Continuous Variables and their Distributions: 2-D

Rebecca Nugent
Department of Statistics, Carnegie Mellon University
http://www.stat.cmu.edu/∼rnugent/PCMI2016

PCMI Undergraduate Summer School 2016

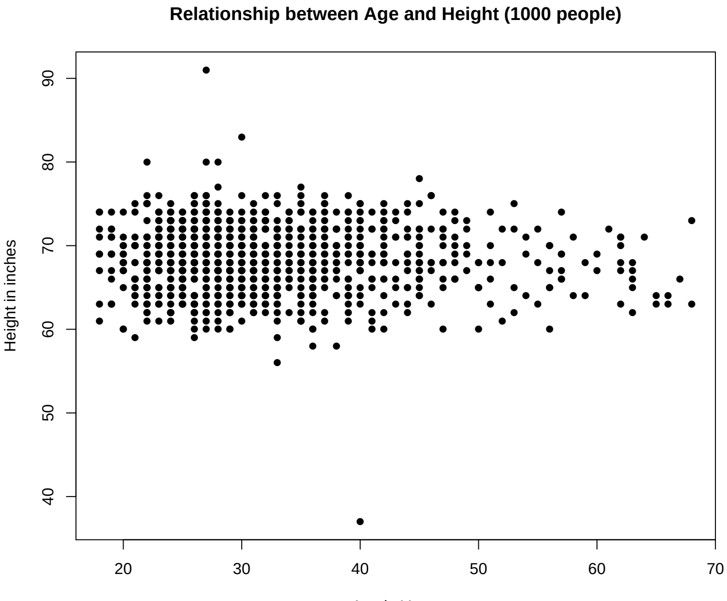
July 5, 2016

What did we think about last time?

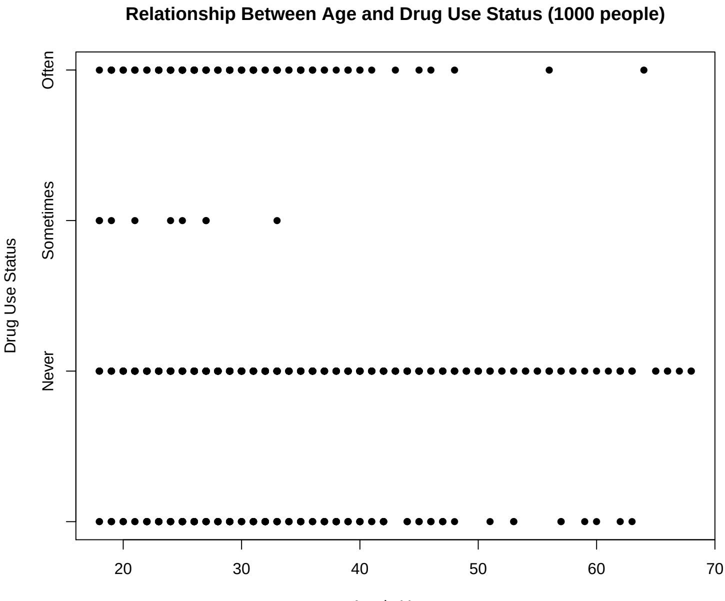
- Distribution of a Continuous Variable
    - Mean, St Dev, Median, Range, IQR
    - Skew, (A)symmetry, Modality, Tails, Outliers
- Histograms: bin width parameter (large, global; small, local); default rules often assume normal data
- Boxplots: can find numerical summaries, miss modality/shape information; many distributions look the same
- Kernel Density Estimates
    - Nonparametric, no assumptions about origin of data
    - Need to choose kernel shape (Gaussian, rect, tri, etc) and bandwidth (large, global; small, local)
    - Can be expensive
- Box-Percentile Plot, Violin Plot, Bean Plot, Conditional Density Plot (for categ var given cont var)

Now thinking about the structure of relationships between variables

Scatterplot: Age vs Height

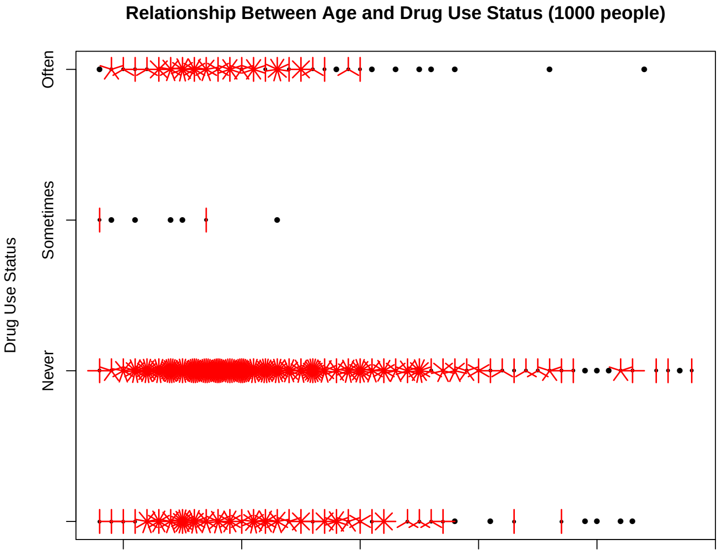


Scatterplot: Age vs Drugs

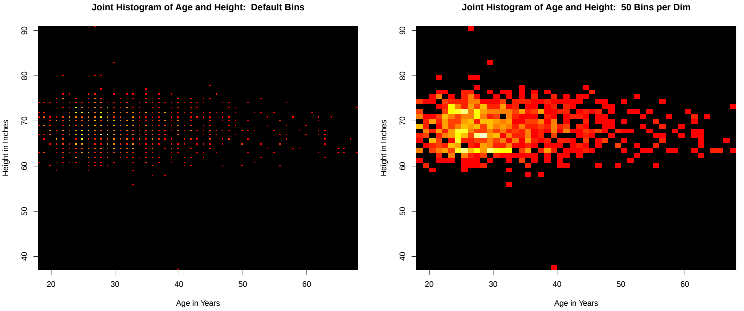


Sunflower Plot: Age vs Drugs

For use with duplicates



Joint Distribution: 2-D Histogram



2-D Boxplots are possible too (have some interpretation issues); check out 2D Bagplots as well

2-Dim Kernel Density Estimate

Reminder of our KDE for one variable: $\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right)$

- Kernel shape dictates contribution to estimate; infinite support (Gaussian), compact support (rect, tri, biweight, triweight); efficiency/theory vs computational issues
- Bandwidth (size of kernel) also dictates contribution but more about smoothness (large, global; small, local)

Moving to 2-Dimension KDE:

$$\hat{f}(\underline{x}) = \frac{1}{nh_1 h_2}\sum_{i=1}^{n} K\left(\frac{x_1 - x_{i1}}{h_1}\right) K\left(\frac{x_2 - x_{i2}}{h_2}\right)$$

- Above treated as product of two kernels; can do multivariate kernel as well
- the K is the same for both dimensions; bandwidths can/should be different

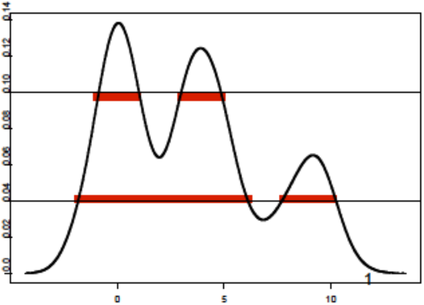
Level Sets of a Density Estimate

Some common structure questions:

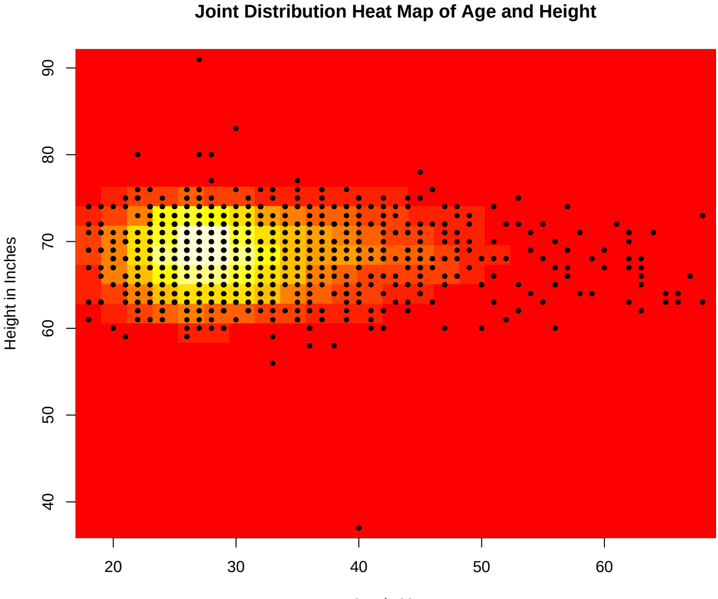
- Where are the high frequency areas? **modes**
- Where are the areas with no data? **valleys**

Can look at/analyze the cross-sections or **level sets** of the density

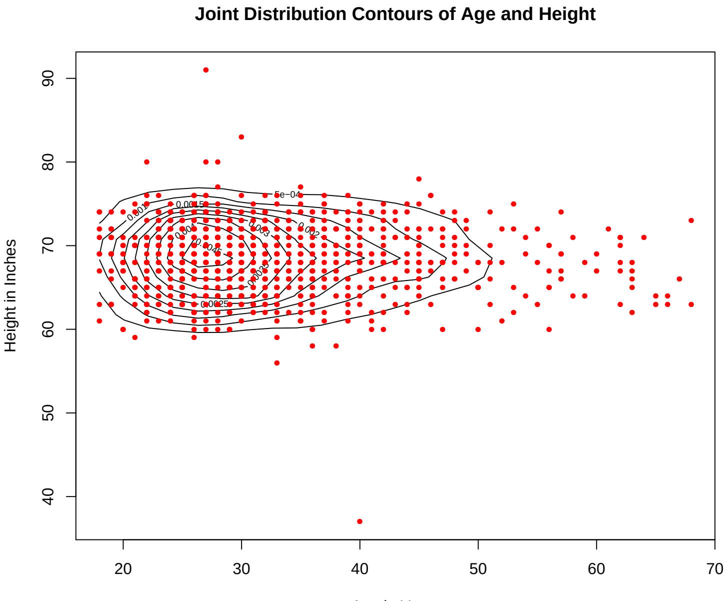
$$L(\lambda; f(x)) = \{x | f(x) > \lambda\}$$



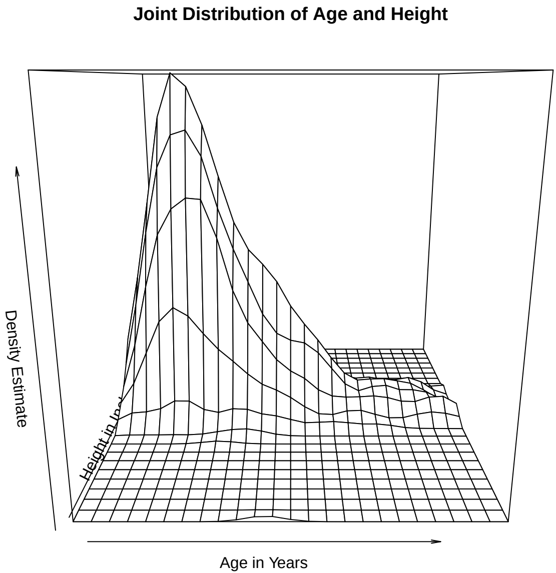
KDE Heat Map: Age vs Height



KDE Contours/Level Sets: Age vs Height



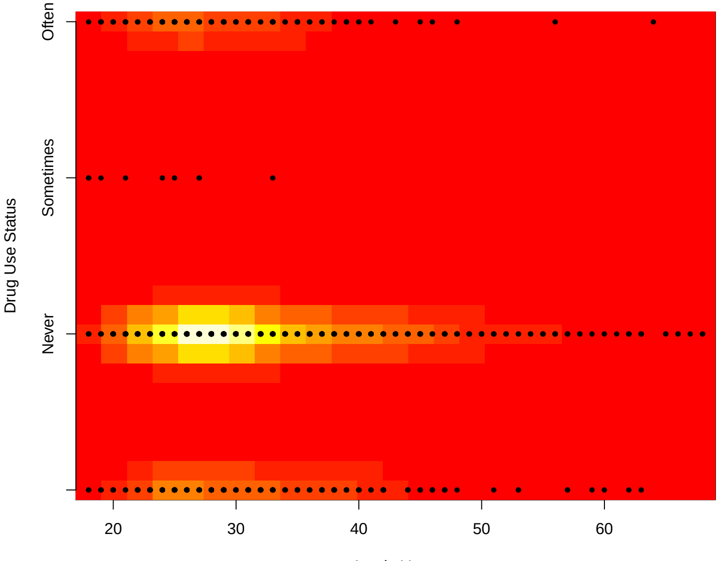
KDE Perspective: Age vs Height



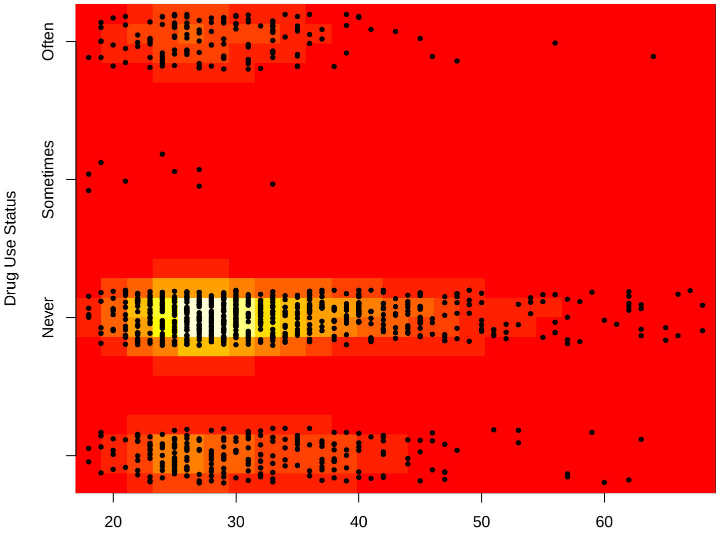
KDE Rotating Perspective: Age vs Height

Often the non-dynamic perspective plot obscures features since much of one dimension ends up “hidden”

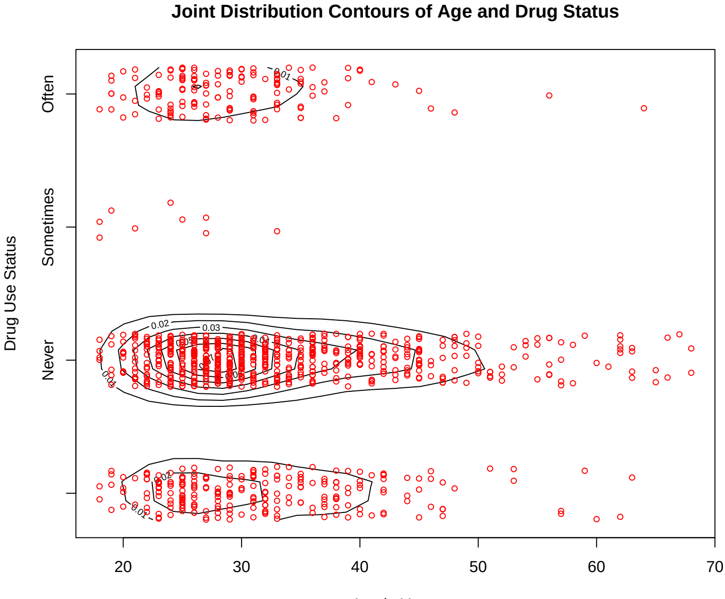
KDE Heat Map: Age vs Drugs



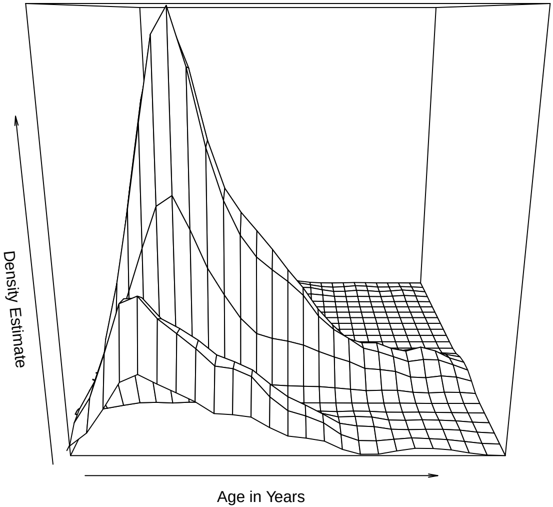
KDE Heat Map: Age vs Drugs (Jittered)



KDE Contours/Level Sets: Age vs Drugs



KDE Perspective: Age vs Drugs



KDE Rotating Perspective: Age vs Drugs

In summary: What did we think about?

- Relationships between Variables
- What do we do with duplicated observations/values?
- Looking at Joint Distributions piecewise constant
- 2-D KDE: Kernels, Bandwidths, Computational Issues
- What happens if one variable (or both) is ordinal/categorical?
- High and low frequency areas; level sets, contours
- Visualizing matrices