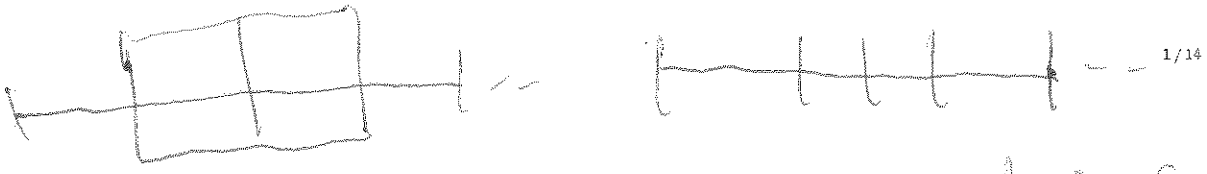
Today: 1-D Continuous

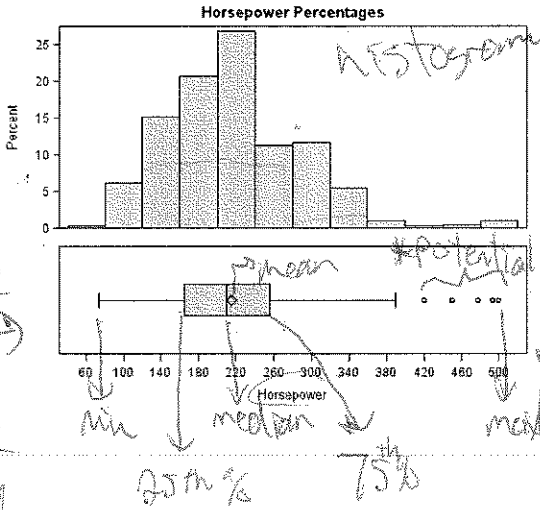
Sam Ventura
36-315
Today: Boxplots, Histograms, and Density Estimates
Conditional Distributions for Continuous Variables
Tartan Data Science Cup – Episode II → Sunday, 10/9

Department of Statistics
Carnegie Mellon University

September 26, 2016



Box Plots vs. Histograms: Advantages and Disadvantages



Boxplots advantages:
- can easily see skew
- get a sense of the center and the spread of the distribution
- get some specific values

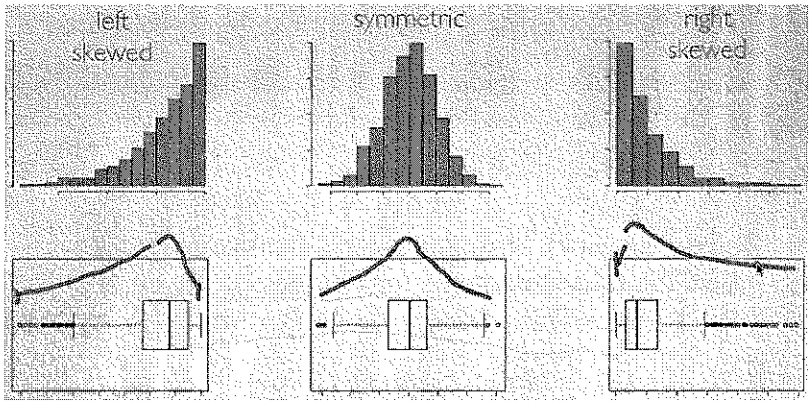
Disadvantages:
- no shape
- modality is difficult to discern
- no idea about sample size

Histograms
Adv: Area actually represents data
- can better see skew
- sample size can be estimated
- frequencies w/in each bin
- can usually see modality

Disadvantages: we have to choose # of bins or the bin width

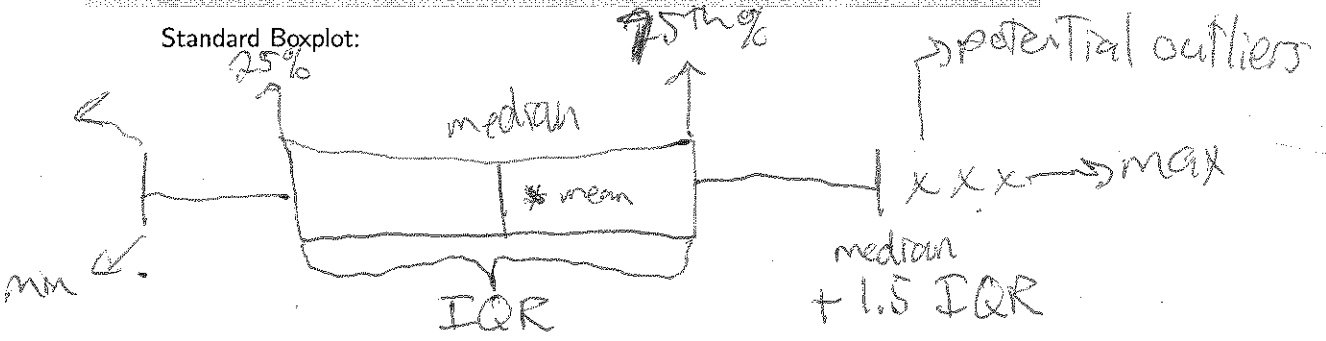
- hard to distinguish between very different distributions

Determining Skew Visually



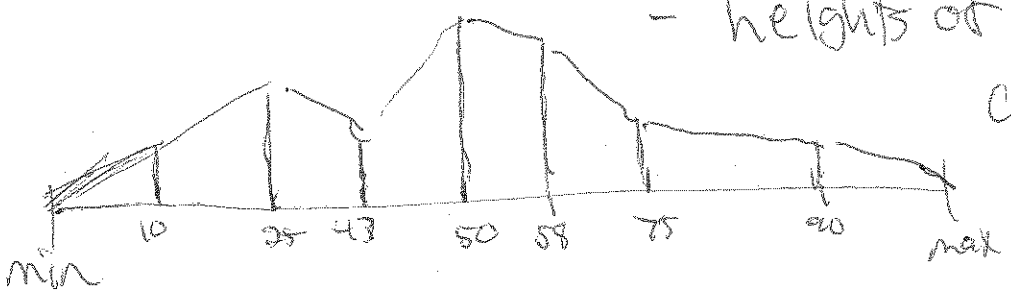
How Can We Improve Boxplots?

Standard Boxplot:



Improved Boxplot?

same thing, but w/ more %-iles
- heights of lines ∝ "density" of observations in that region



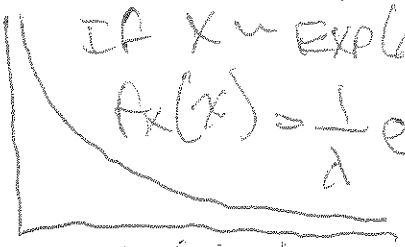
Continuous Densities

"smoothed out function of the data's (variable's) distribution"

Theoretical: X is a random variable, $f_X(x)$ → "probability density function"

If $X \sim N(\mu, \sigma^2)$

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{\sigma^2}}$$



If $X \sim EXP(\lambda)$

$$f_X(x) = \frac{1}{\lambda} e^{-x/\lambda}$$

What if we don't know the underlying distribuion? → "true distribution" → theoretical

How can we estimate the empirical distribution?

↳ based on the data

1-D Kernel Density Estimation