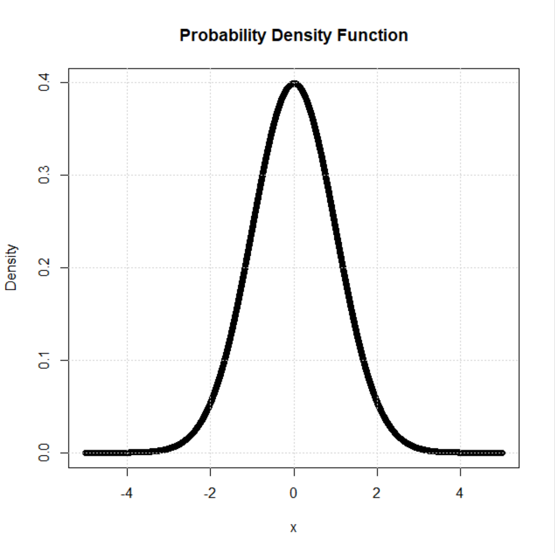
Machine Learning for CE
Probability Distributions

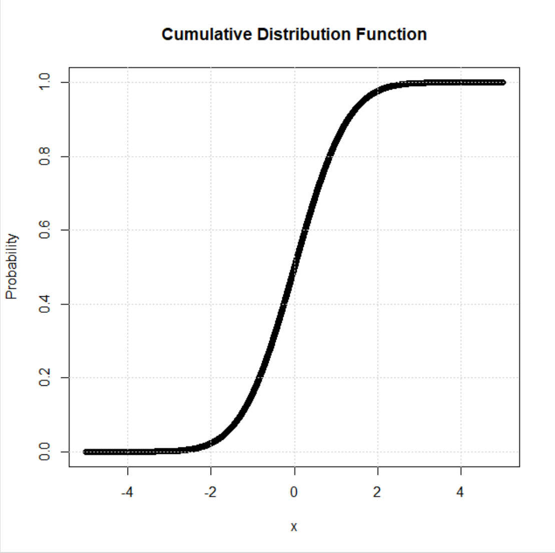
CE 5331 MACHINE LEARNING FOR CIVIL ENGINEERS

Venki Uddameri, Ph.D. , P.E.

Probability Conceptualizations

- There are two probability conceptualizations
  - Probability Density Function (PDF)
    - Measures the relative likelihood of a RV to take an assumed value
    - Can be used to compute probability that a RV will fall between two values
  - Cumulative Probability Distribution Function (CDF)
    - Measures the cumulative probability that a RV has a value equal to or lower than a specified value
    - The Y-axis takes values between 0 and 1

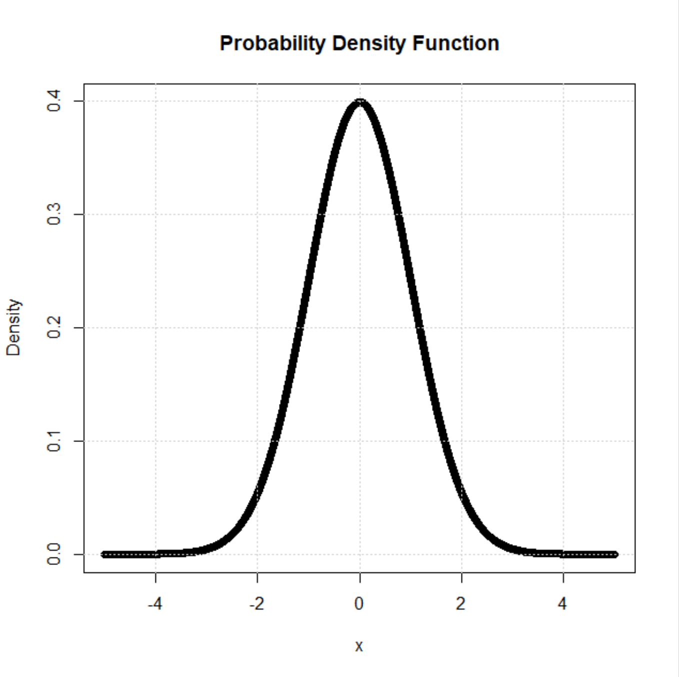




For countably finite discrete variables we can directly plot probability for each value of RV this is called the probability mass function (PMF)

Probability Density Function (PDF)

- PDF measures the relative likelihood of a random variable to take a given value
  - Unlike probability, the Probability density can assume values greater than 1
  - The area under the PDF is ALWAYS equal to 1
- PDF is usually denoted as f(x) (lower-case letters)
  - Probabilities are typically denoted by upper case



$$P\,(X \geq x_a \,\&\, X < x_b) \;=\; \int_{x_a}^{x_b} f(x)dx$$

Area Under a PDF = 1 $\int_{-\infty}^{+\infty} f(x)dx = 1$

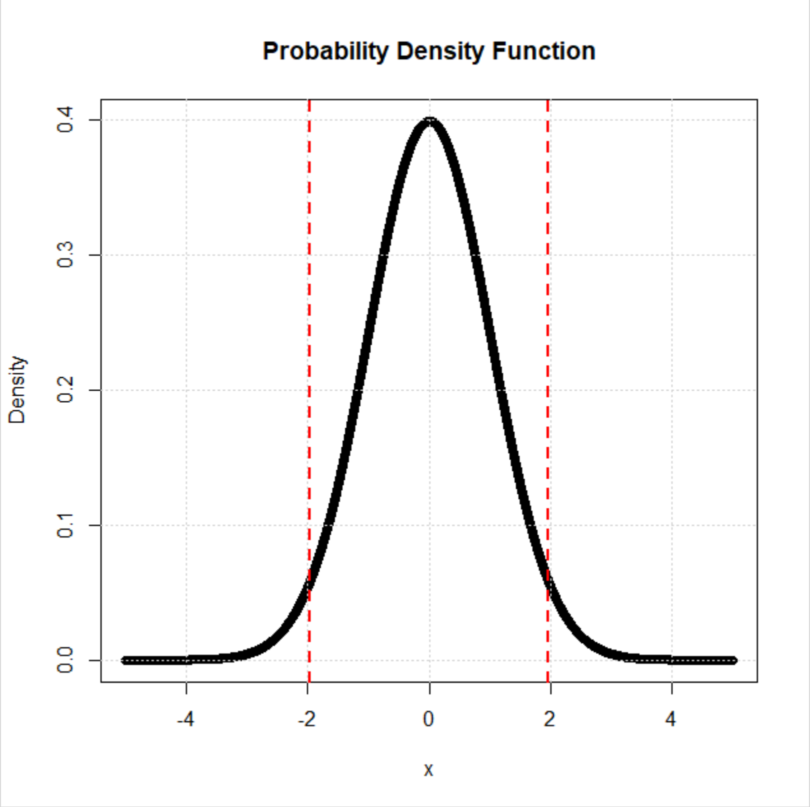
There are many theoretical models to define Probability Density Functions – We will explore them later

Probability Density Function Example

- Calculate the probability of a random variable (X) which follows a normal distribution with a mean zero and standard deviation = 1 between values -1.96 and + 1.96
  - A normal distribution with mean = 0 and standard deviation = 1 is called the standard normal distribution

- We can use the R built in function *dnorm* and *integrate* to do this calculation

```
# Integrate standard normal between -1.96 and + 1.96
P02 <- integrate(dnorm,-1.96,1.96,0,1)
P02
```



```
> 0.9500042 with absolute error < 1e-11
```

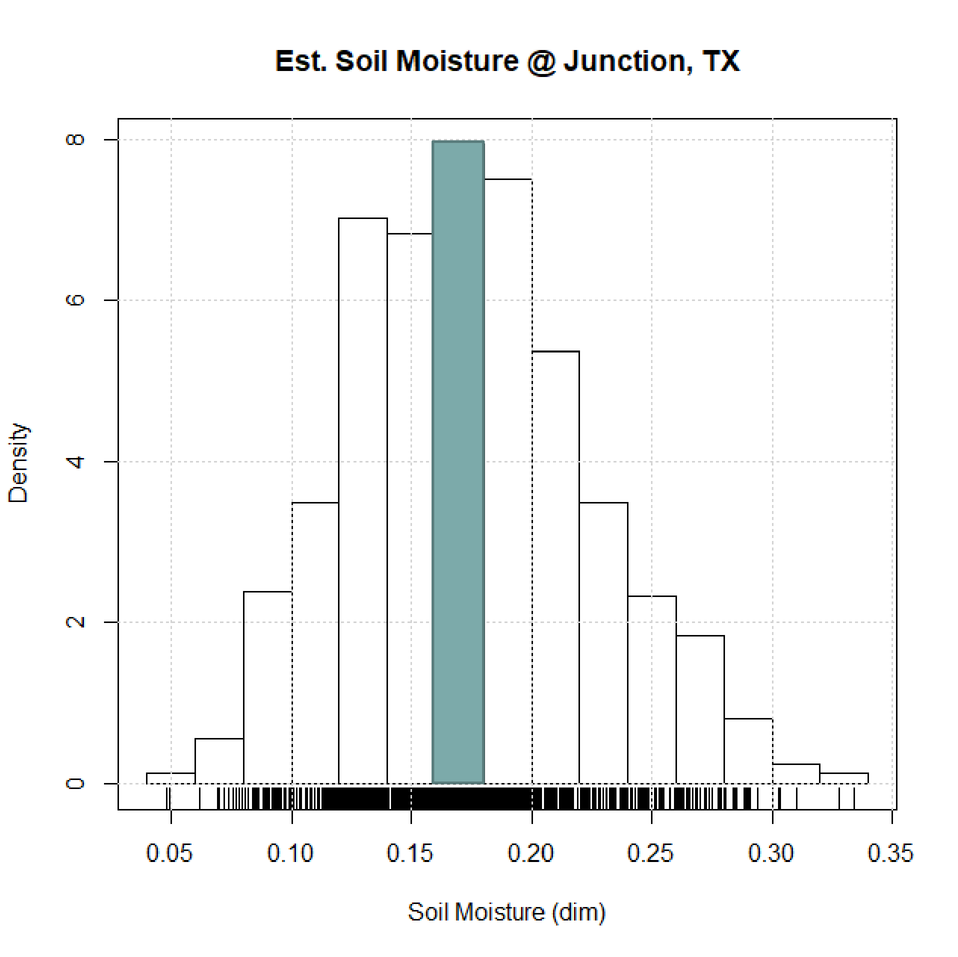
95% of the values of the normal distribution fall within ± ~2 Standard Deviations

Empirical Representation of Probability Functions

- PDF and CDF can be empirically constructed from sample data
- These functions are called empirical PDF and Empirical CDF
- They are based on the assumption that the sampled data provides a reasonable estimate of the population
- They are mostly used for preliminary analysis and visualization
- They are also used to identify appropriate theoretical models
  - Match theoretical model to observed PDF and CDF
  - We shall explore this idea in detail later in the class
- They are limited in that we have not observed everything from the population
  - But they provide the best available observed information we have at hand

Empirical Probability Representation

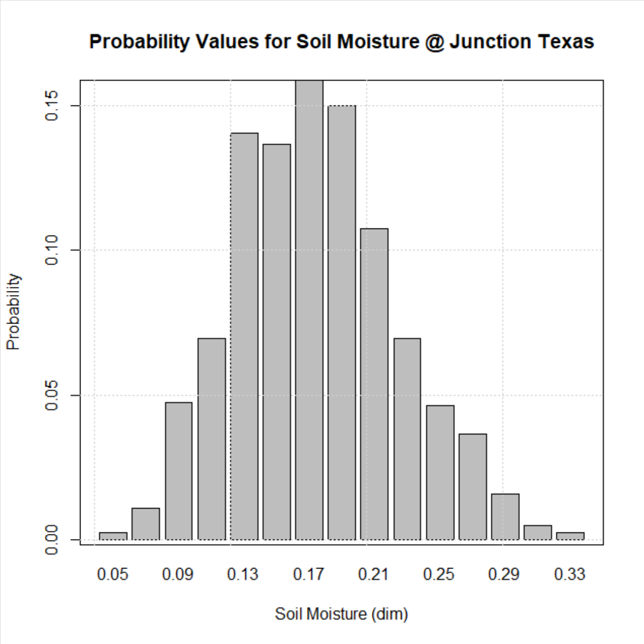
- The histogram is the basic representation of probability
  - Relative Frequency or density is plotted on the Y-axis instead of counts
- Each bar of the histogram depicts the (density) or likelihood of observing values in that bin (range)
- The product of the bin size x density gives the probability of the RV being in that range



$$P\ (\text{soil moisture} \geq 0.16 \text{ and } < 0.18) = \underset{\substack{\uparrow \\ \text{Bin Width}}}{0.02} \text{ x } \underset{\substack{\nwarrow \\ \text{Density}}}{7.94} = 0.158$$

Histogram Computations using R

- R can be used to plot histogram
  - Set freq = FALSE to obtain density plot instead of counts
  - Can be used to obtain breaks and density values
  - Midpoints can be used to plot probability bar plot for illustration

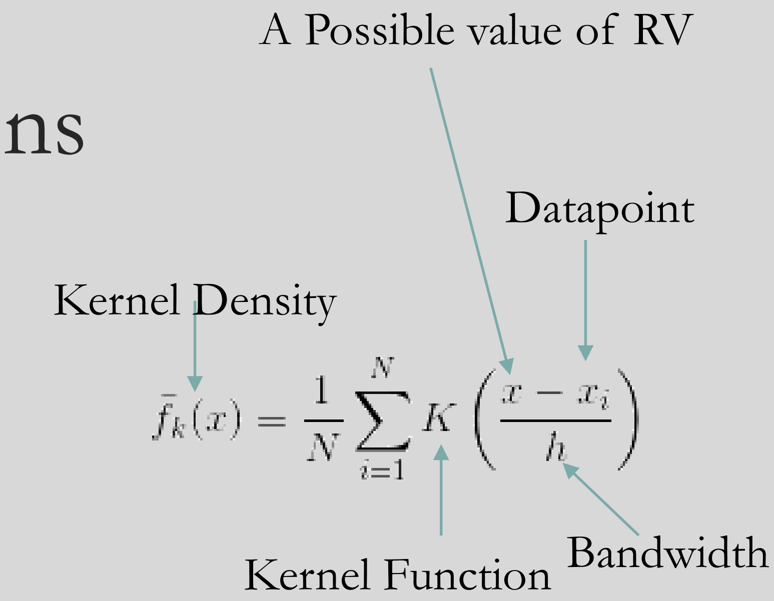


```
# Script to plot histogram and obtain densities
# Written by Venki Uddameri, Ph.D., P.E.


# Set Working Directory
setwd('D:\\Dropbox\\CE5331-Probabilistic Methods- Fall2019')

# Read data
a <- read.csv('soilmoisturedata.csv')

# Extract Junction data from the dataset
JSM <- a$Junction..mm
summary(JSM)
JSM <- JSM/1600  # Divide by 1600 cm (soil column length)

# Plot the histogram with freq-False to get density
sm.hist <- hist(JSM,freq=FALSE,xlab='Soil Moisture (dim)',main='Soil Moisture @ Junction, TX')
box()
grid()

# Extract bin width and densities
sm.bin <- sm.hist$breaks
sm.den <- sm.hist$density
sm.lag <- diff(sm.bin,1)  # Compute the bin width
sm.prob <- sm.lag*sm.den  # Compute probabilities
sum(sm.prob) # check to see if  the sum of  probabilities = 1
```

Note: A bar plot looks like a histogram but it is not!!

Empirical Probability Density Functions

- Kernel Density estimation can also be used to plot empirical density functions from data
- Kernel density plot provides a smoother representation of the PDF
  - Overcomes the jaggedness noted in histograms
- Kernel density is obtained by pivoting a ‘Kernel function’ on each point of the data
  - A Standard Normal distribution (Gaussian Kernel) is often used



$$\bar{f}_k(x) = \frac{1}{N}\sum_{i=1}^{N} K\left(\frac{x - x_i}{h}\right)$$

Kernel Density is sensitive to the choices of Kernel function and the Bandwidth

Choice of the Kernel function is less important than the selection of bandwidth

Small bandwidth will make the density function spikey

Large bandwidth will make the density function too smooth

To be meaningful Kernels must be centered at zero (locally) and integrate to 1

Kernel Density Estimation using R

- R provides density function to do Kernel Density Estimation
  - Default is Gaussian Kernel which can be changed
    - "epanechnikov", "rectangular", "triangular", "biweight", "cosine", "optcosine“
    - Some of these provide more compact bounds than the Gaussian
  - By default the bandwidth is selected based on normal reference rule

$$h = 0.9 \times min\left(sd, \frac{IQR}{1.34n^{-\frac{1}{5}}}\right)$$

  - You can either specify the bandwidth or obtain a best-fit using cross-validation
    - bw = 4 or bw=‘ucv’ (unbiased cross-validation) or bw=‘bcv’ (bias cross-validation estimate)

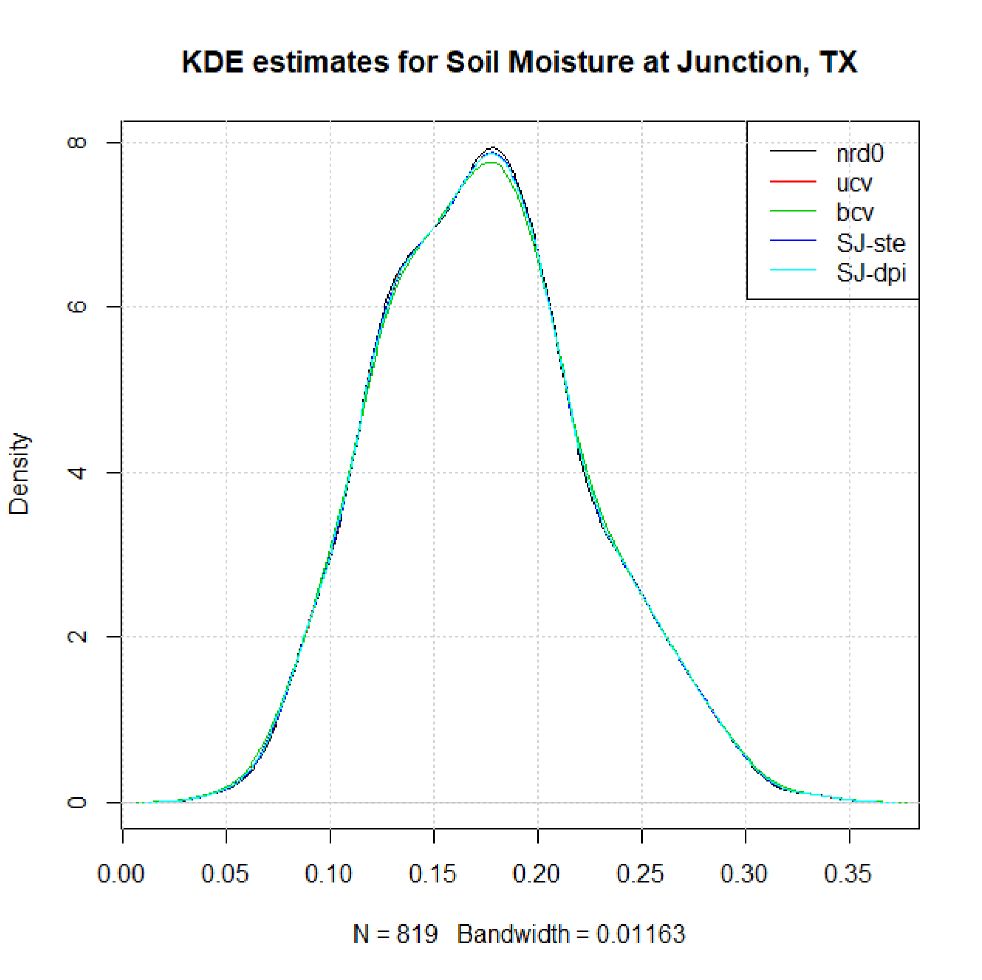
Improper choices for bandwidth can alter the range of the Random Variable

Kernel Density Estimation involves some art and science
Some trial-and-error experimentation is necessary to get an aesthetically pleasing fit
Need to be even more careful if the density is to be used in calculations

KDE with different Bandwidths

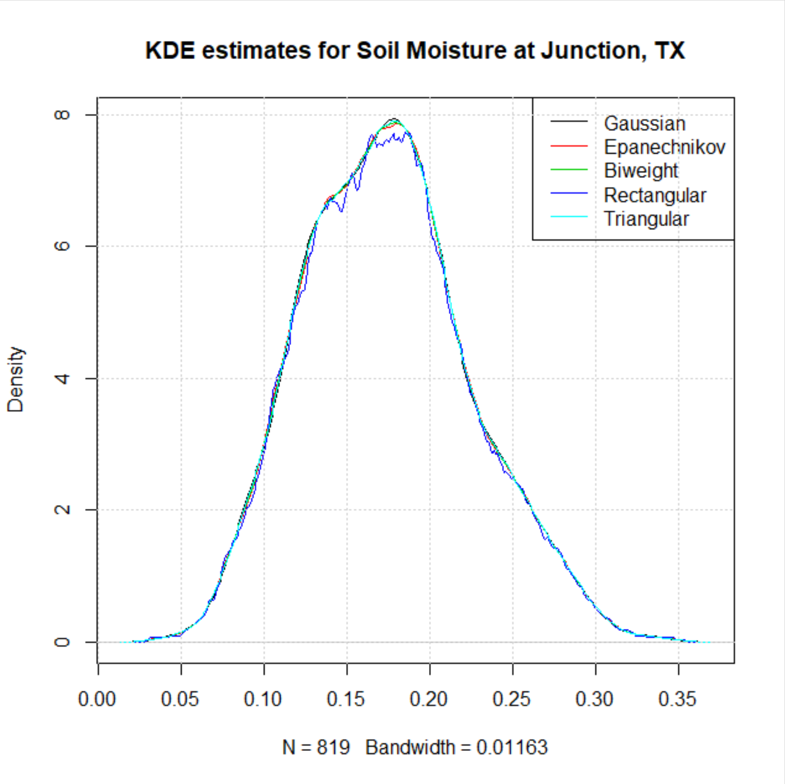
- Construct KDE for soil moisture in Junction, TX using Gaussian KDE but with default bandwidth; bandwidth = 'ucv', bandwidth='bcv' and bandwidth='SJ-ste' and bandwidth='SJ-dpi'

Method	Value	Remarks
nrd0	0.0116	default
ucv	0.0127	Unconstrained cross-validation
bcv	0.0143	Biased cross-validation
SJ-ste	0.0125	Sheater-Jones (solve the equation)
SJ-dpi	0.0127	Sheater-Jones (direct plug-in)



Example of KDE

- Plot KDE for Soil Moisture Data in Junction, TX using different Kernels and default bandwidth



```
# Script to plot KDE using various Kernels
# Assumes default bandwidth
# Venki Uddameri, Ph.D. P.E.
# Set working directory
setwd('D:\\Dropbox\\CE5331-Probabilistic Methods- Fall2019')

# Read the data
a <- read.csv('soilmoisturedata.csv')

# Extract soil moisture for Junction station
JSM <- a$Junction..mm
summary(JSM)
# Normalize it to dimensionless soil moisture
JSM <- JSM/1600
summary(JSM)

# KDE with different density estimators
plot(density(JSM),main="KDE estimates for Soil Moisture at Junction, TX",col=1)
lines(density(JSM,kernel = "epanechnikov"),col=2)
lines(density(JSM,kernel='biweight'),col=3)
lines(density(JSM,kernel='rectangular'),col=4)
lines(density(JSM,kernel='triangular'),col=5)
legend('topright',legend=c('Gaussian','Epanechnikov','Biweight','Rectangular','Triangular’)
, col=c(1,2,3,4,5),lty=c(1,1,1,1,1))
grid()
```

Most functions except ‘Rectangular’ give similar results

KDE - Density

- Compute the area under the curve (AUC) for the densities obtained using different kernels for Soil Moisture Data in Junction, TX
- Density gives values of RV x and corresponding f(x) we can numerically integrate using Trapezoidal Rule
  - Trapezoidal rule divides the curve into a set of triangles and trapezoids and computes the area
  - There is a package in R called caTools that provides a function for trapezoidal rule (trapz)

Method	**Area under the curve**
Gaussian	1.000974
Epanachnikov	1.001093
Biweight	1.000977
Rectangular	0.981428
Triangular	1.000974

Reasonable except for Rectangular

```
# Script to calculate KDE area using various Kernels
# Assumes default bandwidth
# Venki Uddameri, Ph.D., P.E.

# load library (needs to be installed before using it)
library(caTools)

# Set working directory
setwd('D:\\Dropbox\\CE5331-Probabilistic Methods- Fall2019')

# Read the data
a <- read.csv('soilmoisturedata.csv')

# Extract soil moisture for Junction station
JSM <- a$Junction..mm
summary(JSM)
# Normalize it to dimensionless soil moisture
JSM <- JSM/1600

# KDE with different density estimators
JSM.gau <- density(JSM)
JSM.epa <- density(JSM,kernel = "epanechnikov")
JSM.biw <- density(JSM,kernel='biweight')
JSM.rec <- density(JSM,kernel='rectangular')
JSM.tri <- density(JSM,kernel='triangular')

# compute the area under the curve
auc.gau <- trapz(JSM.gau$x,JSM.gau$y)
auc.epa <- trapz(JSM.epa$x,JSM.epa$y)
auc.biw <- trapz(JSM.biw$x,JSM.biw$y)
auc.rec <- trapz(JSM.rec$x,JSM.rec$y)
auc.tri <- trapz(JSM.tri$x,JSM.tri$y)
cat(auc.gau,auc.epa,auc.biw,auc.rec,auc.tri,'\n')
```

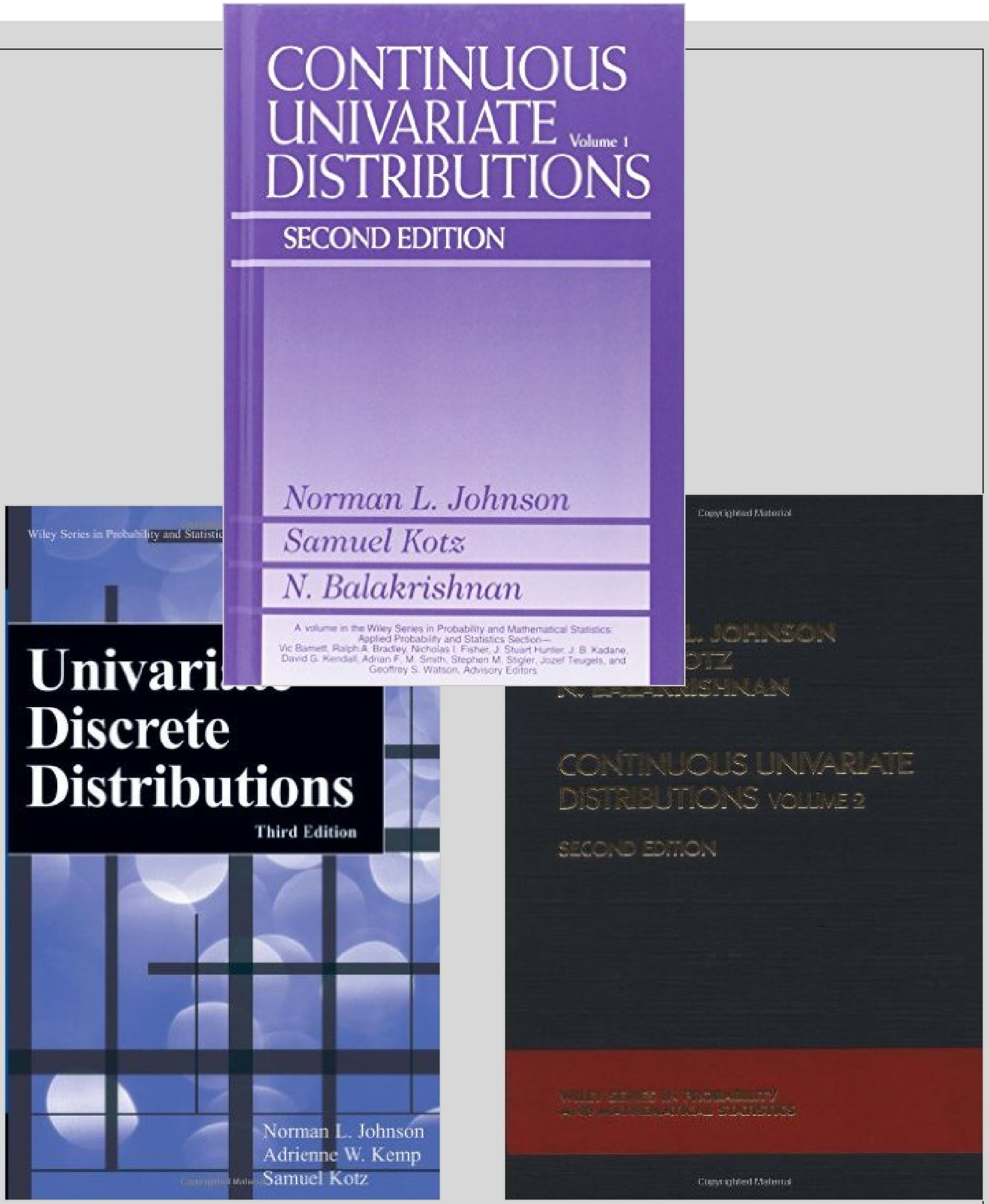
KDE using R packages

- There are several packages available in R for computing Kernel Density Function
  - A good review is provided by Deng and Wickham (2011) - https://vita.had.co.nz/papers/density-estimation.pdf
  - Many packages can fit KDE in multiple dimensions
    - Can be used to model joint distributions
- The use of libraries help improve the visual aesthetics of KDEs
- These libraries can also provide better estimates
  - Follow the rules of probability better

SELECT COMMON DISTRIBUTIONS

Probability Models

- Random variables are characterized using probability distributions
- A probability distribution specifies a relationship between the magnitude of the random variable and its associated probability
- There are several hundred probability distribution functions specified in the literature


CONTINUOUS UNIVARIATE DISTRIBUTIONS Volume 1
SECOND EDITION
Norman L. Johnson
Samuel Kotz
N. Balakrishnan
A volume in the Wiley Series in Probability and Mathematical Statistics
Wiley Series in Probability and Statistics
Univariate Discrete Distributions
Third Edition
Norman L. Johnson
Adrienne W. Kemp
Samuel Kotz
CONTINUOUS UNIVARIATE DISTRIBUTIONS VOLUME 2
SECOND EDITION

Probability Models are always for population

Normal and lognormal Distributions

- The normal distributions is useful to measure central tendencies
  - Defined using two parameters
    - Mean (μ) and Standard Deviation (σ)
  - Leads to symmetric distribution
  - Represents additive processes
- The lognormal distribution assumes the log of random variable x (i.e., log(x) is randomly distributed
  - Does not work for negative data
  - Represents multiplicative processes

$$f(x) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{1}{2} \frac{(x-\mu)^2}{\sigma^2} \right) \quad \forall -\infty \leq x \leq \infty$$

$$F(x) = \int_{-\infty}^{x} f(x)dx$$

The same formulas apply for both normal and lognormal distributions. However, the data is log-transformed first for lognormal distribution

Lognormal distribution is appropriate when the data are skewed

Normal and Lognormal Distributions are best suited to represent central tendencies

Poisson Distribution

- It is a discrete distribution used to represent the number of independent events within a fixed time
  - Number of **independent** rainfall events within a year
- Closely related to Exponential Distribution
  - Continuous distribution for inter-arrival times
- Poisson distribution assumes stationarity
  - Rate at which events occur is constant

$$p\left(k\ events\ in\ a\ \mathrm{int}\ erval\right)=\frac{\lambda^k e^{-\lambda}}{k!}$$

$$F\left(k\right)=\lambda^k\sum_{i=1}^{k}\frac{e^{-i}}{i!}$$

λ =Average number of events per interval

Poisson Distribution is used for count data

For a discrete distribution we use the term probability mass function (PMF) instead of pdf
PMF directly gives us the probability instead of probability density

Exponential Distribution

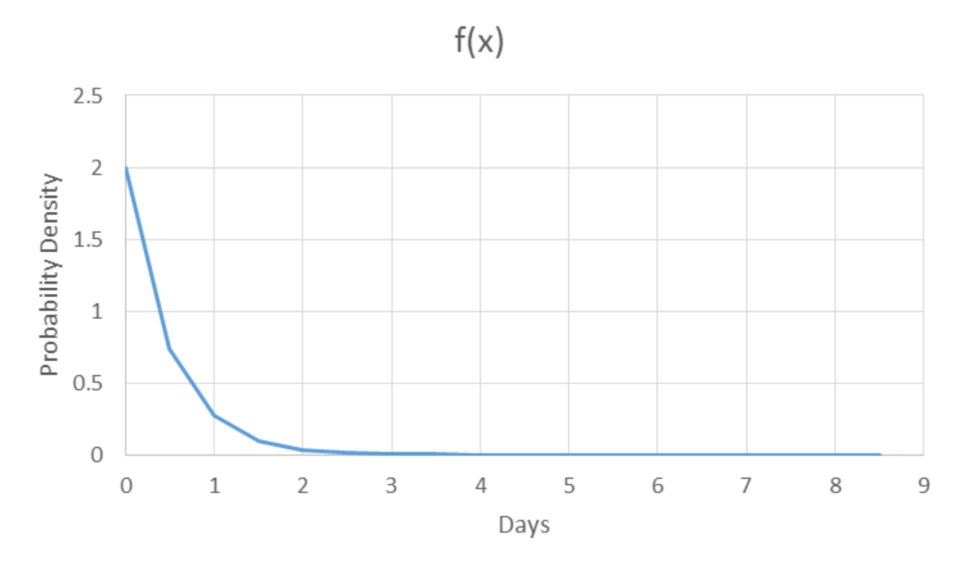
- Often used to model the time between two independent events
  - Time between two rainstorms
- Exponential distribution is represented by the parameter (λ)
  - Reciprocal of Average inter-event time
- Closely related to Poisson Distribution
  - A discrete distribution for number of events in a fixed time

$$f(t) = \lambda e^{-\lambda t}$$

$$F(t) = \int_0^t f(t)dt = 1 - e^{-\lambda t}$$

$$E(t) = 1/\lambda$$

$$Var(t) = 1/\lambda^2$$



Binomial Distribution

- Used when there are two outcomes
  - Success and Failure
- The probability of success for each event is denoted by “p”
  - Often assumed stationary
- The PMF calculates the probability of success of x out of N total events

This distribution is very useful for risk and reliability calculations

$$f(x) = p(x) = \binom{N}{x} p^x (1-p)^{N-x}$$

$$F(x) = \sum_{i=0}^{x} \binom{N}{i} p^i (1-p)^{N-i}$$

This distribution is the basis for modeling discrete binary data

Logistic Distribution

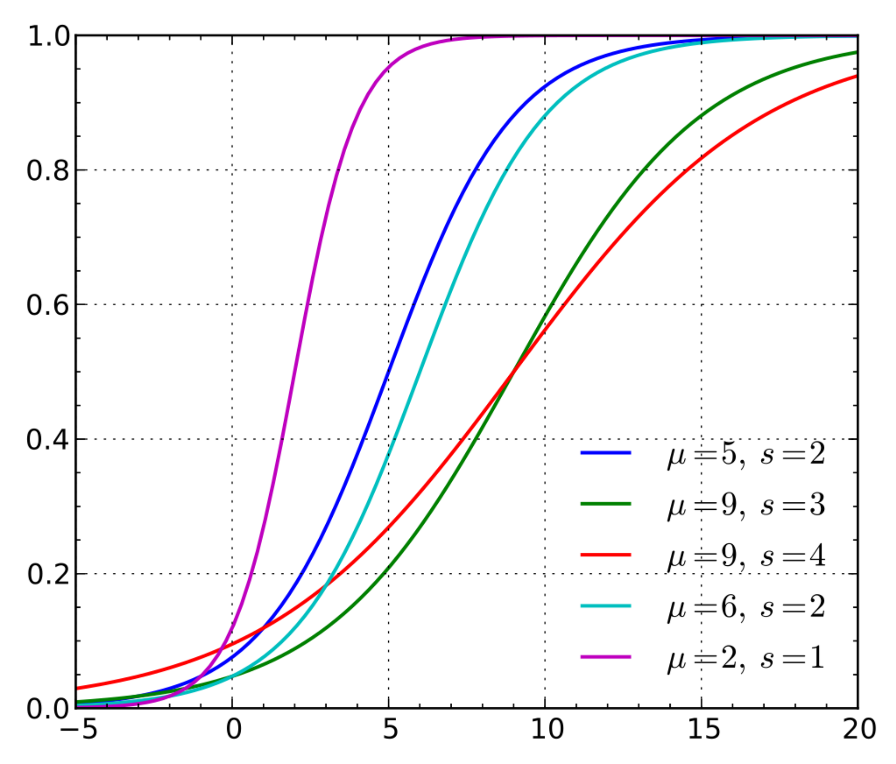
- Useful to model Probability of occurrence of binary data
  - Provides continuous probability function
  - S shaped curve representing growth patterns

$$f(x;\mu,s) = \frac{e^{-(x-\mu)/s}}{s\left(1+e^{-(x-\mu)/s}\right)^2}$$

$$= \frac{1}{s\left(e^{(x-\mu)/(2s)}+e^{-(x-\mu)/(2s)}\right)^2}$$

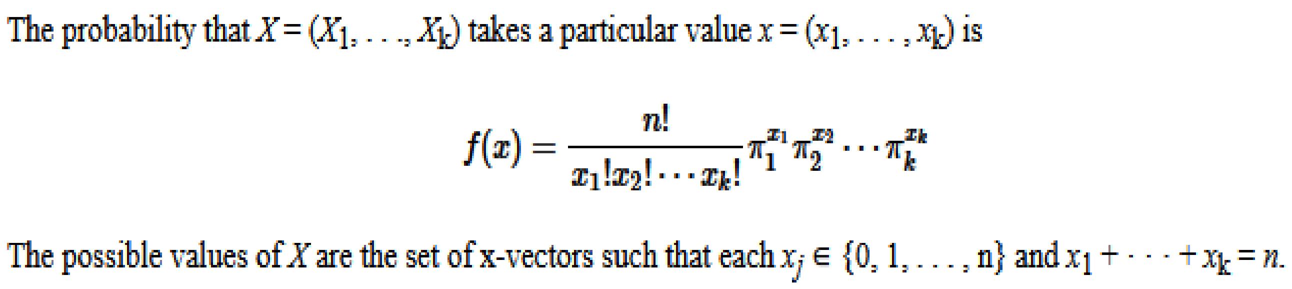
$$= \frac{1}{4s}\operatorname{sech}^2\left(\frac{x-\mu}{2s}\right).$$

$$F(x;\mu,s) = \frac{1}{1+e^{-(x-\mu)/s}} = \frac{1}{2}+\frac{1}{2}\tanh\left(\frac{x-\mu}{2s}\right).$$



Multinomial Distribution

- A generalization of binomial distribution
- Used to model nominal or ordinal data
  - A small set of discrete choices greater than 2
- Generally used to model sampling of k from a set of n with replacement



The probability that $X = (X_1, . . ., X_k)$ takes a particular value $x = (x_1, . . . , x_k)$ is

$$f(x) = \frac{n!}{x_1!x_2!\cdots x_k!}\pi_1^{x_1}\pi_2^{x_2}\cdots\pi_k^{x_k}$$

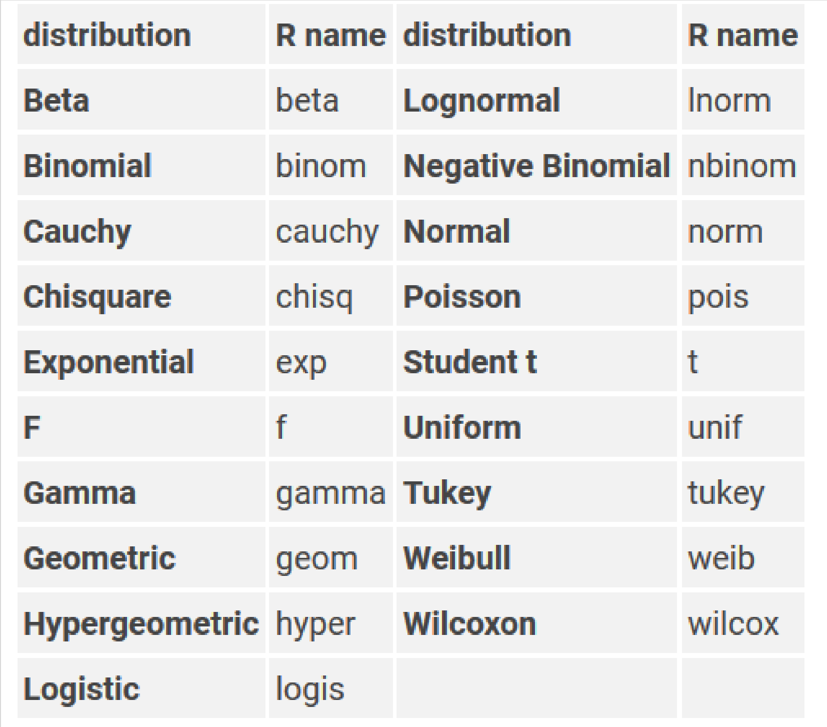
The possible values of X are the set of x-vectors such that each $x_j \in \{0, 1, . . . , n\}$ and $x_1 + \cdots + x_k = n$.

R Statistical Functions

- There are built-in functions for several univariate distributions
- R follows a consistent approach to naming
  - d*foo* → probability density function (pdf)
  - p*foo* → cumulative distribution function (cdf)
  - q*foo* → quantile function
  - r*foo* → random number generation

Where *foo* is the name of the distribution

Common Distributions in R



distribution	R name	distribution	R name
Beta	beta	**Lognormal**	lnorm
Binomial	binom	**Negative Binomial**	nbinom
Cauchy	cauchy	**Normal**	norm
Chisquare	chisq	**Poisson**	pois
Exponential	exp	**Student t**	t
F	f	**Uniform**	unif
Gamma	gamma	**Tukey**	tukey
Geometric	geom	**Weibull**	weib
Hypergeometric	hyper	**Wilcoxon**	wilcox
Logistic	logis		