..count.. } objects returned by ggplot when you use geom_hist(), geom_bar(), geom_density()

↓

telling R to look for variables, then use them in the plot

..count..

Today: 1-D Continuous

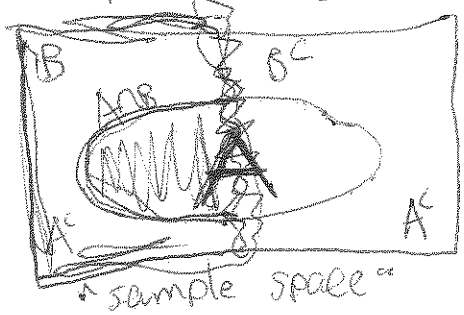
Sam Ventura
36-315
Today: Density Estimates, Kernels, Violin Plots
Rugs, Conditional Distributions, KS Tests

Department of Statistics
Carnegie Mellon University

February 22, 2016

Conditional Distributions

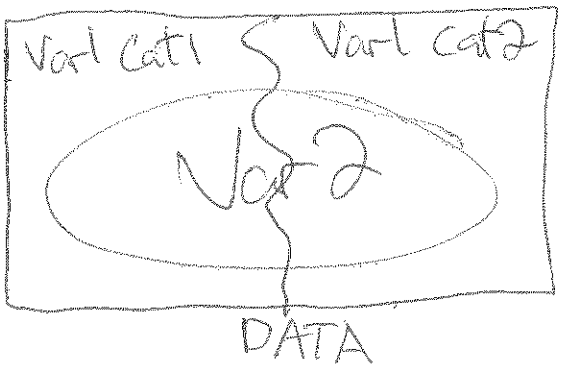
In probability: Ω



$$P(A) = \frac{\text{Area of } A}{\text{Area of } \Omega} \quad \Omega = \text{box}$$

$$P(A|B) = \frac{\text{Area of } A \cap B}{\text{Area of } B}$$

So, in R, get subsets w/ facets

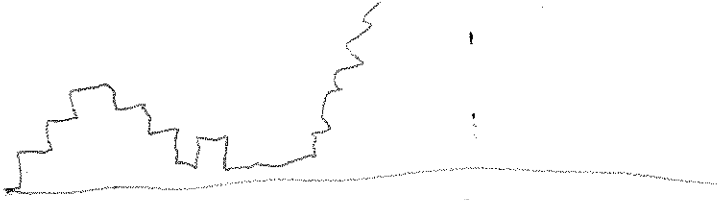


Var1 is categorical w/ 2 cats.

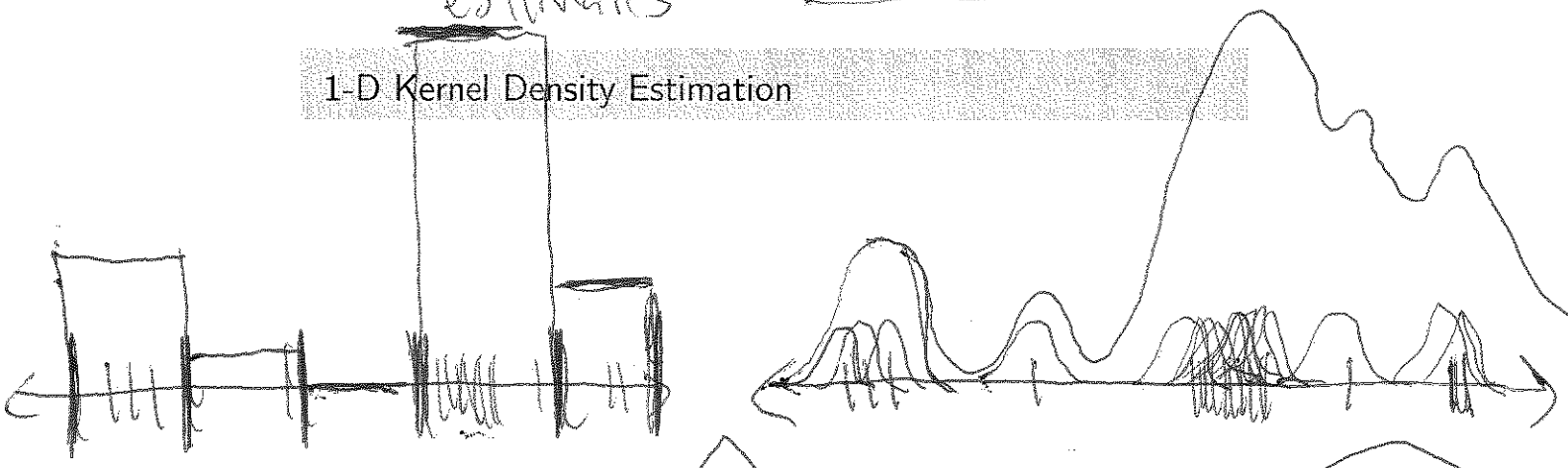
Var2 is continuous

Cond. distn of Var2 | Var1, cat1
↳ taking the subset of data corresponding to Var1 cat1
↳ find distn of Var2 w/in that subset.

Histograms ↪ density estimates



1-D Kernel Density Estimation



Kernels:

Gaussian/normal — smooth, default
Bad:

triangular — most of the mass/area centered on each obs

uniform/boxcar — step function[3][13]

epanechnikov — smooth-ish

Good when we have a distn w/ fixed end points

Gamma, exponential, Beta

KDE → non-parametric estimate of a density
↪ no assumptions about underlying distn
↪ Data was generated randomly and is representative of underlying distn

↳ BASED ON THE DATA