PROBABILITY OF CONTINUOUS VARIABLE

SYRACUSE UNIVERSITY
School of Information Studies

HOW TO ESTIMATE PROBABILITIES OF CONTINUOUS ATTRIBUTES

Two approaches for continuous attributes:

Discretization

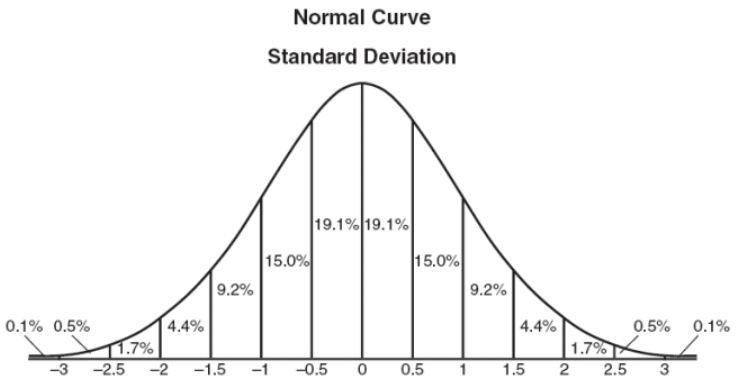
(0, 60k), (60k, 100k), (100k, ...)

Probability density estimation

Assume attribute follows a normal distribution.

Use data to estimate parameters of distribution (e.g., mean and standard deviation).

Once probability distribution is known, can use the probability density function to estimate the conditional probability $P(A_i|c)$.



HOW TO ESTIMATE PROBABILITIES OF CONTINUOUS ATTRIBUTES

Tid	Refund	Marital Status	Taxable Income	Evade
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Normal distribution:

$$P(A_i \mid c_j) = \frac{1}{\sqrt{2}\ _{ij}^{2}} e^{\frac{(A_i\ \ _{ij})^2}{2\ _{ij}^{2}}}$$

One for each (A_i, c_i) pair

For (Income, Class = No):

If Class = No

Sample mean = 110

Sample variance = 2,975

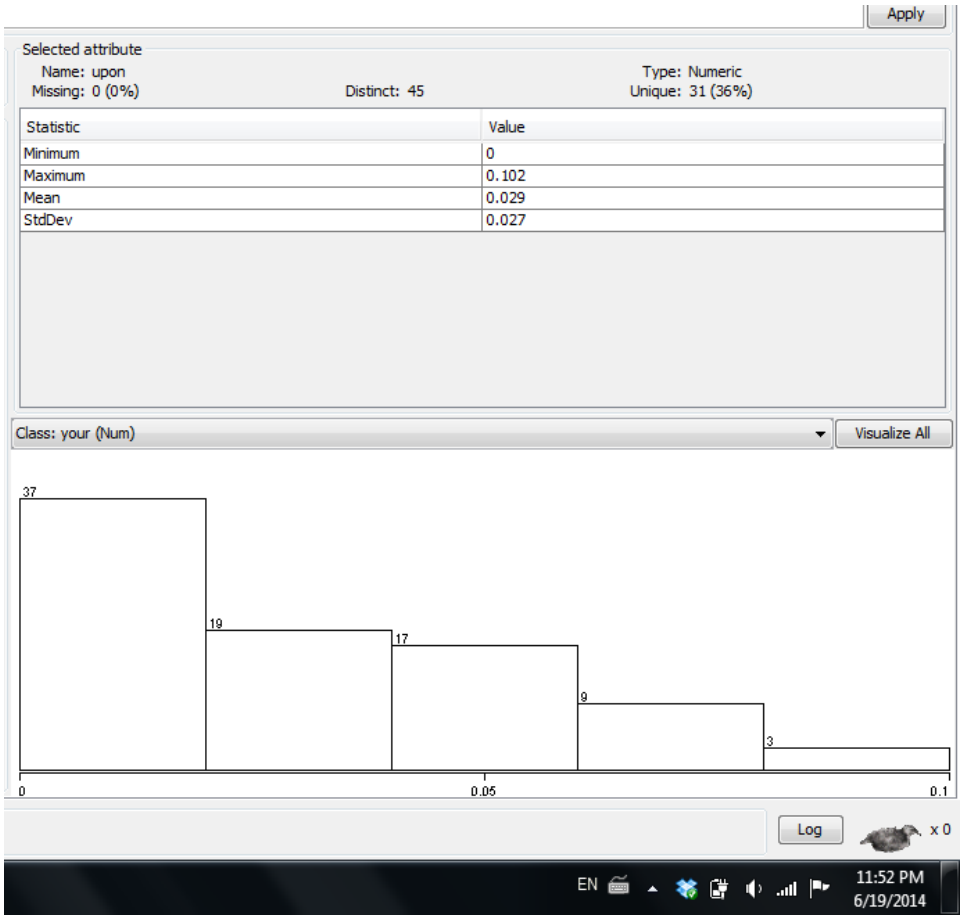
$$P(Income = 120 \mid No) = \frac{1}{\sqrt{2}\ (54.54)} e^{\frac{(120\ \ 110)^2}{2(2975)}} = 0.0072$$

HOW DO I KNOW IF A VARIABLE FOLLOWS NORMAL DISTRIBUTION?

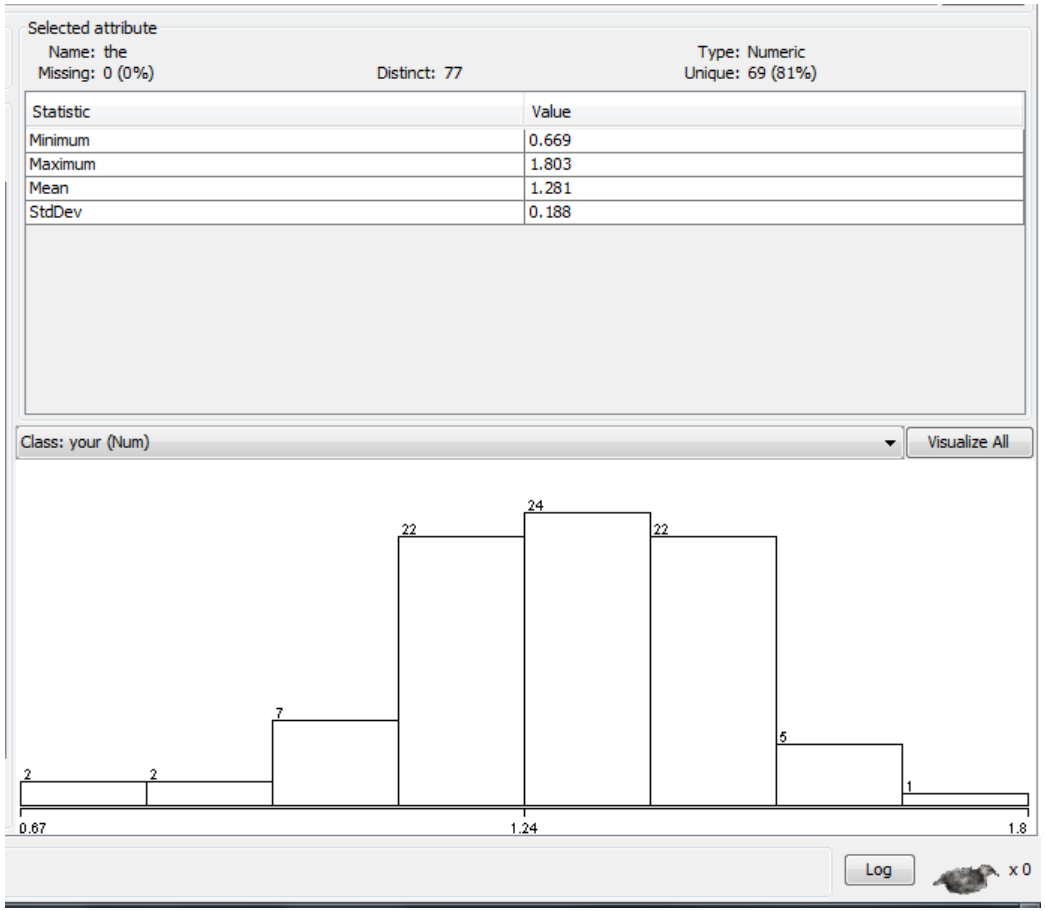
Approach 1: Use statistics test.

Approach 2 (recommended to our class): Use visualization. (Does it look like a bell curve?)

A VARIABLE THAT SEEMS NOT TO FOLLOW NORMAL DISTRIBUTION



A VARIABLE THAT SEEMS TO FOLLOW NORMAL DISTRIBUTION



NORMAL VS. KERNEL DENSITY ESTIMATION

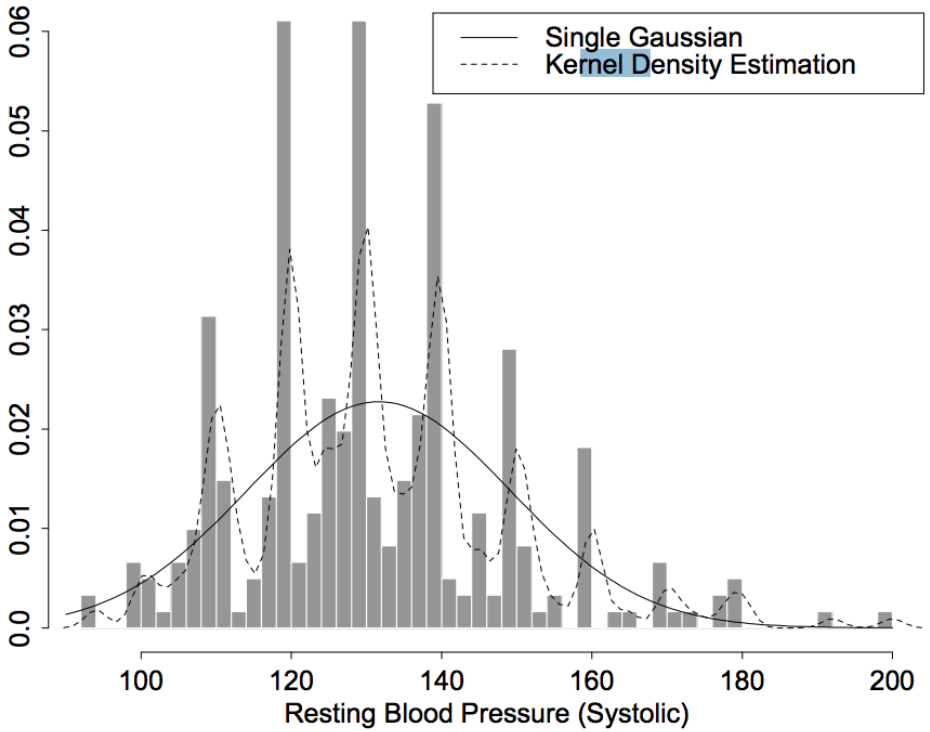


Figure 3: Systematic measurement errors in the Cleveland heart disease database.

Normal distribution and Gaussian distribution are two names for the same thing.

http://www.cs.iastate.edu/~honavar/bayes-continuous.pdf