

Continuous Variables and their Distributions

Rebecca Nugent

Department of Statistics, Carnegie Mellon University

<http://www.stat.cmu.edu/~rnugent/PCMI2016>

PCMI Undergraduate Summer School 2016

July 2, 2016

Continuous Distributions

$$x = \{x_1, x_2, \dots, x_n\} \in \mathfrak{R}, n \text{ observations}$$

If we wanted to see all of the observations, options are

- ▶ dot chart
- ▶ stripchart (like a one-dim scatterplot)
- ▶ stem-and-leaf

Often can't possibly visualize all the observations;
interested in their overall *distribution*

What kind of features/structure would we be interested in?

Histograms

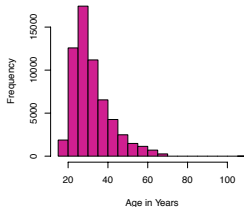
- [illegible]

Boxplots

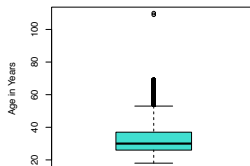
- ▶ Common picture of a five number summary: 25th perc, median, 75th perc, upper/lower hinges
- ▶ Can get idea of possible outliers
- ▶ Advantages? Disadvantages?

Histograms and Boxplots

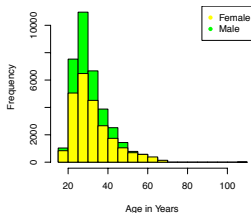
Distribution of Age



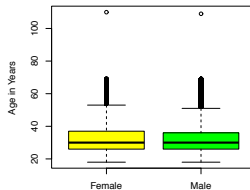
Distribution of Age



Distribution of Age conditioned on Gender



Distribution of Age conditioned on Gender



Estimating the Distribution

If underlying distribution of data is known (normal, uniform, exponential, etc), just need to estimate the parameters

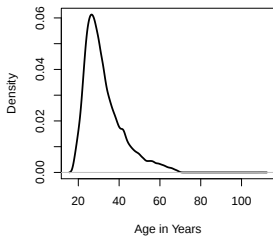
In real life, we never really know anything. #statprobs #lifeprobs
Have to estimate instead

Kernel Density Estimate

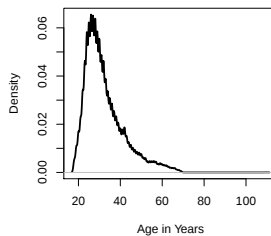
- ▶ Nonparametric, no assumptions
- ▶ Requires choice of kernel
- ▶ Requires bandwidth choice

Kernel Density Estimates

Distribution of Age



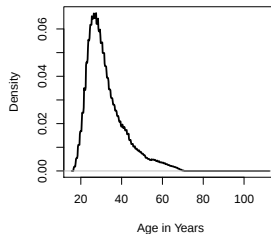
Distribution of Age, BW = 0.4



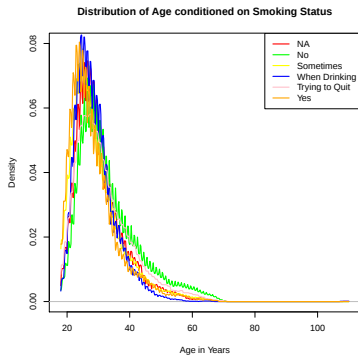
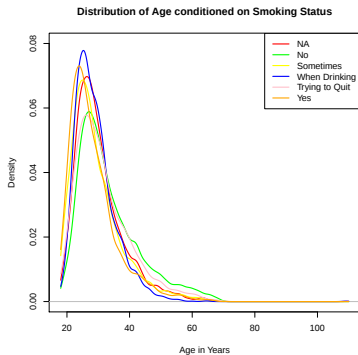
Distribution of Age, BW = 3



Distribution of Age, Kernel = Rectangle



KDE: Age Conditioned on Smoking Status



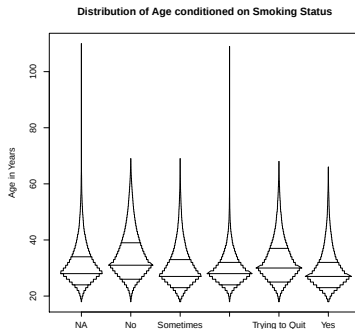
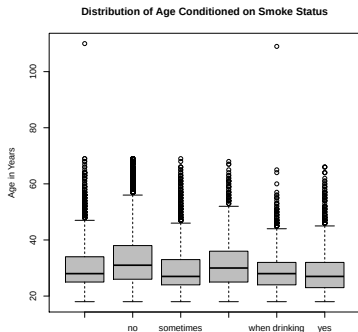
Box-Percentile Plots

Combining quartiles w/ features of a cumulative distr fxn

Width is prop to probability of seeing that value or more extreme

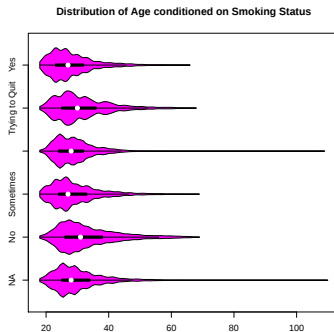
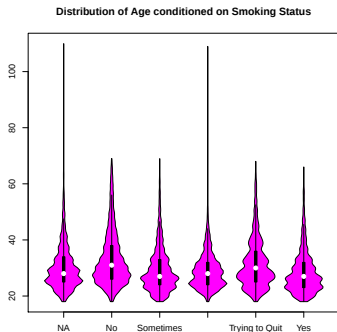
No change in width = no observations

Sudden width changes = groups of points, high freq area



Violin Plots

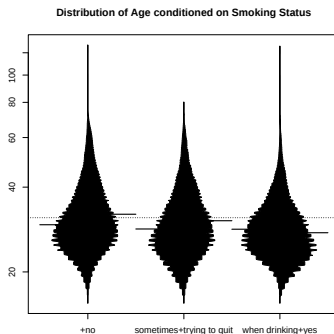
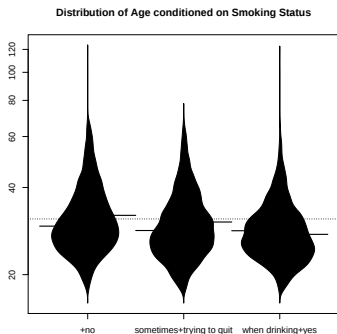
Combining kernel density estimates with boxplots



Bean Plots

Often used for comparing densities of subgroups; includes:

- ▶ Overall average value
- ▶ Subgroup average values
- ▶ Can include the actual values as well



Have to choose kernel, bandwidth for density estimate

Conditional Density Plot

How does conditional distribution of categorical variable change over a continuous variable? *flipping things around a little*

