36-303: Sampling, Surveys and Society

Exam, References, Graphs & Models

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

Handouts & Announcements

- These Lecture Notes
- Additional handouts in the Week 13 area of the website!
- Second Round: Peer Citizenship Review
  - Please fill out for each (other) member of your team
  - EMAIL to me on Apr 29 – Same day as final papers

Outline

- Exam Results
  - Solutions online in “exam2” area of website
- Upcoming Events
- References in Scholarly Articles
- Making Graphs with Weighted Data
- Regression Models with Weighted Data

Exam Results

- I did not have a chance to look over the graded exams
- I will hand graded exams back on Thurs
- Solutions will be posted later today.

Upcoming Events

- ***Today*** is the last formal lecture in the class
- ***Next three class periods*** – final presentations
  - see order on next slide
  - attendance (and participation) is again mandatory
- ***Fri April 29***:
  - Email me final reports (one pdf per group!)
  - Email me second round of peer reviews for the course
    - Peer evaluation blanks are posted online as before
- ***May 4***: Meeting of the Minds Research Conference
  - Seven groups from this class are presenting posters!
  - Please feel free to arrange meeting with me next week if you'd like advice on making a good poster

Order of Final Presentations

- ***Thu Apr 23***
  - Team F - *Caffiene consumption on campus*
  - Team C - *Faculty attitudes toward student attendance and performance*
  - Team G - *Faculty attitudes toward +/- grading*
- ***Tue Apr 26***
  - Team B - *Students' attitudes toward alcohlic energy drinks*
  - Team A - *Students' choice of majors*
  - Team H - *Undergrad prospects after graduation*
- ***Thu Apr 28***
  - Team D - *Student involvement at Carnegie Mellon*
  - Team E - *Accuracy of Pittsburgh bus schedules*
  - Team I - *School childrens' familiarity with architecture concepts*

References in Scholarly Articles

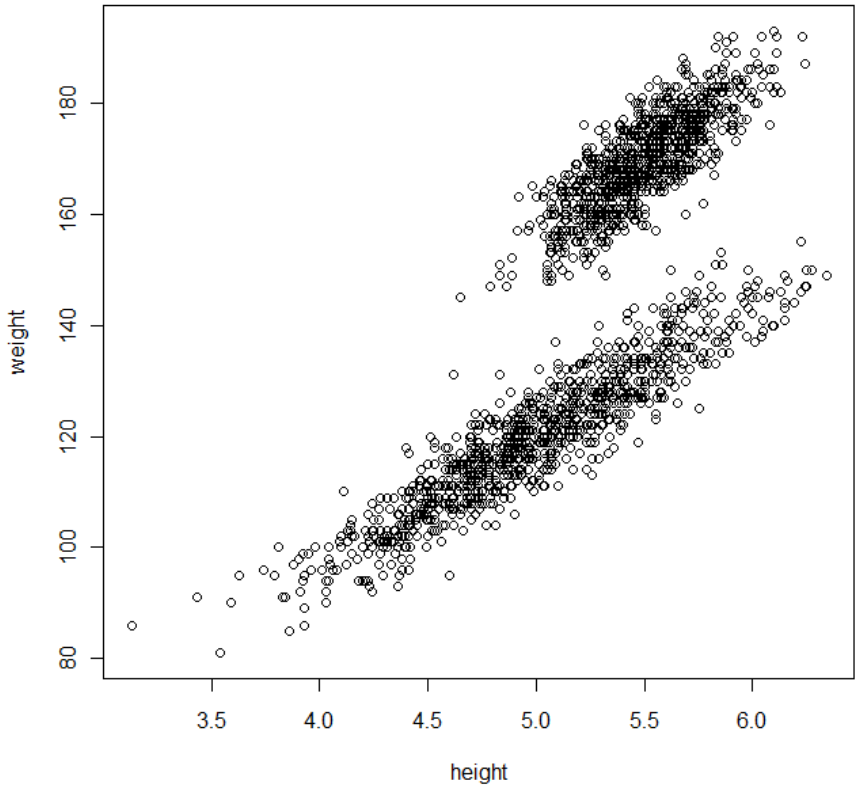
- Different fields have different conventions
- In Psychology, Social Sciences and Statistics there is a fairly common set of conventions:
  - "Note that Smedley (1887) previously conduced a survey like this…"
  - "In a survey similar to ours (Smedley, 1887), men reported more…"
- REFERENCES:
  - Smedley, 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

- Post-stratification weights are important when we are worried about “representativeness”
- We know they are a pain for variance calculations
  - Taylor Series
  - Jackknife
- How do we handle weights 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

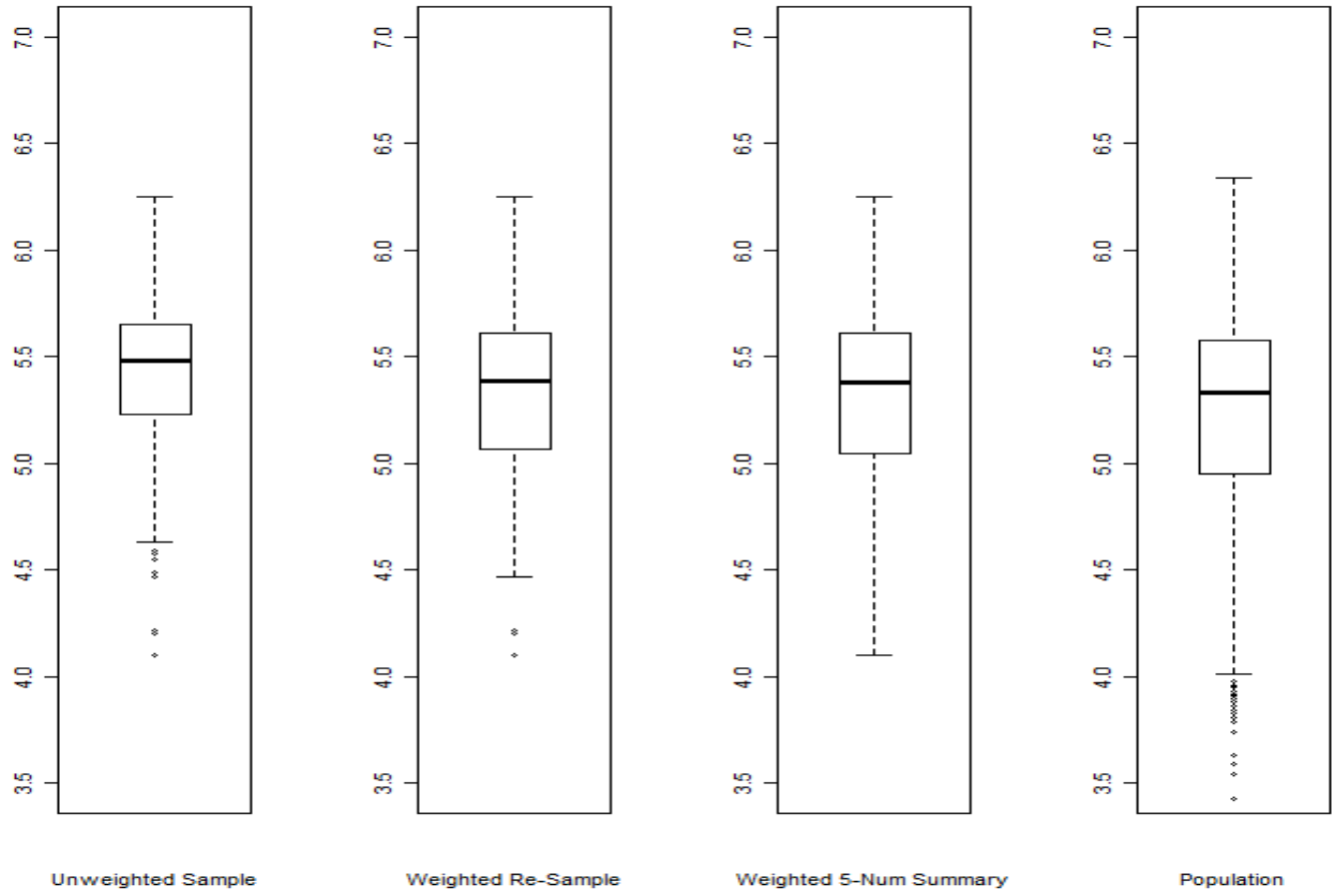
  0.667, 0.667, ..., 0.667, 2.000, …, 2.000

- 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

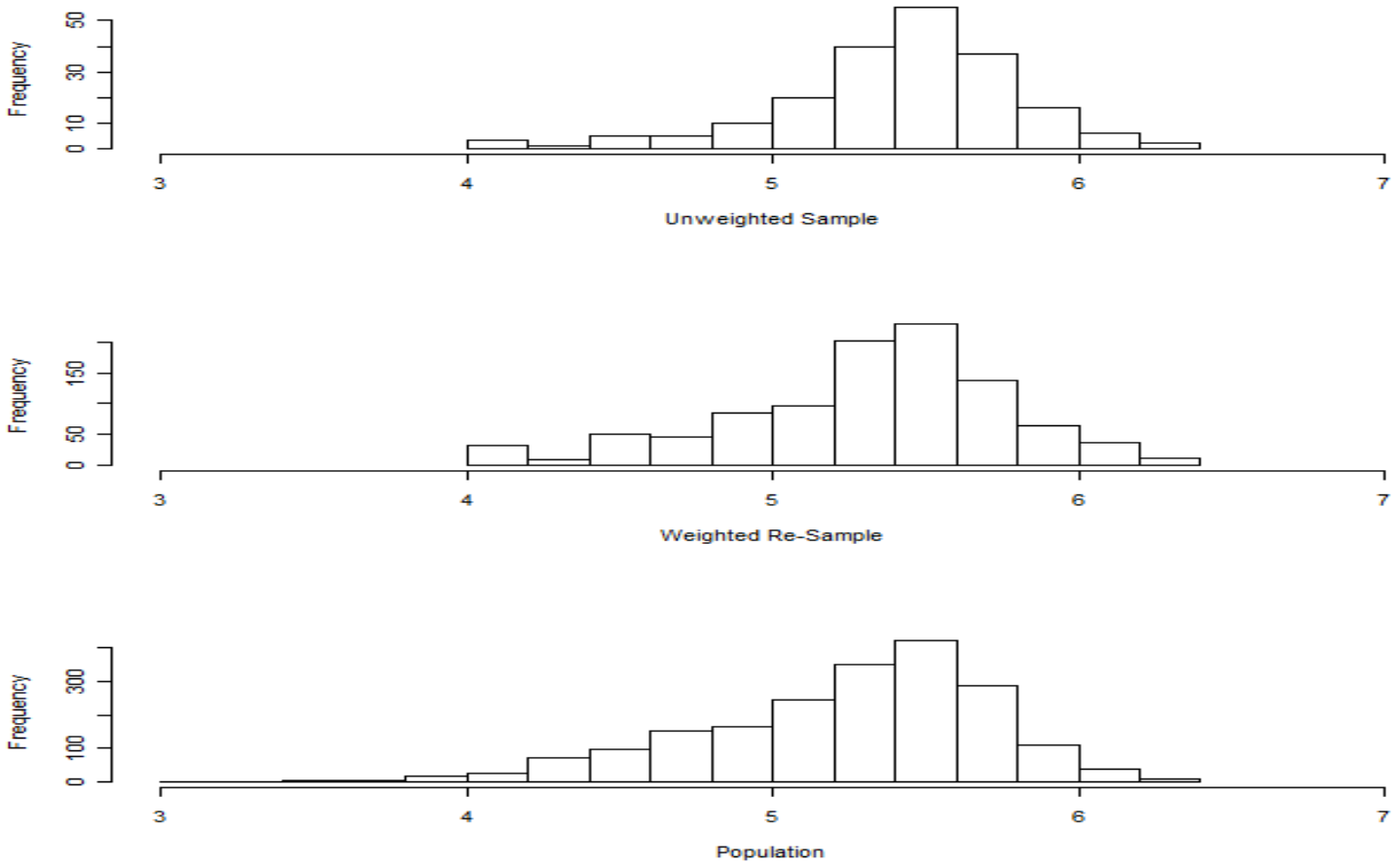
Compare the boxplots (for heights)…



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!
- But it is probably easier to just use the resampling idea

Compare the histograms (for heights)…

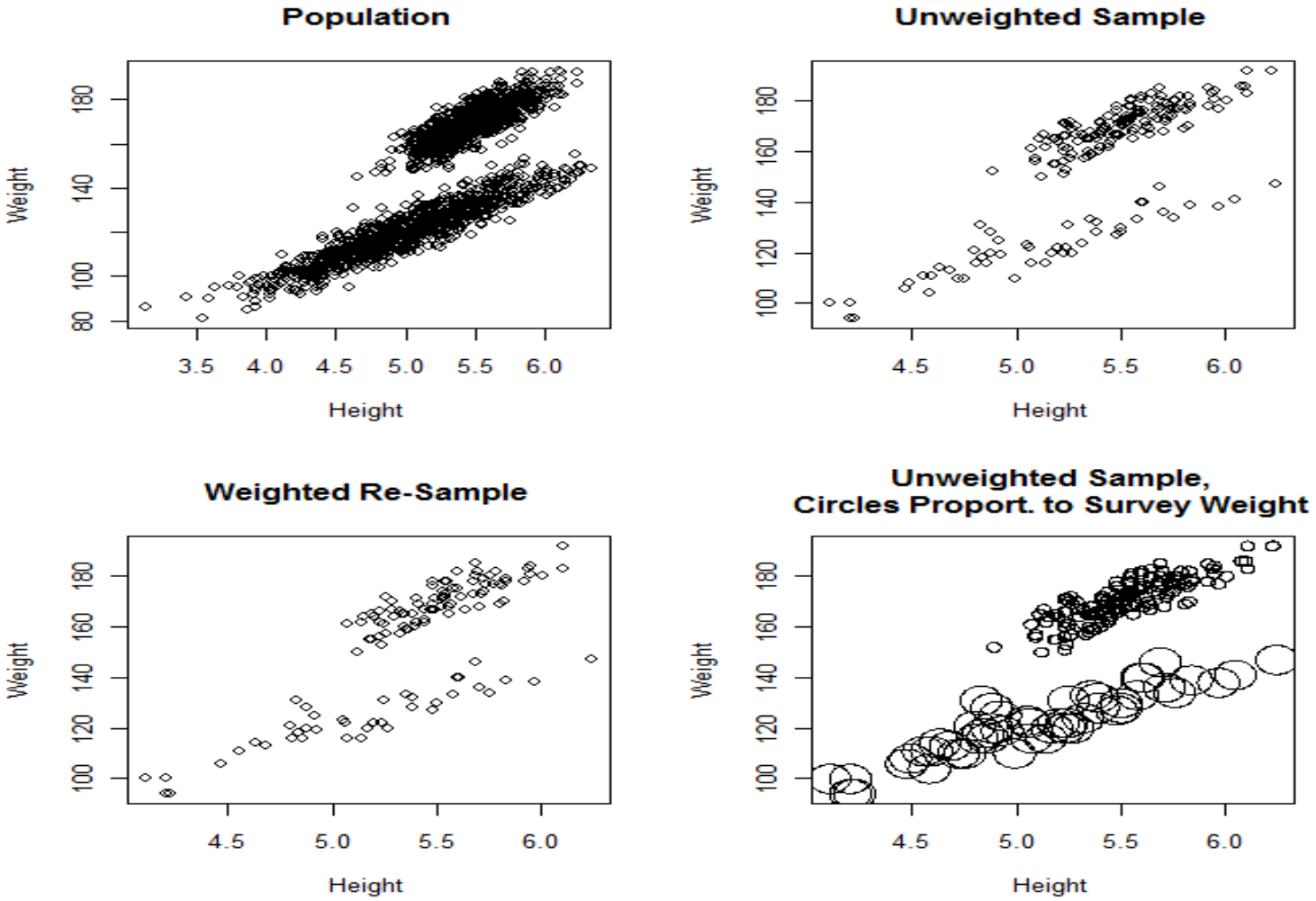


Scatterplots…

- We can resample proportional to the weights again
- 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…)



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)	-122.94	14.64
height	51.95	2.71

Weighted Regression:

	Estimate	Std. Error
(Intercept)	-119.05	14.13
height	50.00	2.67

Jackknifed Regression:

	Estimate	Std. Error
(Intercept)	-91.84	17.25
height	44.85	3.34

Resampled Regression:

	Estimate	Std. Error
(Intercept)	-67.67	15.39
height	39.94	2.88

Mean Resamp. Regr.'s:

	Estimate	Std. Error
(Intercept)	-91.33	12.61
height	44.75	2.49

Population Regression:

	Estimate
(Intercept)	-102.68
height	47.16

How can you do this??

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

    “plotting and regression with weights.r”
  - The R package “survey” from CRAN does all this and more, automagically!

Review

- Exam Results
- Upcoming Events – Final Presentations
- Peer Reviews
  - Due on Apr 29 (along with final papers)
- Final Drafts of Papers
  - Email one pdf per team, Apr 29
- References in Scholarly Articles
- Making Graphs with Weighted Data
- Regression Models with Weighted Data