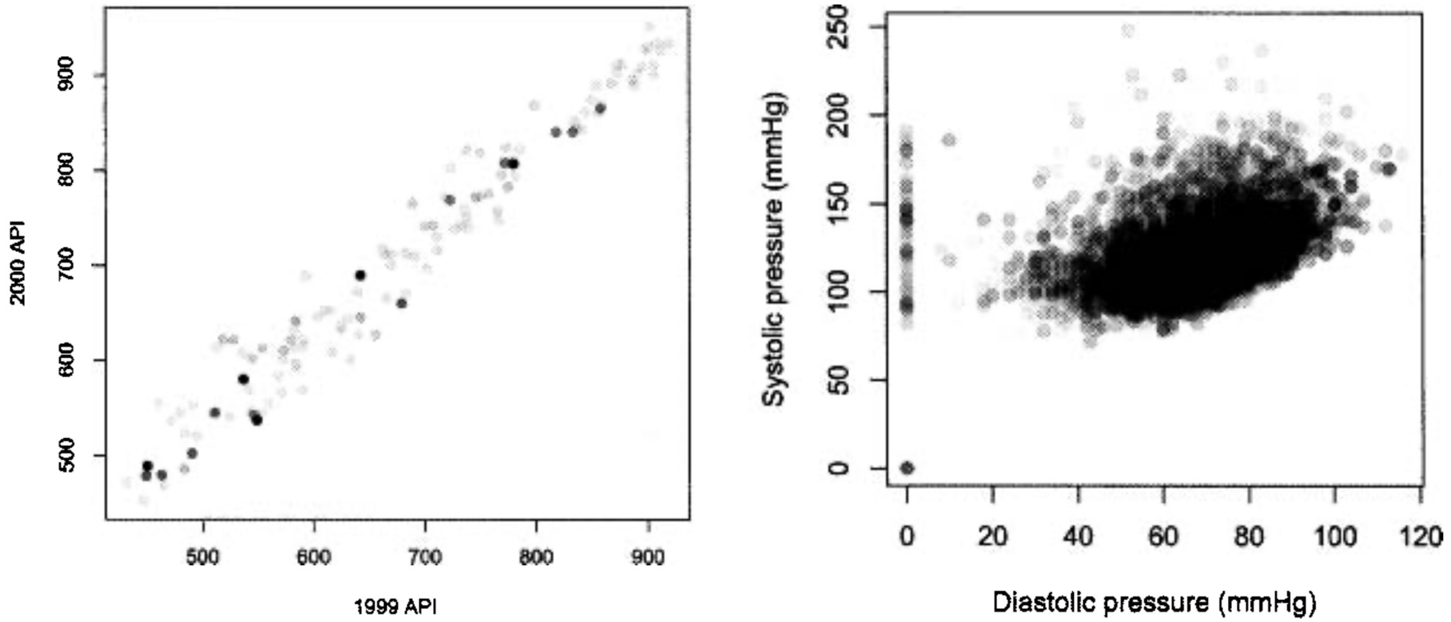


Figure 4.14 Representing sampling weights by shading: change in Academic Performance Index from 1999 to 2000, and relationship between systolic and diastolic blood pressure in NHANES 2003–2004

4.4 Conditioning Plots

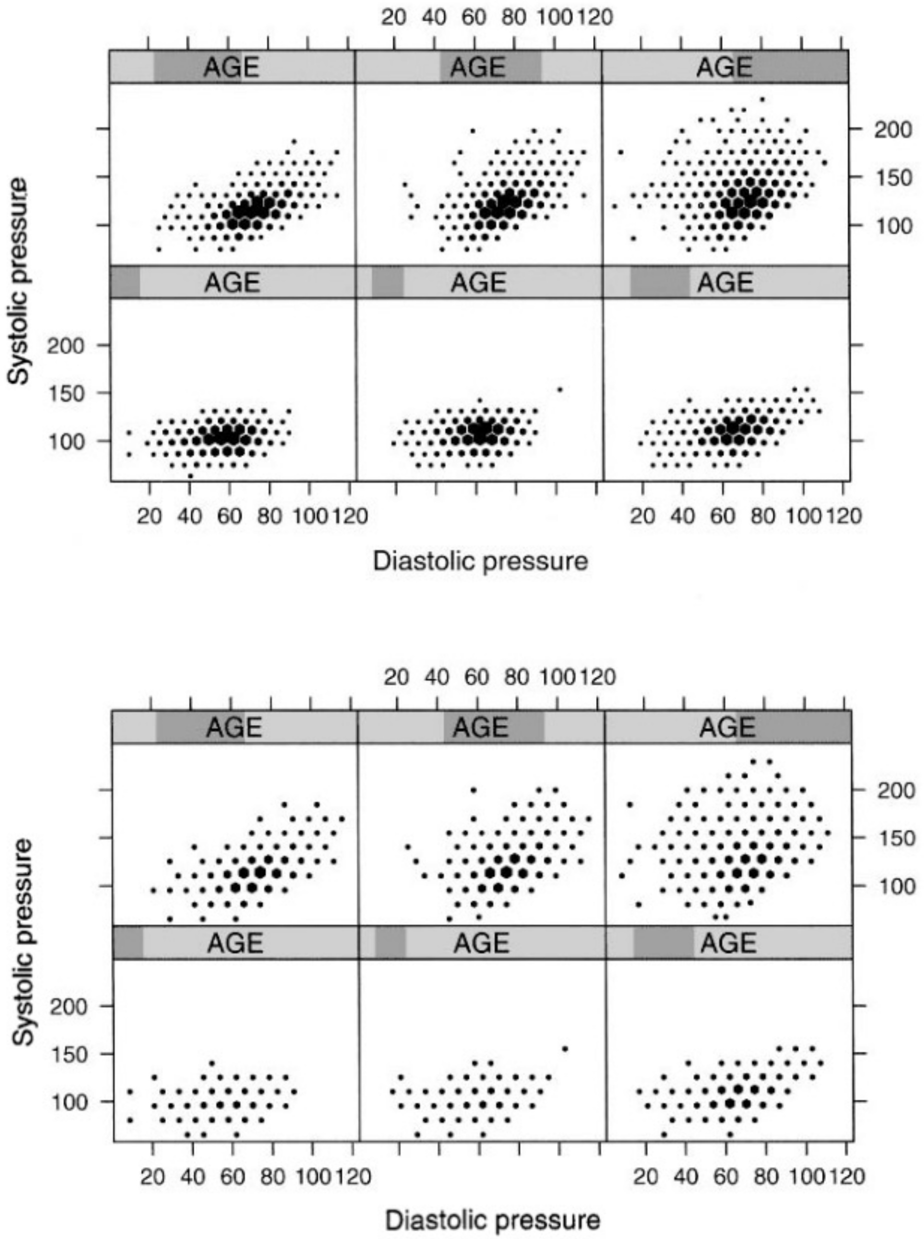


Figure 4.20 Relationship between systolic and diastolic blood pressure by age, using data from NHANES 2003–2004. In the upper plot the hexagons are scaled separately for each panel, in the lower plot the scales are the same across all panels

5. Ratios and Linear Regression

5.1 Ratio Estimation

5.2 Linear Regression

5.1 Ratio Estimation

I. Estimating Ratios

- *E.g. Estimating the proportion of students who took the Academic Performance Index (API) exams*

II. Ratios for Subpopulation estimates

- *E.g. the proportion of people over 65 who have high blood pressure*

III. Ratio estimators of totals

- *E.g. From a sample of 200 schools we estimated that 83.69%∓0.77% of students take the API tests. Suppose we know the total enrollment for all schools in California (3811472), and we want to estimate the total number who took the API tests*

5.2 Linear Regression

- Linear regression summarizes the difference in mean of a *response* variable Y over different values of one or more *predictor* variables X. The working model for linear regression comes in two parts. The *systematic* part of the model describes the differences in mean

$$E[Y] = \alpha + X\beta$$

The random part of the working model says that the variance of Y is constant

$$\mathrm{var}[Y] = \sigma^2$$

- Estimate parameters by sampling-weighted least squares. To minimize the sum of squared residuals over the population

$$RSS = \sum_{i=1}^{N} (Y_i - \alpha - X_i\beta)^2$$

- With data from a complex sample, estimate the population sum of squared residuals using the sampling weights :

$$\widehat{RSS} = \sum_{i=1}^{n} \frac{1}{\pi_i} (Y_i - \alpha - X_i\beta)^2$$

- Consider a data set with just two points (x_1, y_1) and (x_2, y_2)

$$\hat{\beta}_{1,2} = \frac{y_1 - y_2}{x_1 - x_2}$$

- If we have *n* points from a simple random sample there are *n(n* - 1)/2 such pairs of points, so we could compute *n(n*- 1)/2 different slopes, with $\hat{\beta}_{i,j}$ being the slope of the line between (x_i, y_i) and (x_j, y_j)

$$\hat{\beta} = \frac{\sum_{i,j=1}^{n} w_{ij} \hat{\beta}_{ij}}{\sum_{i,j=1}^{n} w_{ij}}$$

where

$$w_{ij} = (x_i - x_j)^2$$

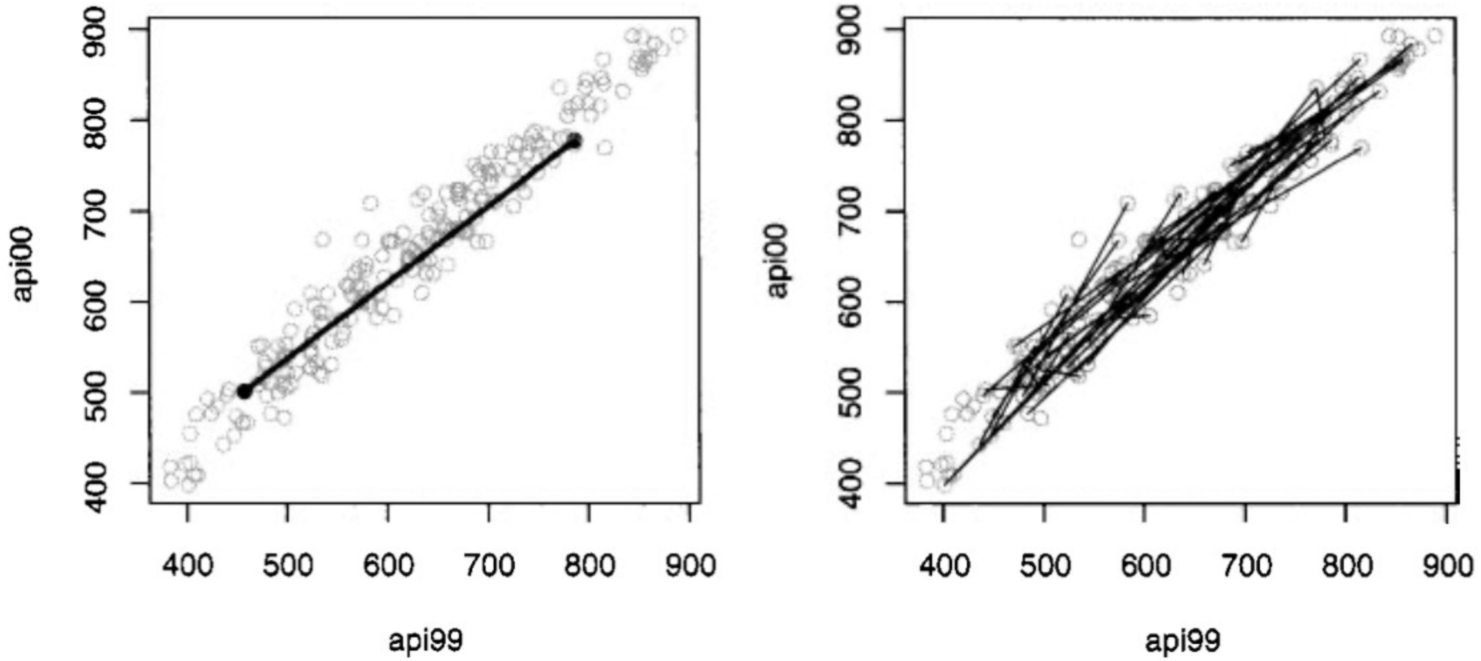


Figure 5.5 The ordinary linear regression estimator is a weighted average of all pairwise slopes

6. Categorical Data Regression

6.1 Logistic Regression

6.2 Ordinal Regression

6.3 Loglinear Models

6.1 Logistic Regression

The logistic regression model for a binary response variable Y and predictor variables $X_1, X_2, \dots, X_p$ is

$$\operatorname{logit} P[Y = 1] = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$$

Where the logit function is defined as $\operatorname{logit}(p) = \log\left(\frac{p}{1-p}\right)$

The logit of P [Y = 1] is also the logarithm of the *odds* of Y = 1, where the odds is defined as P[Y = 1] / P[Y = 0].

The parameters β are now average differences in $logit\ P[Y = 1]$ for a one-unit difference in X. If is a difference in the logarithm of the odds of $Y = 1$ for one-unit differences in X, then $\boxed{e^{\beta}}$ is a ratio of the odds of $Y = 1$ for a one-unit difference in X.

Odds ratio

6.2 Ordinal Regression

A variable with an ordered set of K categories could be turned into a binary variable by dichotomizing at any of the $K-1$ breaks between categories. Dichotomizing at $\mathrm{Y} \leq k$ vs $\mathrm{Y} > k$ would give a logistic regression model

$$\operatorname{logit} \Pr[Y > k] = \alpha_k + x\beta_k$$

In the $K-1$ models arising from the $K-1$ possible choices of k it is possible that β_k is approximately the same, but it is impossible for α_k to be the same since α_k is the log odds of $Y > k$ at $x = 0$. A simplified model could then be

$$\operatorname{logit} \Pr[Y > k] = \alpha_k + x\beta$$

It defines the *proportional odds model* or *ordinal logistic model.* The parameter estimates are log odds ratios for one-unit differences in x averaged over observations and over cutpoints k.

Other possible *link functions*: $\log(-\log \Pr[Y > k]) = \alpha_k + x\beta$

6.3 Loglinear Models

Loglinear models are a class of multivariate models for categorical data.

For example, the health insurance status (Yes/No) by smoking status (Current/Former/Never) in the California Health Interview Survey.

Let us define p_{ij} for the estimated population probability of being in smoking category i and insurance category j.

- Model 1: $\log p_{ij} = \log p_0 + \beta_i + \gamma_j$
- Model 2: $\log p_{ij} = \log p_0 + \beta_i + \gamma_j + \kappa_{ij}$

where each Greek letter represents a set of scores that add up to zero and are uncorrelated with the other scores.

7. Tests in Contingency Tables

- Pearson chi-squared statistic
- Wald test

8. Missing Data

8.1 Item Non-Response

8.2 Two-Phase Estimation for Missing Data

8.3 Imputation of Missing data

8.1 Item Non-Response

Two broad classes of approach to item non-response:

1) Model the non-response as part of the sampling mechanism, in a two-phase design in which some variables are measured on the whole sample and others on a subsample.
2) Impute the missing data, using the observed information on each subject as a guide to plausible values for the missing information, like multiple imputation and reweighting.

8.2 Two-Phase Estimation for Missing Data

We can think about this as the result of a two-phase design: the original sample is selected and the variables with no missing data are measured, then a subset of people is selected and the remainder of the variables are observed

8.3 Imputation of Missing Data

Multiple Imputation

Two steps:

1) Construct a model that can be used to predict the missing data, and fit this model to the observed data.
2) Once this model is constructed, the missing data are sampled from the predictive distribution of the model

Reference

- Complex Surveys: A Guide to Analysis Using R by Thomas Lumley,
- NHANES Tutorials
- Survey Data: Design and Examples
- Analysis of Health Surveys by Korn, Edward Lee., Graubard, Barry I.
- Analysis of Survey Data by Chambers, R. L. (Ray L.); Skinner, C. J.
- The 3 books are available at NIH library