36-303: Sampling, Surveys and Society

References, Graphs & Models

Brian W. Junker
132E Baker Hall
brian@stat.cmu.edu

Handouts & Announcements

- These Lecture Notes
- Korn & Graubard: Scatterplots with Survey Data
- Additional handouts in the Week 12 area of the website!
- Exam next Tue (review this Thu)

Outline

- References in Scholarly Articles
- Making Graphs with Weighted Data
- Regression Models with Weighted Data

References in Scholarly Articles

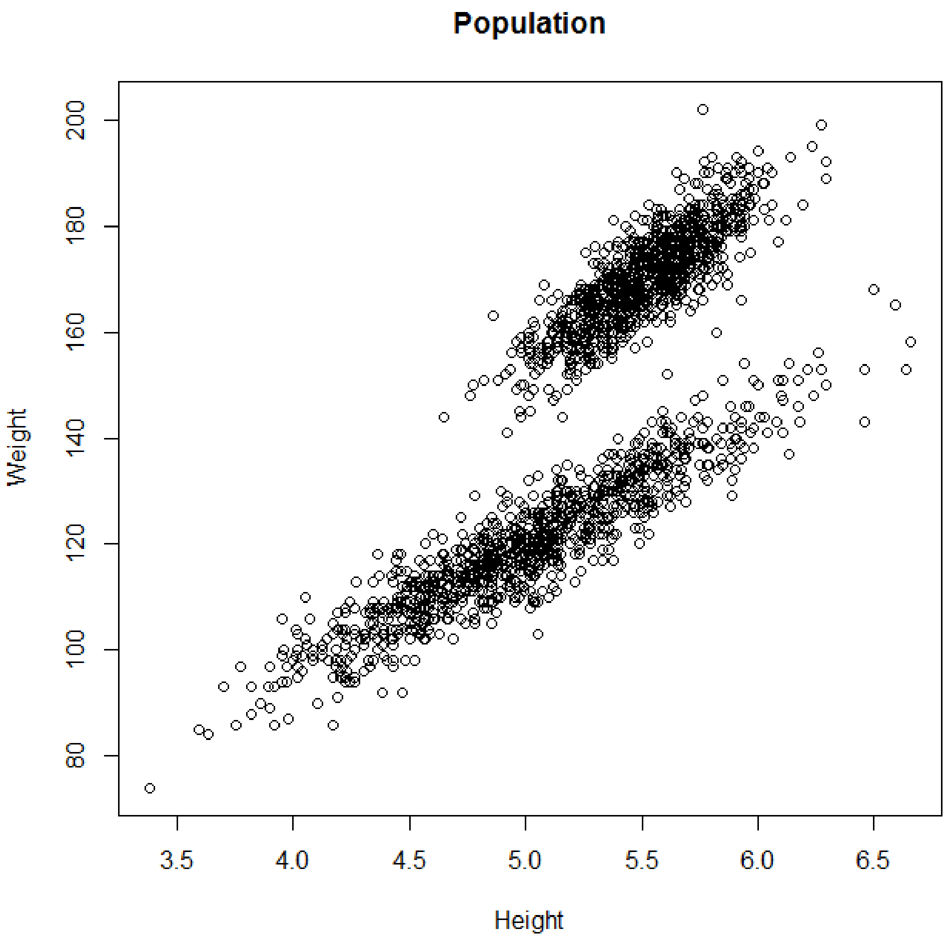
- Different fields have different conventions
- In Psychology, Social Sciences and Statistics there is a fairly common set of conventions:
  - “Note that Smedley (1887) previously conduced a survey like this…”
  - “In a survey similar to ours (Smedley, 1887), men reported more…”
- REFERENCES:
  - Smedley, F.T. (1887). A social survey of attitudes toward so-called “horseless carriages”. *Social Survey Quarterly, 13,* 15-22. Obtained April 1, 2008 from http://www.irreproducible-results.org
  - Author (Date). Title. *Source,* pages. Web-citation.
- See Bem article (on writing research reports) for more examples!
- Good quick reference:
  - http://www.library.cornell.edu/resrch/citmanage/apa

Weights in Plots and Linear Regression

- We have encountered weights in two settings:
  - Design stratification weights (strata and weights determined before we collect data)
    - Variance calculations more complicated but not too bad
  - Post-stratification weights (strata and weights determined after we collect data, when we are worried about “representativeness”
    - Variance calculations involve Taylor Series (Delta Method) or Jackknife
- How do we handle weights, generally, in
  - Plots (Boxplots, Histograms, Scatter plots)
  - Linear regression models: lm(), aov()

Example…

- I constructed a fake population of size N=2000
  - 1000 men
  - 1000 women
  - Fake heights and weights for each
- I took a biased sample of
  - 50 women
  - 150 men



Example (cont'd)

- Post-stratification weights
  - Men: (1000/2000)/(150/200) = 0.6667
  - Women: (1000/2000)/(50/200) = 2
- We will explore
  - Boxplots
  - Histograms
  - Scatter Plots
  - Linear Regression models

Boxplots

- Three options:
  - Plot the unweighted, biased sample
  - Use the weights instead of raw counts to compute quartiles, and make boxplot based on “weighted quartiles”
  - Re-sample the data proportional to the weights
- Compare to population boxplot

Boxplots: Using the weights to calculate quartiles

- Quartiles: sort the data, then…
  - 1st quartile – 25% of the data lie below this
  - median – 50% of the data lie below this
  - 3rd quartile – 75% of the data lie below this
- Weighted quartiles: sort the data, then…
  - 1st quartile – 25% of the *weights* lie below this
  - median – 50% of the *weights* lie below this
  - 3rd quartile – 75% of the *weights* lie below this

Boxplots: Resampling proportional to weights

- The weights are

  0.667, 0.667, ..., 0.667, 2.000, …, 2.000

- Convert them to probabilities by dividing by the sample size (200, = sum of the weights!)

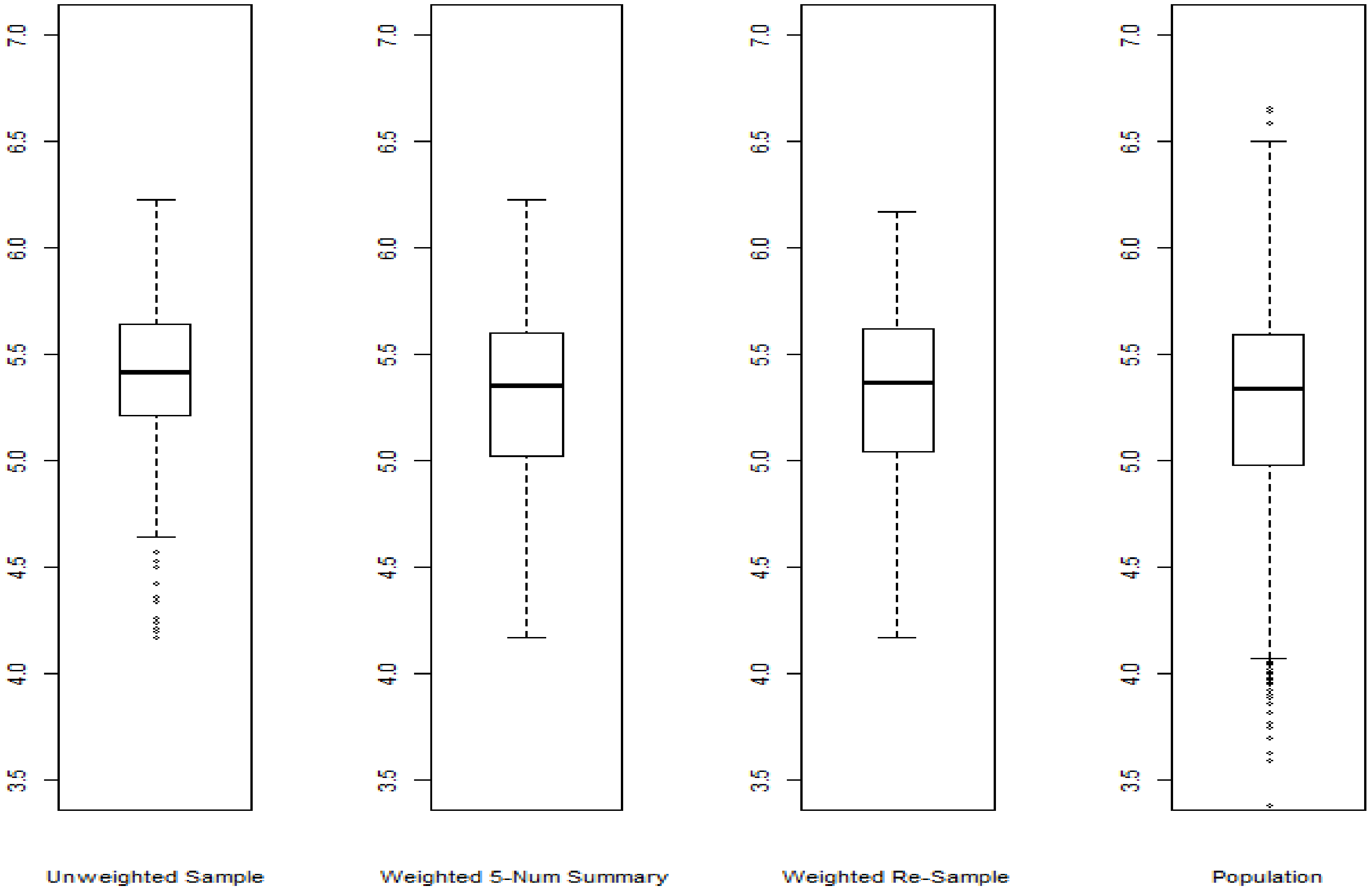
  0.003, 0.003, …, 0.003, 0.010, …, 0.010

- Take an SRS (with replacement!) where each observation in the original sample can be in the new sample with probabilities p above

Compare the 5-number Summaries (for Heights)

	Min	Q1	Med	Q3	Max
Unweighted 5-number Summary	4.17	5.215	5.420	5.64	6.23
Weighted 5-number Summary	4.17	5.020	5.350	5.60	6.23
Weighted Resample	4.17	5.040	5.365	5.62	6.17
Population 5-number Summary	3.38	4.980	5.340	5.59	6.66

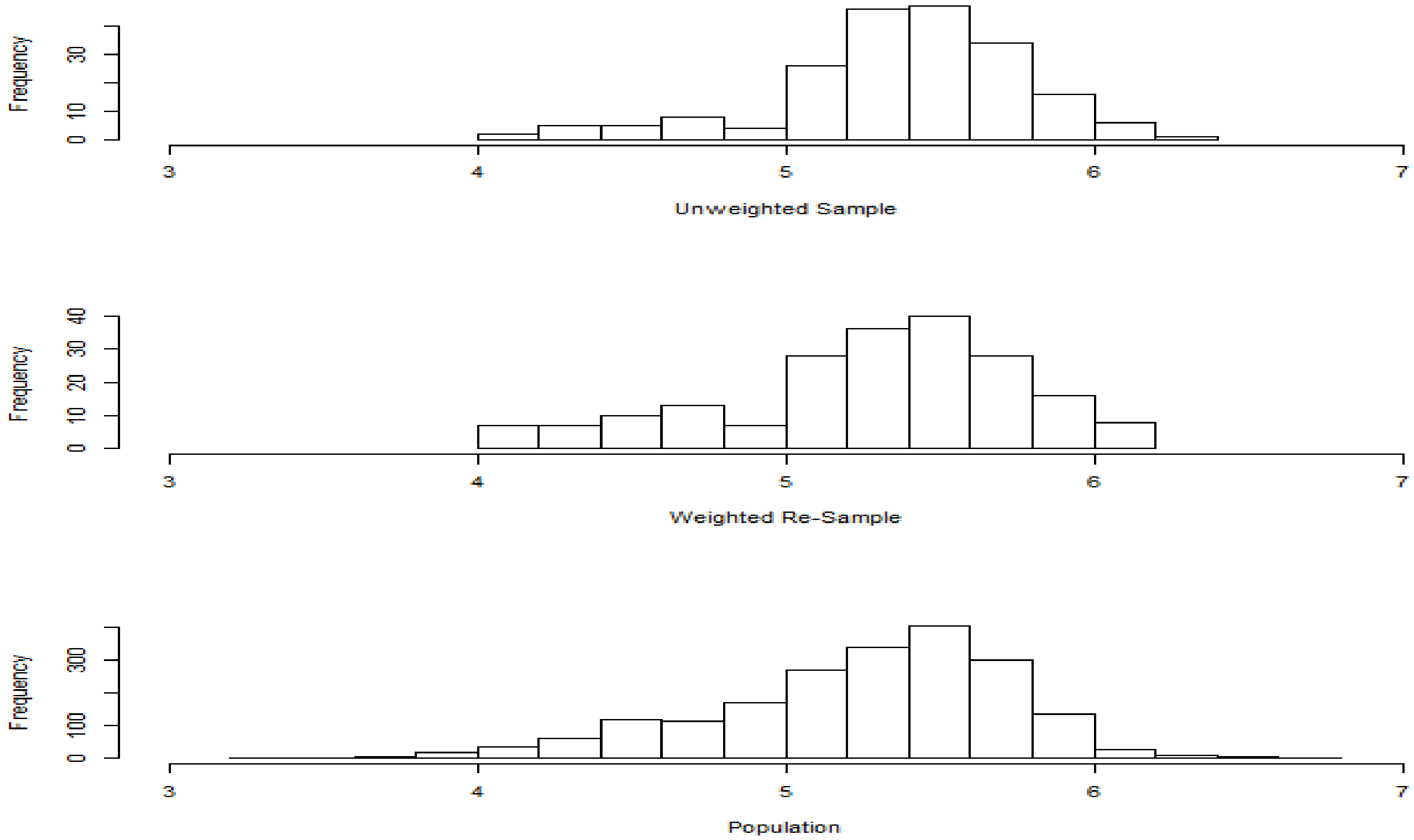
Compare the Boxplots (for Heights)…



Histograms…

- We could use the weights to adjust the heights of the bars in a histogram
  - Just like using the weights to adjust the quartiles for a boxplot!
  - Height of each bar is the sum of the weights for observations in that interval
    - (rather than the count of observations in the interval)
- But it is probably easier to just use the resampling idea

Compare the Histograms (for Heights)…

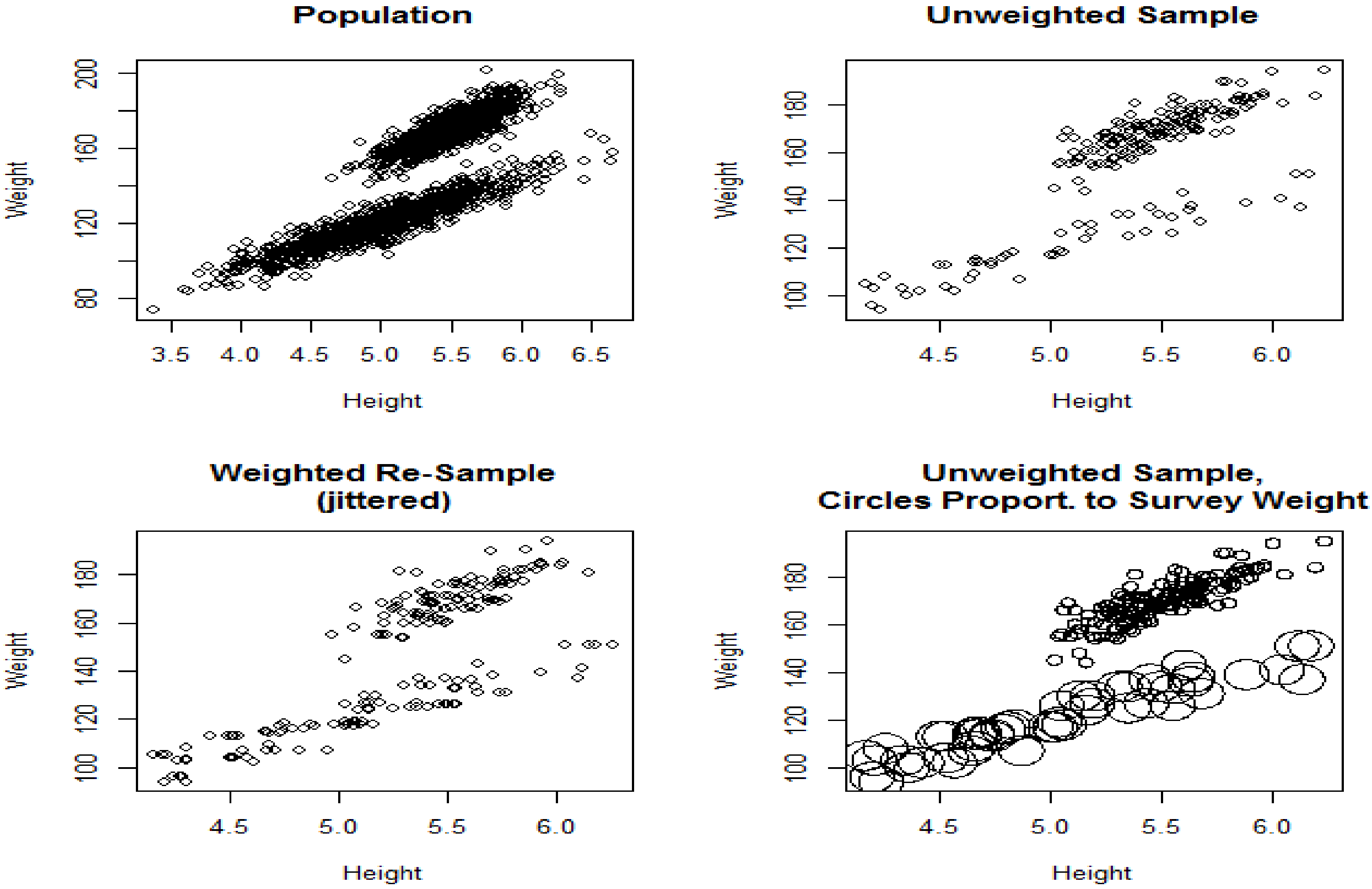


Scatterplots…

- We can resample proportional to the weights again
  - I "jittered" this plot since the resampling can produce duplicate points.
- Another approach would be to
  - plot the unweighted data, but
  - make plotting symbols that are proportional to the size of the post-stratification weights

(this allows us to "see" the real data in the sample, but also to see how much of the population each sampled data point is supposed to represent!)

Compare scatterplots (for heights…)



Population

Unweighted Sample

Weighted Re-Sample (jittered)

Unweighted Sample, Circles Proport. to Survey Weight

Linear Regression

- Here there are (at least!) four options:
  - Run regression on the unweighted data
  - Most regression functions allow you to include weights for each data point, so run the regression on the weighted data
  - Use the jackknife method with weighted jackknife samples to improve point estimates and standard errors, for the weighted regression
  - Resample the data proportional to the weights and run the regression on the resampled data

If regression functions allow you to use weights, why jackknife or resample??

- Regression functions in most statistical packages (R, Minitab, SPSS) allow you to add weights for each observation

- The regression functions assume that the weights represent identical replicated observations
  - *bigger weights* -> bigger sample size -> *smaller standard error*

- But survey weights are like imputation: they tell you how many more people you are assigning this value (height, etc.). Since you cannot be sure this is the right value for them
  - *bigger weights* -> more uncertainty -> *bigger standard error*

- For survey weights, weighted regression gives the right point estimates but the wrong standard errors…

Comparing Linear Regression Results

$$(weight)_i = \beta_0 + \beta_1 (height)_i + \varepsilon_i$$

```
Unweighted Regression:
            Estimate Std. Error
(Intercept)   -92.34      14.14
height         46.42       2.62
```

```
Weighted Regression:
            Estimate Std. Error
(Intercept)   -84.23      13.38
height         43.56       2.53
```

```
Jackknifed Regression:
            Estimate Std. Error
(Intercept)   -84.23      16.76
height         43.56       3.29
```

```
Resampled Regression:
            Estimate Std. Error
(Intercept)   -66.78      14.87
height         40.00       2.80
```

```
Population Regression:
            Estimate
(Intercept)   -98.80
height         46.40
```

How can you do this??

- The plots are fairly easy to make “by hand” in Minitab, Excel, SPSS, R, etc.
- Applying Jackknife to regression takes a little more effort
- If someone on your team knows R…
  - Online handout:

    “plotting and regression with weights.r”

Summary

- For graphs that “count things”, best results by adding up weights instead of counting
  - If it is impossible (scatterplots) or inconvenient (histograms) then resampling proportional to weights is OK
    - But it introduces additional sampling error
- For regression and similar calculations, best results by jackknife or delta method
  - If it is difficult to use jackknife or delta method then resampling proportional to weights is OK
    - But it introduces additional sampling error

Review

- References in Scholarly Articles
- Making Graphs with Weighted Data
- Regression Models with Weighted Data
- Exam next Tue (review this Thu)