Today: 1-D Continuous

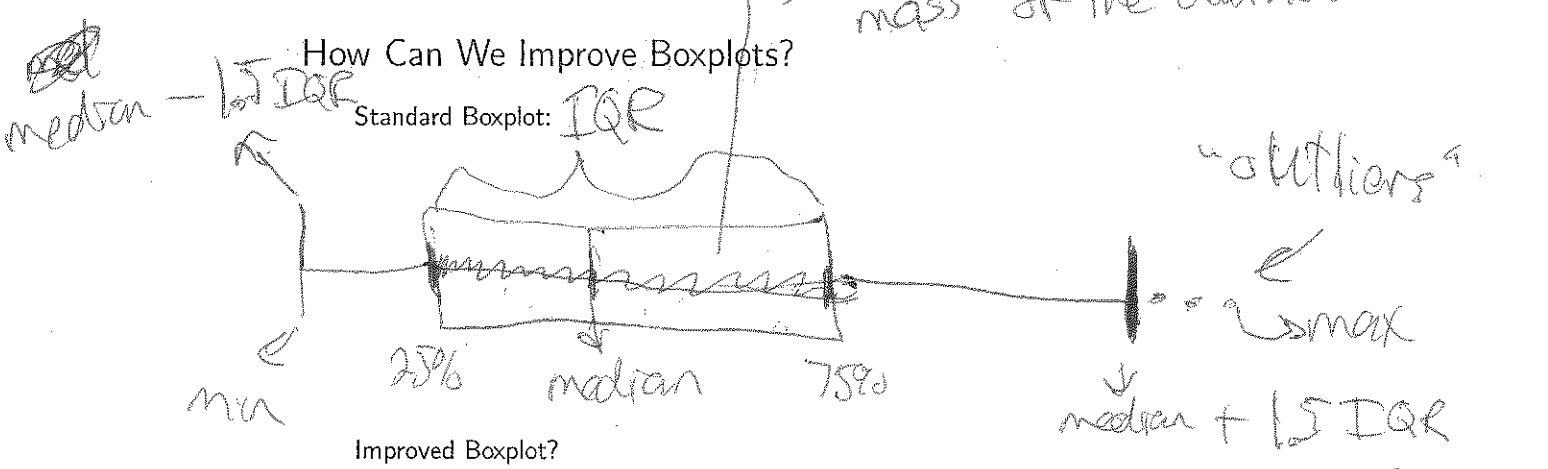
Sam Ventura
36-315
Today: Density Estimates
Box-Percentile Plots

Department of Statistics
Carnegie Mellon University

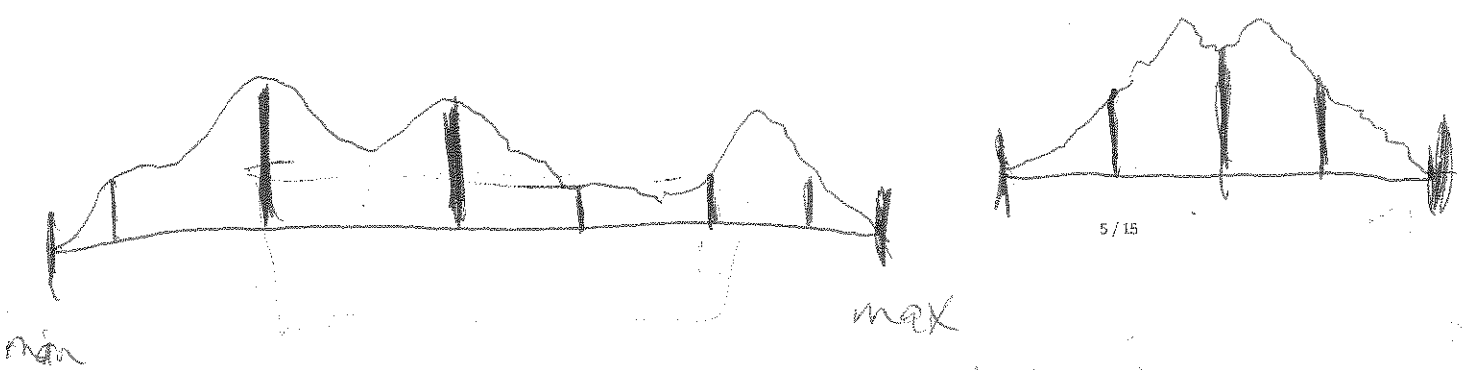
February 17, 2016

How Can We Improve Boxplots?

Standard Boxplot:



Improved Boxplot?



"probability density function"

Continuous Densities → Smoothed-out function of the data's distribution

Theoretical: X is RV, $f_X(x)$ e.g. $f_X(x=1)$

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

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

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

How can we estimate the empirical distribution? → "observed" → based on the data

"^" "estimate" → so $\hat{f}$ means "our estimate of f, the continuous density"

parametric statistics: making assumption about the underlying distribution

1-D Kernel Density Estimation → Non-parametric → No assumptions → based on the data.

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right)$$

- n = # of observations in dataset
- h = "bandwidth" → "bin size" / "bin width" → this is a parameter that we get to choose. It dictates "smoothness" vs. "roughness"/"jaggedness" of the density estimate.
  - K contributes less to shape of the density estimate.
  - Small h: bumpy densities
  - large h: smoother densities, K contributes more to shape of density estimate
- K = "kernel function" → we get to choose this as well (parameter) → different functions give us different features of our resulting density estimate.
  - Ex) Boxcar kernel, Gaussian kernel, uniform kernel, etc.
- x: the point at which we are estimating the density; should be w/in the range of the continuous variable