This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike License. Your use of this material constitutes acceptance of that license and the conditions of use of materials on this site.

Copyright 2009, The Johns Hopkins University and John McGready. All rights reserved. Use of these materials permitted only in accordance with license rights granted. Materials provided "AS IS"; no representations or warranties provided. User assumes all responsibility for use, and all liability related thereto, and must independently review all materials for accuracy and efficacy. May contain materials owned by others. User is responsible for obtaining permissions for use from third parties as needed.

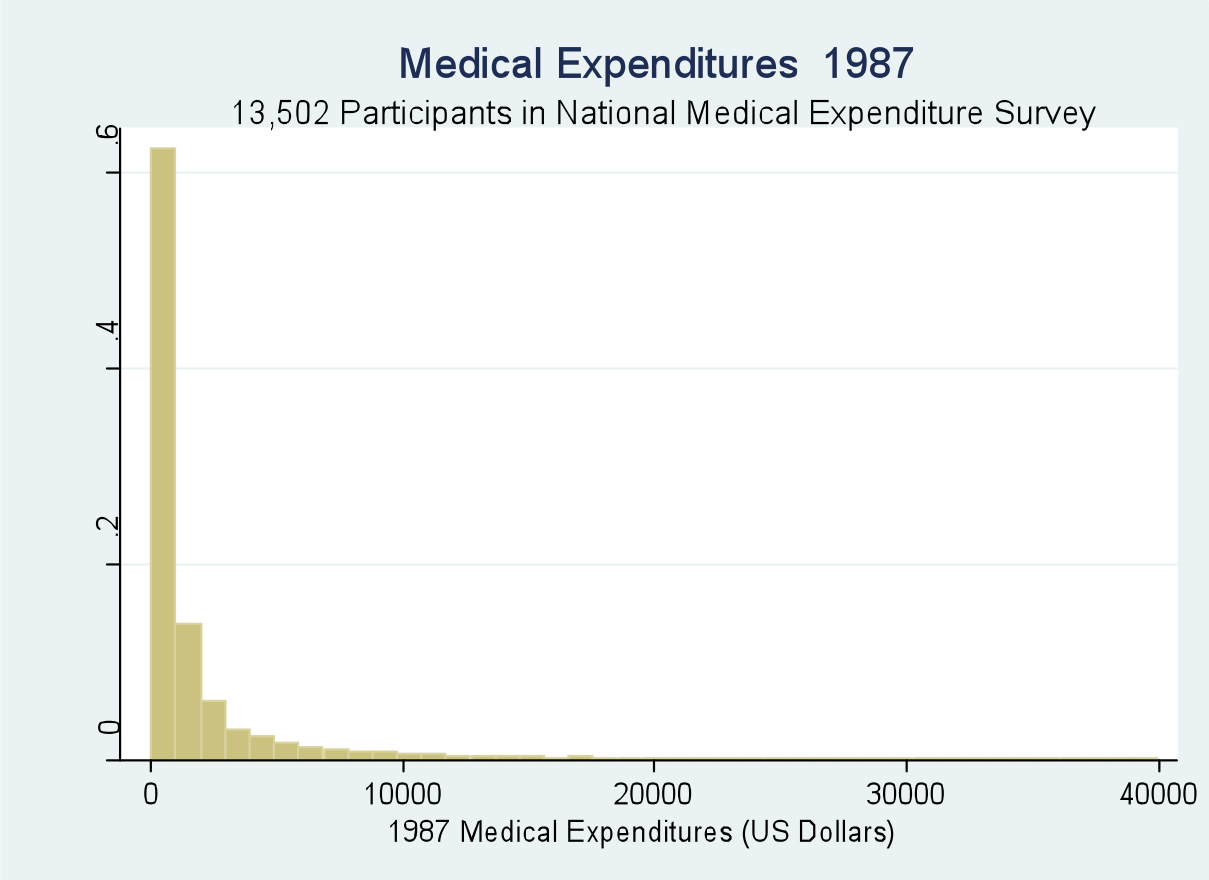
JOHNS HOPKINS
BLOOMBERG
SCHOOL of PUBLIC HEALTH

Lecture 1f: Practice Problem Solutions

John McGready
Johns Hopkins University

Practice Problems

1. The following histogram shows the distribution of medical expenditures (in U.S. dollars in the year 1987) for participants in the National Medical Expenditures Survey (NMES)
   - Mean: $2,300; Median $588; SD $4,957



Practice Problems

a) How would you characterize this distribution (symmetric, right skewed, etc.)?
   - While this is somewhat of a subjective question, I assume most of us would not hesitate to call this a “right-skewed” distribution
   - The extreme values are large relative to the majority; and the mean is much larger than the median

Practice Problems

b) Suppose you take a random sample of 100 observations from this "population" of over 13,000 respondents. What shape will the histogram of the these sample values likely have?
   - As the sample is random, it's characteristics should mimic the characteristics of the larger population of values form which it is selected
   - As such, the distribution of the 100 sample values will also likely be right skewed

Practice Problems

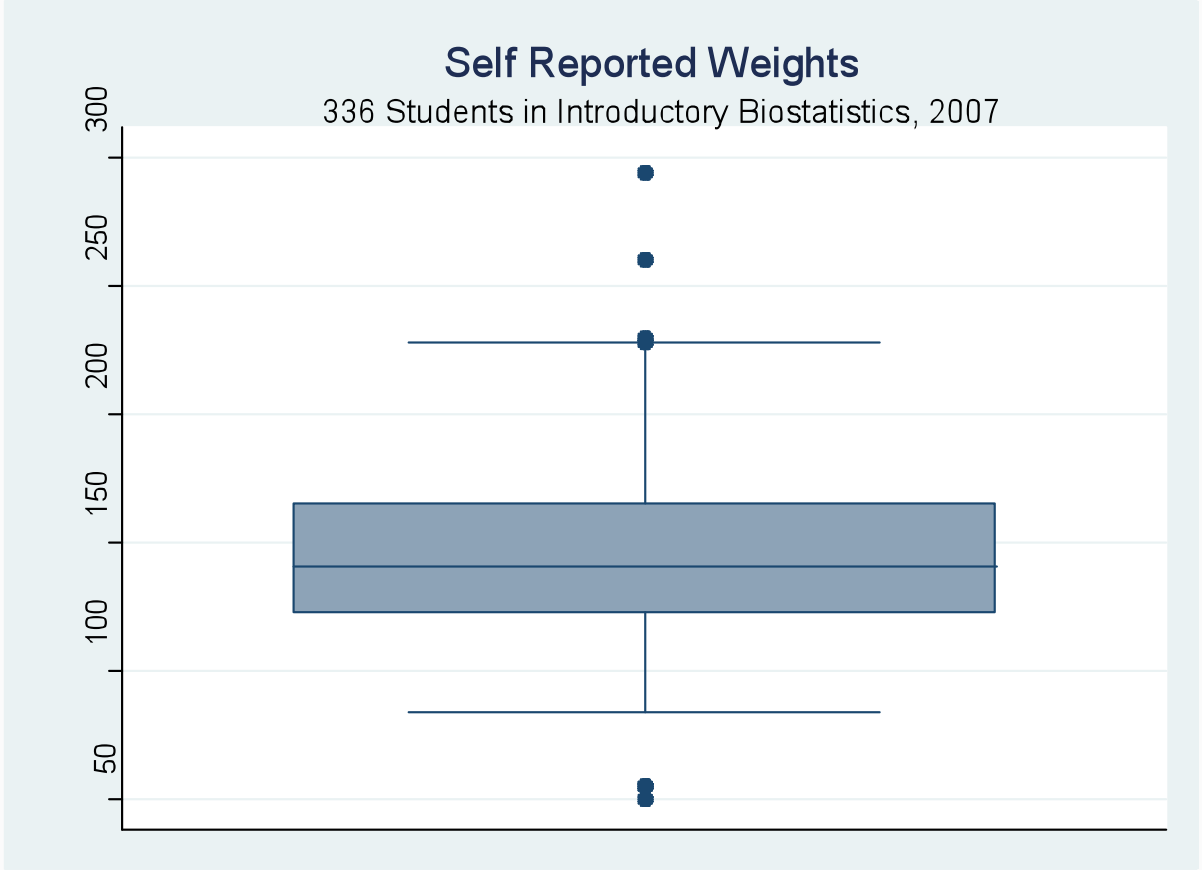
c) Suppose you take a random sample of 2,000 observations from this "population" of over 13,000 respondents. What shape will the histogram of the these sample values likely have?
   - Again, as the sample is random, it's characteristics should mimic the characteristics of the larger population of values from which it is selected
   - As such, the distribution of the 2,000 sample values will also likely be right skewed

Practice Problems

d) Which random sample, the sample of 100 or the sample of 2,000 will have a larger sample standard deviation?
   - There is no way to answer this question without knowing the sample standard deviations for each of the sample sets
   - While these sample standard deviations will likely differ in value, they are estimating the same underlying population quantity
   - There is no systematic link between the sample size and the relative magnitude of sample estimates

Practice Problems

2. The following boxplot shows the distribution of self-reported weights (in pounds) of 336 students enrolled in an introductory biostatistics course at JHBSPH in year 2007 (not 611!)
   - Mean: 145 lbs; median 141 lbs; SD 31 lbs



Practice Problems

a) How would you characterize this distribution (symmetric, right skewed, etc.)?
   - Again, a subjective question
   - My take: a relatively symmteric distribution but with a slight right skew
   - The median is slightly closer to the 25th percentile than the 75th percentile, and is smaller in value than the mean (but not by a relatively large difference as we saw with the medical expenditures data)
   - However, the outliers are balanced between small and large extremes and the smallest and largest non-outlying values are similar distance from the “box”

Practice Problems

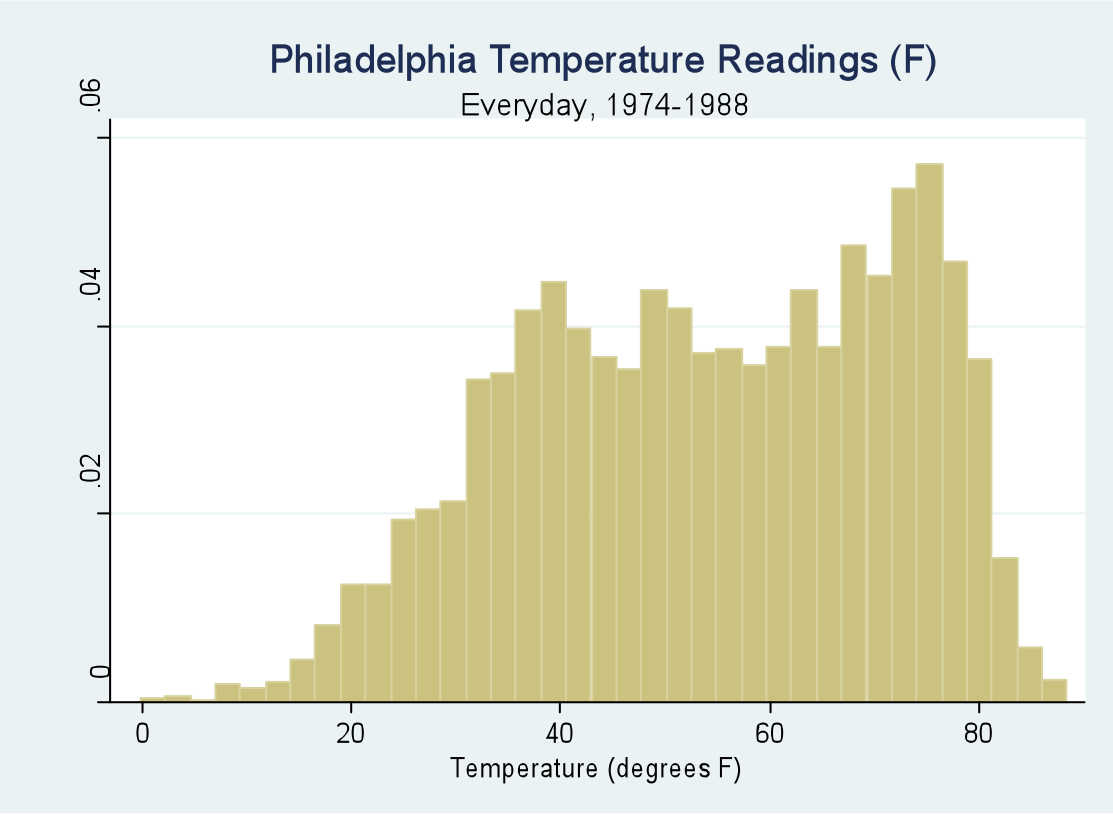
b) Suppose this sample is representative of all graduate students at JHBSPH enrolled in 2007. What does this sample suggest about the distribution shape for the weights of all graduate students in 2007 at JHBSPH?
   - If this sample is representative, then the distribution of population values should have the same characteristics as the sample distribution (i.e., relatively symmetric with a slight right skew)

Practice Problems

c) Students self selected to participate in this survey (there were more than 336 enrollees in this other introductory class). How might this impact the representativeness of the sample with regards to all JHBSPH graduate students in 2007?
   - Some possibilities:
     - Heavier (or lighter) persons most likely not to participate
     - Students enrolled in a certain degree program collectively decide not to participate
     - MPH students much more likely to participate than any other degree seeking students

Practice Problems

3. The following histogram shows the temperature measured at 12 noon on everyday of a fifteen year period for the U.S. city of Philadelphia. (5,471 days)
   - Mean: 54; median 55; SD 18



Practice Problems

a) How would you characterize this distribution? (symmetric, right skewed, etc.)
   - This distribution defies classification!
   - Well at least simple classification
   - There is a slight left skew in the histogram (and the mean is slightly lower in value than the median)
   - However, ignoring this minor left skew, the majority of the distribution is relatively uniform—however others may argue that its multi-modal