Today: 1-D Continuous

Sam Ventura
36-315
Today: Boxplots, Histograms, and Density Estimates
Conditional Distributions for Continuous Variables

Department of Statistics
Carnegie Mellon University

February 13, 2017

Time, Age, Temperature, height, weight, 1/19
probabilities / percentages, rates, Money, distance

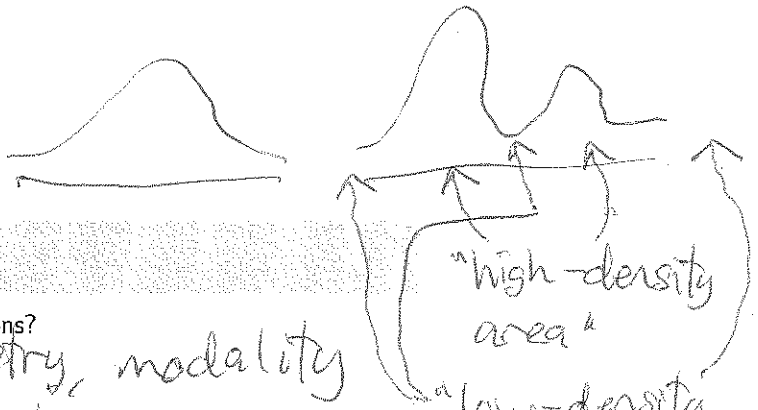
1-D Continuous Data

Structure: $X = \{X_1, X_2, \cdots X_n\}$, $X_i \in \mathbb{R}$

Summary:

mean / average, median → center
weighted average, mode
standard deviation, variance → spread
range (max − min), IQR (~~75% rule~~ per
~~In R:~~
maximum, minimum, percentiles, quantiles
skew / symmetry, "outliers" 2/19

mean(), median(), range(), etc



1-D Continuous Distributions

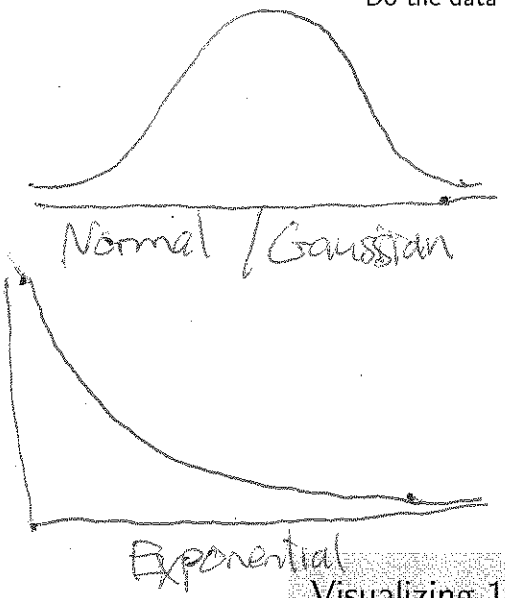
How do we describe continuous distributions?

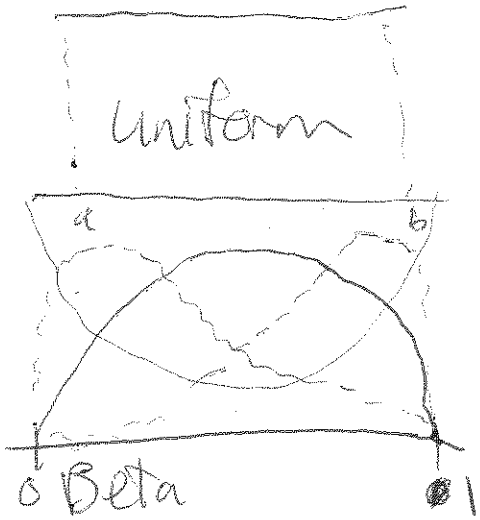
Shape → skew, symmetry, modality

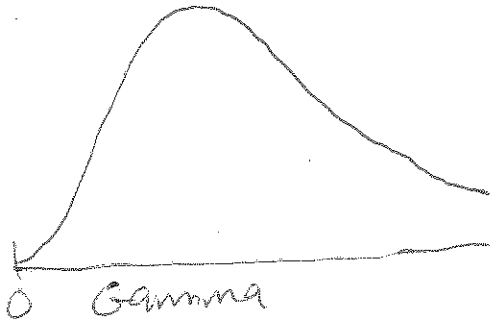
center → mean, median, locations of modes

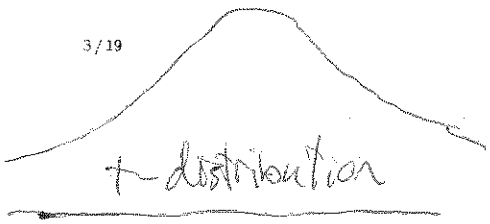
spread → SD, Var, range

Do the data appear to fit some common/known distribution:









3/19

Visualizing 1-D Continuous Distributions

How do we visualize continuous distributions?

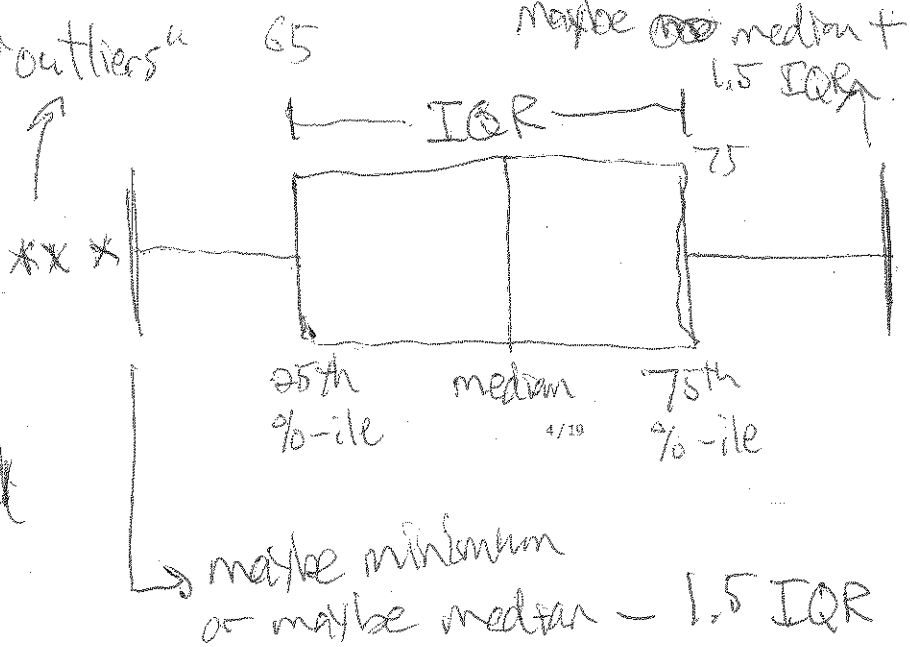
Histogram: Break continuous variable into "bins", then we count # of obs. in each bin.

Boxplots

maybe Mx

max.

maybe ~~Mx~~ median + 1.5 IQR



4/19

Boxplots:

Adv) - "rough" estimate of "outlier" locations

- fairly easy to compare boxplots across conditional distributions
- easy to see spread, skew, location of median, obtain %-iles

Disadvantages) Distortion!

- we can't see shape of the distribution (unimodal vs. bimodal)
- modality is impossible to discern
- Can't see mean
- no idea about sample size

Histograms

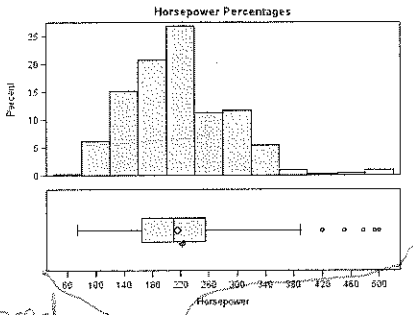
Adv) Easy to see shape, skew, modality, range,

- No (or very limited) distortion
- Area actually represents data
- Sample size is easy
- we can add lines to mark mean, median, etc. → lines

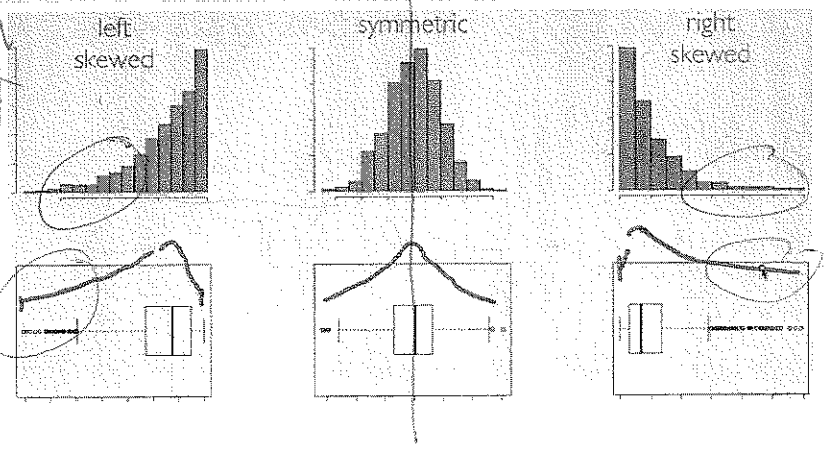
Disadvantages

- We have to choose a parameter that may affect the way the graph is shown
  - ↳ bin width, # of bins
- Bad choices yield bad graphs

Box Plots vs. Histograms: Advantages and Disadvantages

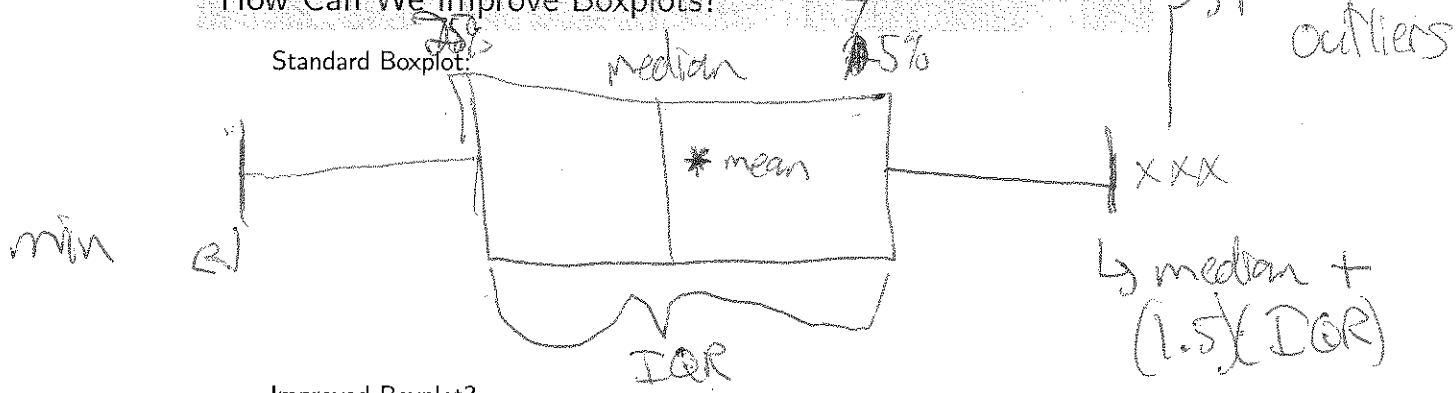


Determining Skew Visually



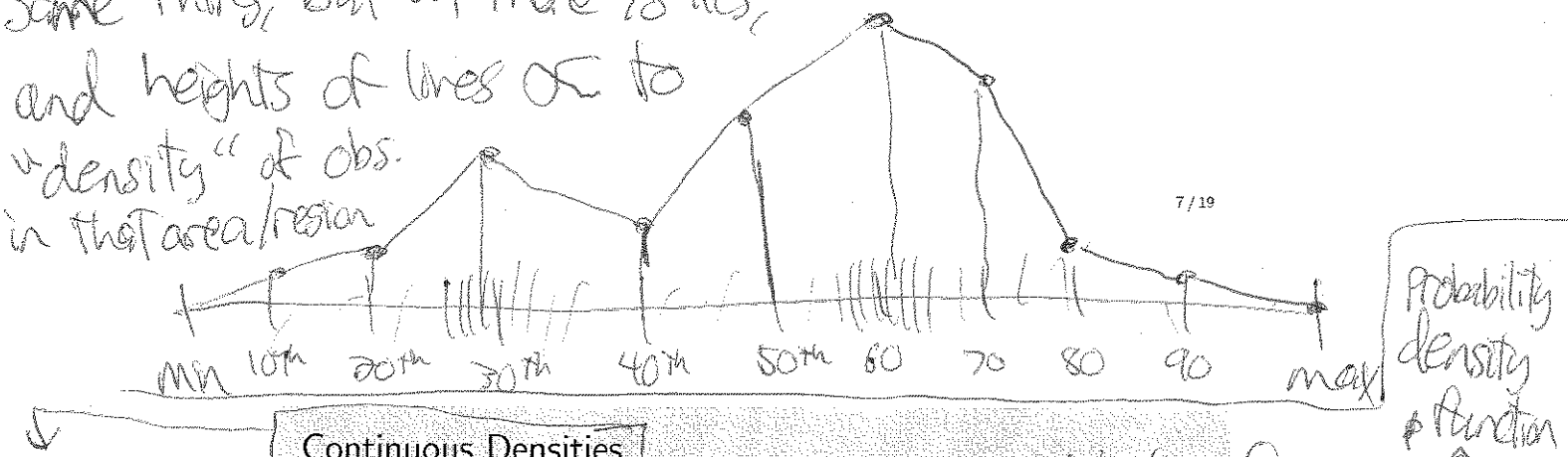
How Can We Improve Boxplots?

Standard Boxplot:



Improved Boxplot?

Same thing, but w/ more %-iles, and heights of lines etc to "density" of obs. in that area/region

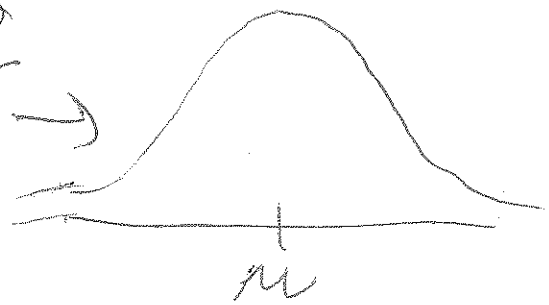


Continuous Densities

Theoretical: X is a "random variable", $f_X(\cdot)$

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

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



What if we don't know the underlying distribuion?

How can we estimate the empirical distribution?

parametric statistics: making assumptions about the underlying distribution of the data, and then finding the optimal parameters that match data + assumptions

Non-parametric statistics

- No assumptions (or at least very limited assumptions)
- 100% based on data

$\hat{}$ = estimate

$\hat{f}_x$ = estimated probability density function

→ Goal: estimate $f_x(x)$ given data, no assumptions

1-D Kernel Density Estimation

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

- n = sample size / # of observations
- h = "bandwidth" "bin width in histogram"
  - ↳ this is a parameter that you choose
  - ↳ it dictates the "smoothness" vs. "rigidness" of the resulting density estimate

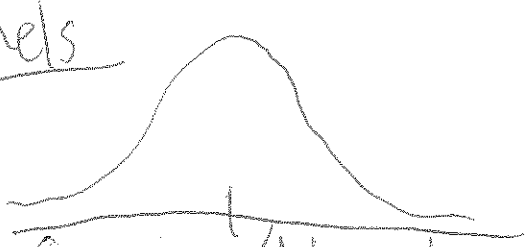
Small h: rigid density estimate

Large h: smooth density estimate

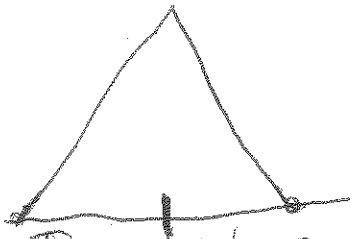
- $K()$ = "kernel function" → we get to choose this as well (parameter)
  - ↳ contributes to the shape of the density estimate
  - ↳ different kernel functions will give us different features in the density estimate

x: the point at which we are ~~evaluating~~ estimating the density
- ↳ should be within the range of our observed data (approximately)

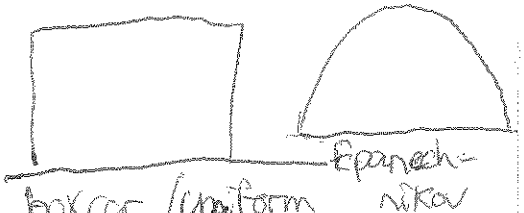
Kernels



Gaussian / Normal
- smooth, default



Triangular
- most of the mass is above the point itself
- fixed endpoints



boxcar / uniform

Epanechnikov
- DEs will be a step function
- smooth
- fixed endpoints

↳ Good when you have fixed endpoints in data (time, etc)

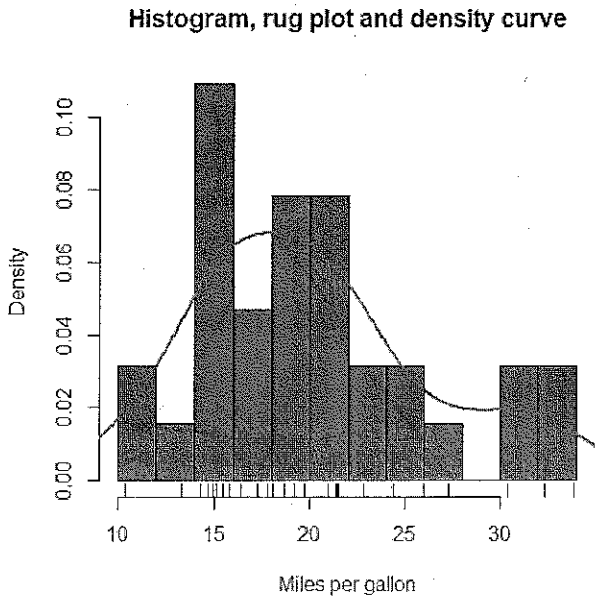
Comparing 1-D Continuous Densities

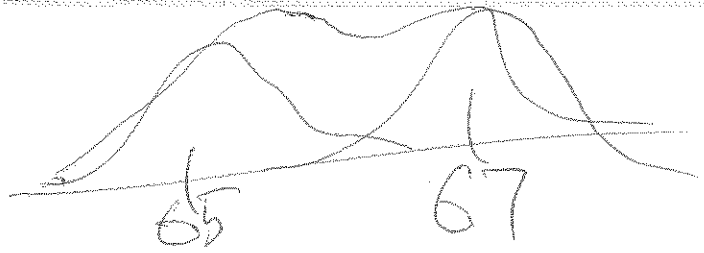
Kolmogorov-Smirnoff (KS) Test:
nonparametric test of the equality of
multiple 1-D continuous distributions

In R: **ks.test**

Kullback-Liebler Divergence:
Measure of the difference between two probability distributions
(discrete/categorical or continuous)

Histogram with Density Curve and Rug





1-D Kernel Density Estimation

Disadvantages

- shape of the DE depends on your choice of kernel and bandwidth.
- NO sample size in DE (some less-than-ideal ways around this)

Advantages

- SMOOTH
- put some mass between observed points
- we can get specific estimates of the probability density function at any point on the x-axis.
- Easy to see modality

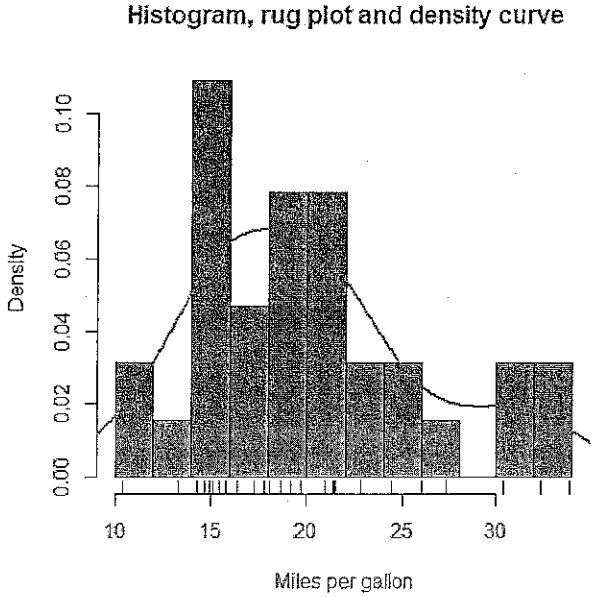
Comparing 1-D Continuous Densities

Kolmogorov-Smirnoff (KS) Test:
nonparametric test of the equality of
multiple 1-D continuous distributions

In R: **ks.test**

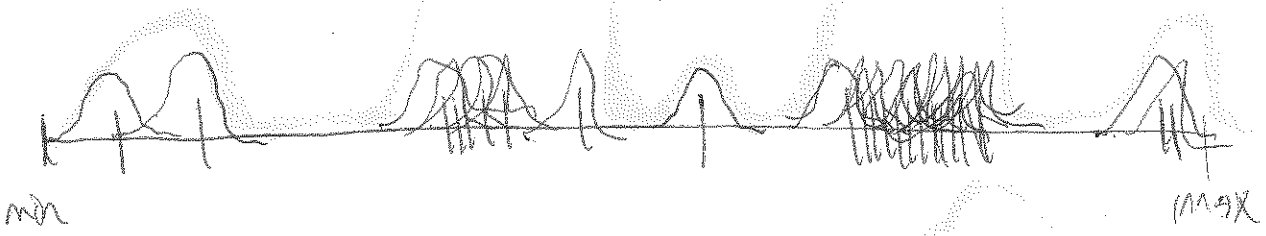
Kullback-Liebler Divergence:
Measure of the difference between two probability distributions
(discrete/categorical or continuous)

Histogram with Density Curve and Rug

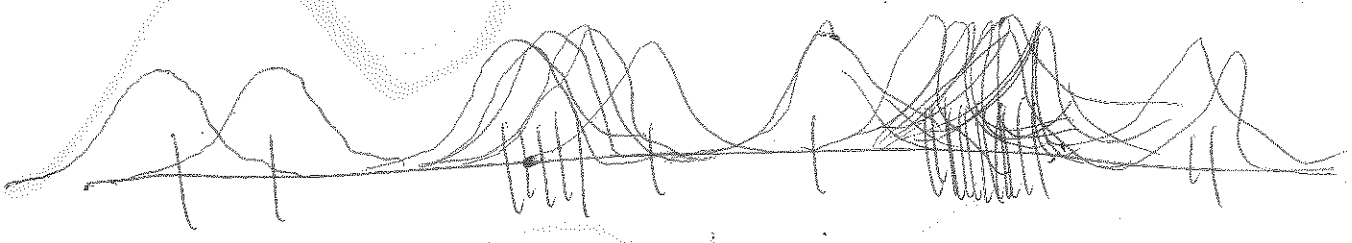


1-D Kernel Density Estimation

small bw



medium bw



large bw

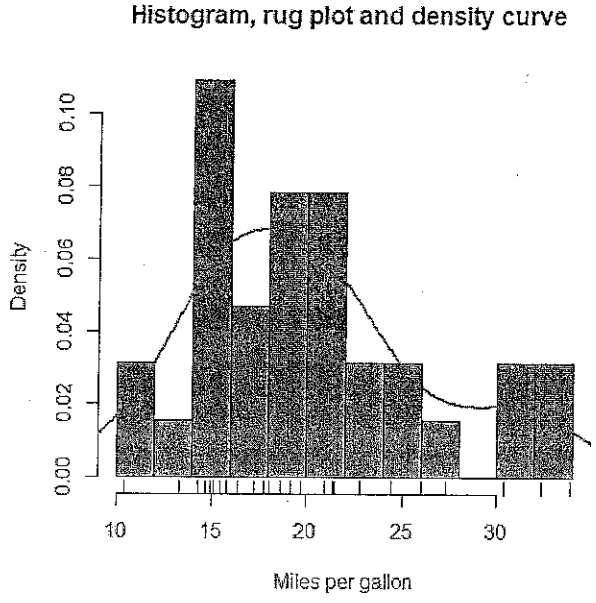
Comparing 1-D Continuous Densities

Kolmogorov-Smirnoff (KS) Test:
nonparametric test of the equality of
multiple 1-D continuous distributions

In R: **ks.test**

Kullback-Liebler Divergence:
Measure of the difference between two probability distributions
(discrete/categorical or continuous)

Histogram with Density Curve and Rug



1-D Continuous Distributions

How do we describe continuous distributions?

Do the data appear to fit some common/known distribution:

Visualizing 1-D Continuous Distributions

How do we visualize continuous distributions?