# CS 147:
# Computer Systems Performance Analysis
## Summarizing Variability and Determining Distributions





CS 147:
Computer Systems Performance Analysis
Summarizing Variability and Determining Distributions

# Overview







Overview

Introduction
Indices of Dispersion
Range
Variance, Standard Deviation, C.V.
Quantiles
Miscellaneous Measures
Choosing a Measure
Identifying Distributions
Histograms
Kernel Density Estimation
Quantile-Quantile Plots
Statistics of Samples
Meaning of a Sample
Guessing the True Value

# Summarizing Variability

- A single number rarely tells entire story of a data set
- Usually, you need to know how much the rest of the data set varies from that index of central tendency





Summarizing Variability

- A single number rarely tells entire story of a data set
- Usually, you need to know how much the rest of the data set varies from that index of central tendency

# Why Is Variability Important?

- Consider two Web servers:
  - Server A services all requests in 1 second
  - Server B services 90% of all requests in .5 seconds
    - But 10% in 55 seconds
  - Both have mean service times of 1 second
  - But which would you prefer to use?





Why Is Variability Important?

- Consider two Web servers:
  - Server A services all requests in 1 second
  - Server B services 90% of all requests in .5 seconds
    - But 10% in 55 seconds
  - Both have mean service times of 1 second
  - But which would you prefer to use?

# Indices of Dispersion

- Measures of how much a data set varies
  - Range
  - Variance and standard deviation
  - Percentiles
  - Semi-interquartile range
  - Mean absolute deviation





Indices of Dispersion

- Measures of how much a data set varies
  - Range
  - Variance and standard deviation
  - Percentiles
  - Semi-interquartile range
  - Mean absolute deviation

# Range

- Minimum & maximum values in data set
- Can be tracked as data values arrive
- Variability characterized by difference between minimum and maximum
- Often not useful, due to outliers
- Minimum tends to go to zero
- Maximum tends to increase over time
- Not useful for unbounded variables





Range

- Minimum & maximum values in data set
- Can be tracked as data values arrive
- Variability characterized by difference between minimum and maximum
- Often not useful, due to outliers
- Minimum tends to go to zero
- Maximum tends to increase over time
- Not useful for unbounded variables

# Example of Range

- For data set 2, 5.4, -17, 2056, 445, -4.8, 84.3, 92, 27, -10
  - Maximum is 2056
  - Minimum is -17
  - Range is 2073
  - While arithmetic mean is 268





Example of Range

- For data set 2, 5.4, -17, 2056, 445, -4.8, 84.3, 92, 27, -10
  - Maximum is 2056
  - Minimum is -17
  - Range is 2073
  - While arithmetic mean is 268

# Variance (and Its Cousins)

- Sample variance is

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

- Expressed in units of the measured quantity, squared
  - Which isn't always easy to understand
- Standard deviation and coefficient of variation are derived from variance





Variance (and Its Cousins)

- Sample variance is

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

- Expressed in units of the measured quantity, squared
  - Which isn't always easy to understand
- Standard deviation and coefficient of variation are derived from variance

# Variance Example

- For data set 2, 5.4, -17, 2056, 445, -4.8, 84.3, 92, 27, -10
- Variance is 413746.6
- You can see the problem with variance:
  - Given a mean of 268, what does that variance indicate?





Variance Example

- For data set 2, 5.4, -17, 2056, 445, -4.8, 84.3, 92, 27, -10
- Variance is 413746.6
- You can see the problem with variance:
  - Given a mean of 268, what does that variance indicate?

# Standard Deviation

- Square root of the variance
- In same units as units of metric
- So easier to compare to metric





Standard Deviation
- Square root of the variance
- In same units as units of metric
- So easier to compare to metric

# Standard Deviation Example

- For sample set we've been using, standard deviation is 643
- Given mean of 268, standard deviation clearly shows lots of variability from mean





Standard Deviation Example

- For sample set we've been using, standard deviation is 643
- Given mean of 268, standard deviation clearly shows lots of variability from mean

# Coefficient of Variation

- Ratio of standard deviation to mean
- Normalizes units of these quantities into ratio or percentage
- Often abbreviated C.O.V. or C.V.





Coefficient of Variation
- Ratio of standard deviation to mean
- Normalizes units of these quantities into ratio or percentage
- Often abbreviated C.O.V. or C.V.

# Coefficient of Variation Example

- For sample set we've been using, standard deviation is 643
- Mean is 268
- So C.O.V. is $643/268 \approx 2.4$





Coefficient of Variation Example

- For sample set we've been using, standard deviation is 643
- Mean is 268
- So C.O.V. is $643/268 \approx 2.4$

# Percentiles

- Specification of how observations fall into buckets
- E.g., 5-percentile is observation that is at the lower 5% of the set
  - While 95-percentile is observation at the 95% boundary
- Useful even for unbounded variables





Percentiles
- Specification of how observations fall into buckets
- E.g., 5-percentile is observation that is at the lower 5% of the set
  - While 95-percentile is observation at the 95% boundary
- Useful even for unbounded variables

# Relatives of Percentiles

- Quantiles - fraction between 0 and 1
  - Instead of percentage
  - Also called fractiles
- Deciles—percentiles at 10% boundaries
  - First is 10-percentile, second is 20-percentile, etc.
- Quartiles—divide data set into four parts
  - 25% of sample below first quartile, etc.
  - Second quartile is also median



2015-06-15



Relatives of Percentiles

- Quantiles - fraction between 0 and 1
  - Instead of percentage
  - Also called fractiles
- Deciles—percentiles at 10% boundaries
  - First is 10-percentile, second is 20-percentile, etc.
- Quartiles—divide data set into four parts
  - 25% of sample below first quartile, etc.
  - Second quartile is also median

# Calculating Quantiles

To estimate $\alpha$-quantile:

- First sort the set
- Then take $[(n-1)\alpha + 1]^{\text{th}}$ element
  - 1-indexed
  - Round to nearest integer index
  - Exception: for small sets, may be better to choose “intermediate” value as is done for median





Calculating Quantiles

To estimate $\alpha$-quantile:

- First sort the set
- Then take $[(n-1)\alpha + 1]^{\text{th}}$ element
  - 1-indexed
  - Round to nearest integer index
  - Exception: for small sets, may be better to choose “intermediate” value as is done for median

# Quartile Example

- For data set 2, 5.4, -17, 2056, 445, -4.8, 84.3, 92, 27, -10 (10 observations)
- Sort it: -17, -10, -4.8, 2, 5.4, 27, 84.3, 92, 445, 2056
- First quartile, Q1, is -4.8
- Third quartile, Q3, is 92





Quartile Example

- For data set 2, 5.4, -17, 2056, 445, -4.8, 84.3, 92, 27, -10 (10 observations)
- Sort it: -17, -10, -4.8, 2, 5.4, 27, 84.3, 92, 445, 2056
- First quartile, Q1, is -4.8
- Third quartile, Q3, is 92

# Interquartile Range

- Yet another measure of dispersion
- The difference between Q3 and Q1
- Semi-interquartile range is half that:

$$\text{SIQR} = \frac{Q_3 - Q_1}{2}$$

- Often interesting measure of what's going on in middle of range
  - Basically indicates distance of quartiles from median





Interquartile Range
- Yet another measure of dispersion
- The difference between Q3 and Q1
- Semi-interquartile range is half that:
SIQR = (Q3 − Q1)/2
- Often interesting measure of what's going on in middle of range
  - Basically indicates distance of quartiles from median

# Semi-Interquartile Range Example

For data set -17, -10, -4.8, 2, 5.4, 27, 84.3, 92, 445, 2056

- Q3 is 92
- Q1 is -4.8

$$\text{SIQR} = \frac{Q_3 - Q_1}{2} = \frac{92 - (-4.8)}{2} = 48$$

- Compare to standard deviation of 643
  - Suggests that much of variability is caused by outliers





Semi-Interquartile Range Example

For data set -17, -10, -4.8, 2, 5.4, 27, 84.3, 92, 445, 2056
- Q3 is 92
- Q1 is -4.8

$$\text{SIQR} = \frac{Q_3 - Q_1}{2} = \frac{92 - (-4.8)}{2} = 48$$

- Compare to standard deviation of 643
  - Suggests that much of variability is caused by outliers

# Mean Absolute Deviation

- Yet another measure of variability
- Mean absolute deviation = $\frac{1}{n}\sum_{i=1}^{n}|x_i - \overline{x}|$
- Good for hand calculation (doesn't require multiplication or square roots)





Mean Absolute Deviation

- Yet another measure of variability
- Mean absolute deviation = $\frac{1}{n}\sum_{i=1}^{n}|x_i - \overline{x}|$
- Good for hand calculation (doesn't require multiplication or square roots)

# Mean Absolute Deviation Example

For data set -17, -10, -4.8, 2, 5.4, 27, 84.3, 92, 445, 2056

- Mean absolute deviation is

$$\frac{1}{10}\sum_{i=1}^{10}|x_i - 268| = 393$$





Mean Absolute Deviation Example
For data set -17, -10, -4.8, 2, 5.4, 27, 84.3, 92, 445, 2056
Mean absolute deviation is
$$\frac{1}{10}\sum_{i=1}^{10}|x_i - 268| = 393$$

# Sensitivity To Outliers

- From most to least,
- Range
- Variance
- Mean absolute deviation
- Semi-interquartile range

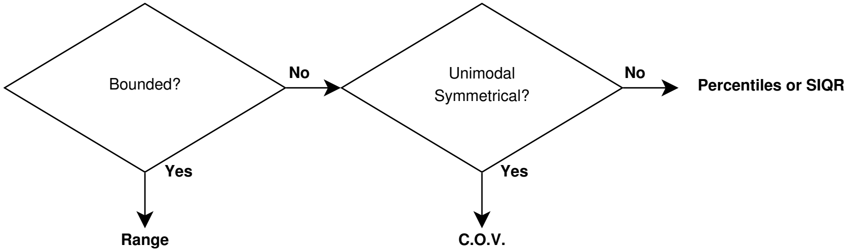




Sensitivity To Outliers

- From most to least,
- Range
- Variance
- Mean absolute deviation
- Semi-interquartile range

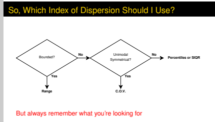
# So, Which Index of Dispersion Should I Use?




But always remember what you're looking for







# Finding a Distribution for Datasets

- If a data set has a common distribution, that's the best way to summarize it
- Saying a data set is uniformly distributed is more informative than just giving mean and standard deviation
- So how do you determine if your data set fits a distribution?





Finding a Distribution for Datasets

- If a data set has a common distribution, that's the best way to summarize it
- Saying a data set is uniformly distributed is more informative than just giving mean and standard deviation
- So how do you determine if your data set fits a distribution?

# Methods of Determining a Distribution

- Plot a histogram
- Kernel density estimation
- Quantile-quantile plot
- Statistical methods (not covered in this class)





Methods of Determining a Distribution
- Plot a histogram
- Kernel density estimation
- Quantile-quantile plot
- Statistical methods (not covered in this class)

# Plotting a Histogram

Suitable if you have relatively large number of data points

Procedure:

1. Determine range of observations
2. Divide range into buckets
3. Count number of observations in each bucket
4. Divide by total number of observations and plot as column chart





Plotting a Histogram

Suitable if you have relatively large number of data points

Procedure:

1. Determine range of observations
2. Divide range into buckets
3. Count number of observations in each bucket
4. Divide by total number of observations and plot as column chart

# Problems With Histogram Approach

- Determining cell size
  - If too small, too few observations per cell
  - If too large, no useful details in plot
- If fewer than five observations in a cell, cell size is too small





Problems With Histogram Approach

- Determining cell size
  - If too small, too few observations per cell
  - If too large, no useful details in plot
- If fewer than five observations in a cell, cell size is too small

# Kernel Density Estimation

- Basic idea: any observation represents probability of high *near* near that observation
- Example:
  - Seeing 7 means pdf is high all around 7
  - Seeing 6.5 *also* means pdf is high near 7
- "Average out" observations to get smooth histogram





Kernel Density Estimation

- Basic idea: any observation represents probability of high *near* near that observation
- Example:
  - Seeing 7 means pdf is high all around 7
  - Seeing 6.5 *also* means pdf is high near 7
- "Average out" observations to get smooth histogram

# KDE Equations

- Want to estimate continuous $p(x)$:

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

- Where $K(x)$ is *kernel function*
- Must integrate to unity: $\int_{-\infty}^{\infty} K(x)\,dx = 1$
- Purpose is to select nearby samples
- $h$ is *bandwidth* parameter
- Controls how many nearby samples selected
- Large bandwidth $\Rightarrow$ more smoothing, less detail





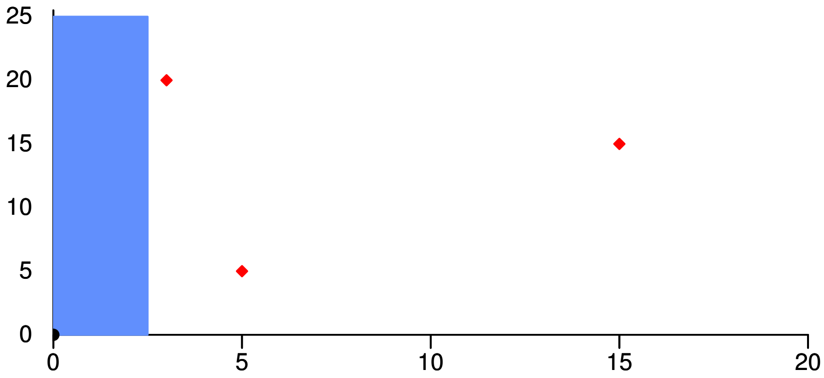
KDE Equations

- Want to estimate continuous $p(x)$:

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

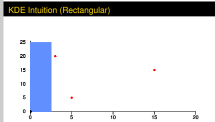
- Where $K(x)$ is *kernel function*
- Must integrate to unity: $\int_{-\infty}^{\infty} K(x)\,dx = 1$
- Purpose is to select nearby samples
- $h$ is *bandwidth* parameter
- Controls how many nearby samples selected
- Large bandwidth $\Rightarrow$ more smoothing, less detail

# KDE Intuition (Rectangular)

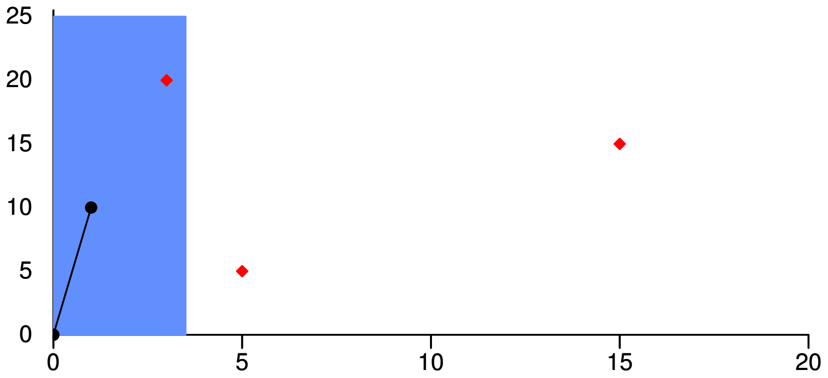





2015-06-15
CS147
Identifying Distributions
Kernel Density Estimation
KDE Intuition (Rectangular)

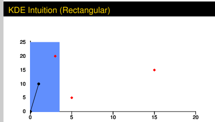


# KDE Intuition (Rectangular)

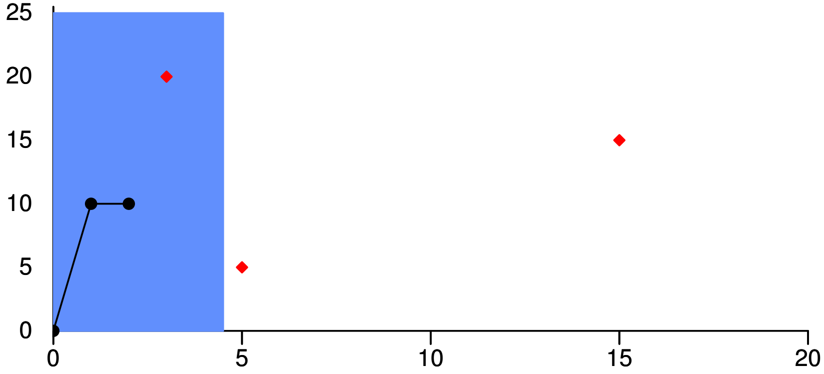






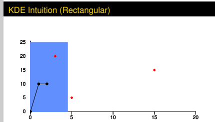
# KDE Intuition (Rectangular)



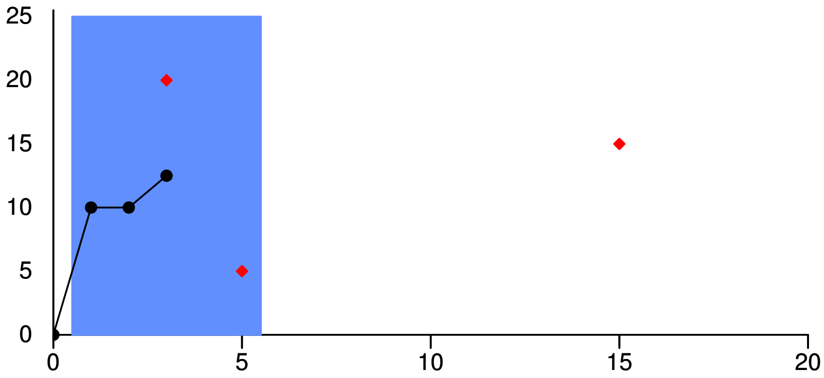



2015-06-15

CS147
└Identifying Distributions
└Kernel Density Estimation
└KDE Intuition (Rectangular)

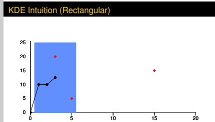


# KDE Intuition (Rectangular)

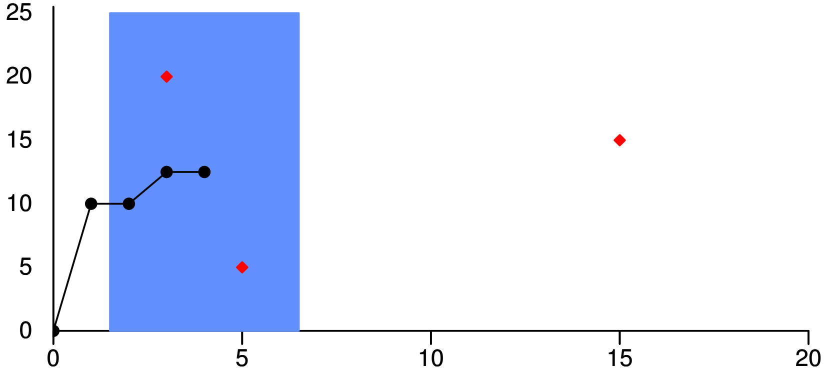






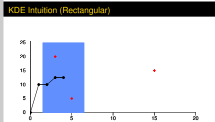



# KDE Intuition (Rectangular)

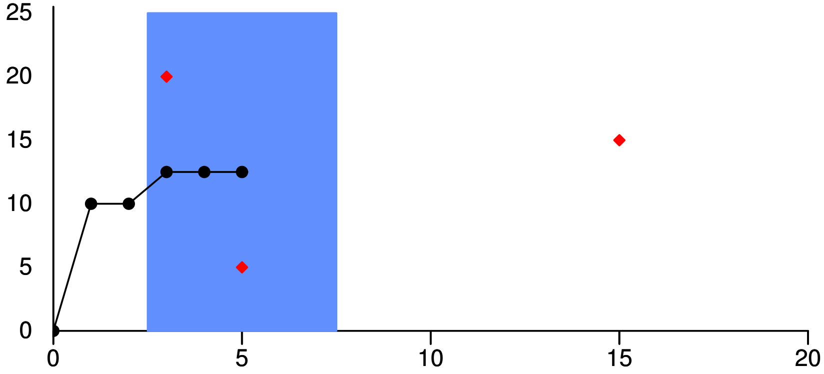






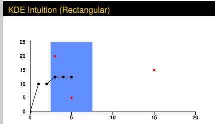



# KDE Intuition (Rectangular)

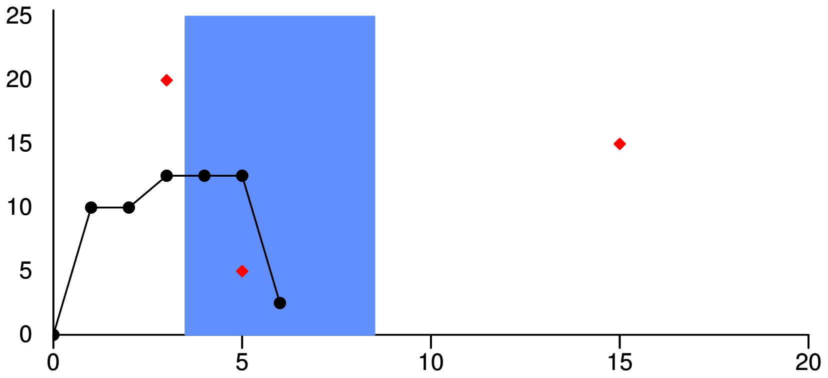






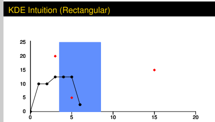



# KDE Intuition (Rectangular)

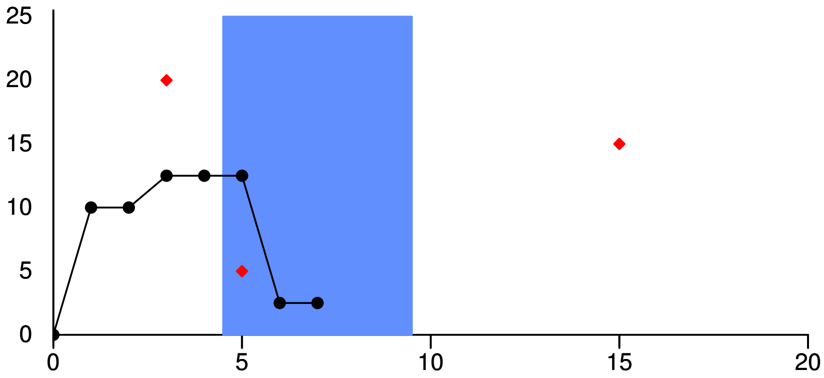






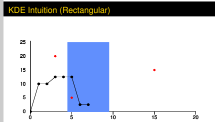



# KDE Intuition (Rectangular)

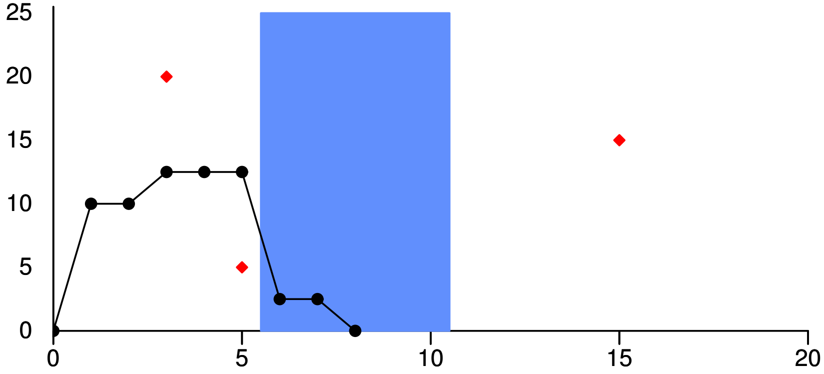






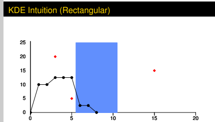



# KDE Intuition (Rectangular)

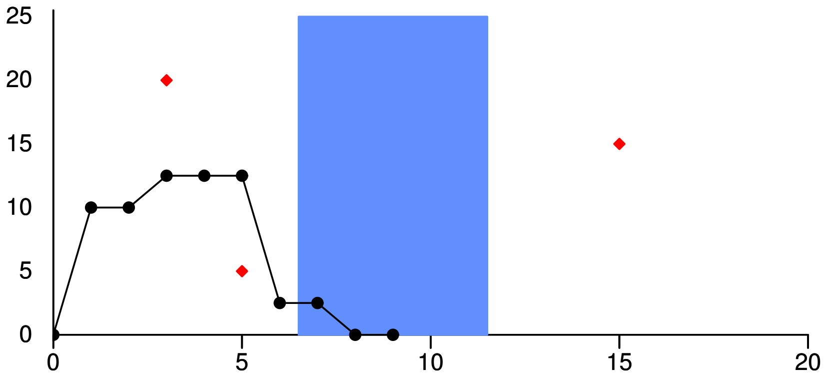






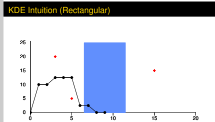



# KDE Intuition (Rectangular)

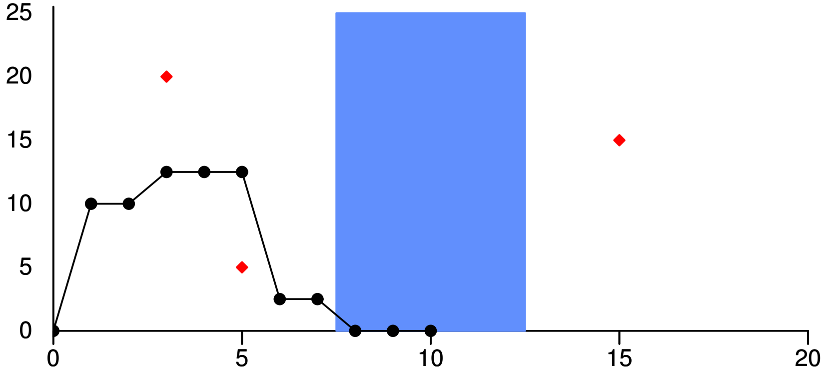






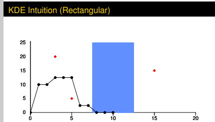



# KDE Intuition (Rectangular)

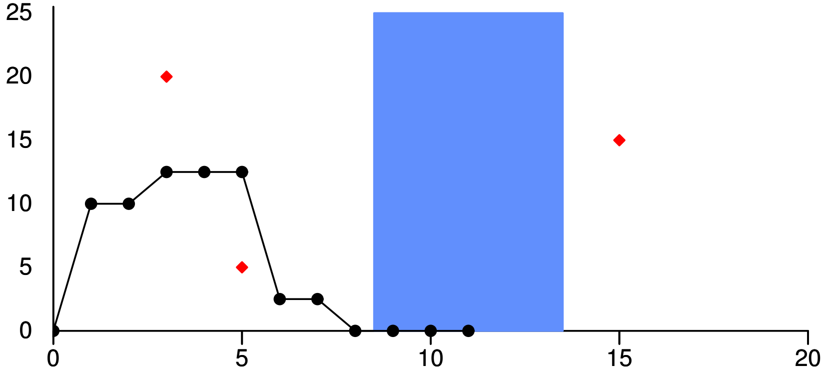






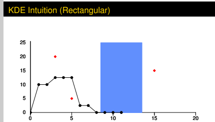



# KDE Intuition (Rectangular)

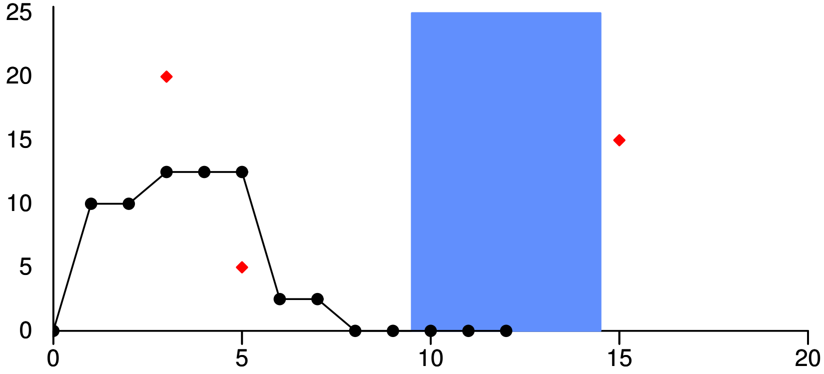






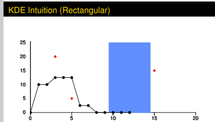
# KDE Intuition (Rectangular)



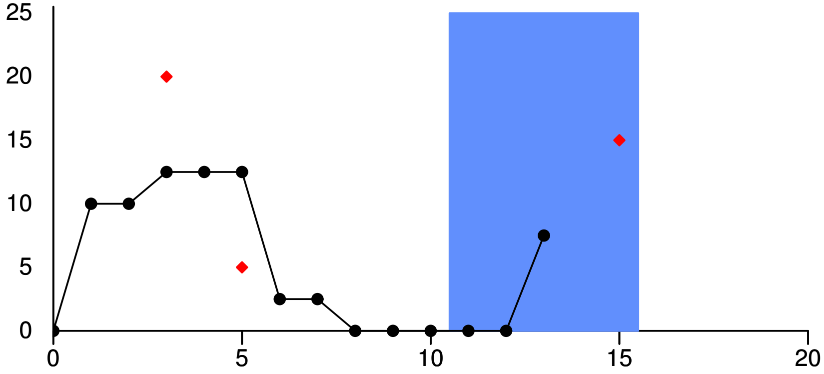



2015-06-15

CS147
└Identifying Distributions
 └Kernel Density Estimation
  └KDE Intuition (Rectangular)



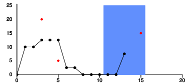
# KDE Intuition (Rectangular)



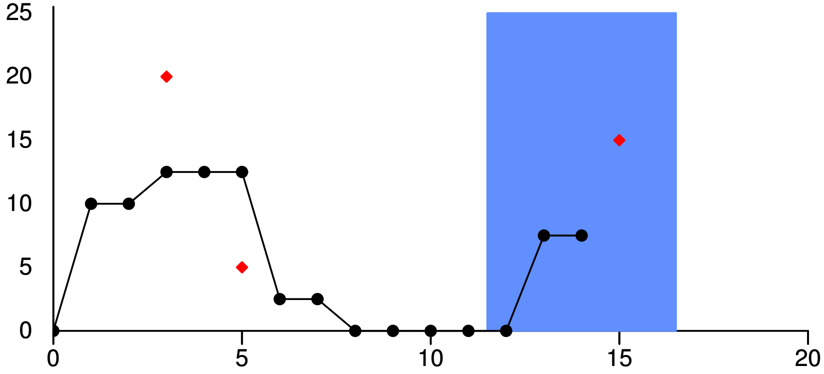





KDE Intuition (Rectangular)



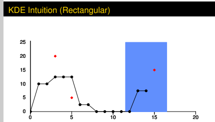
# KDE Intuition (Rectangular)



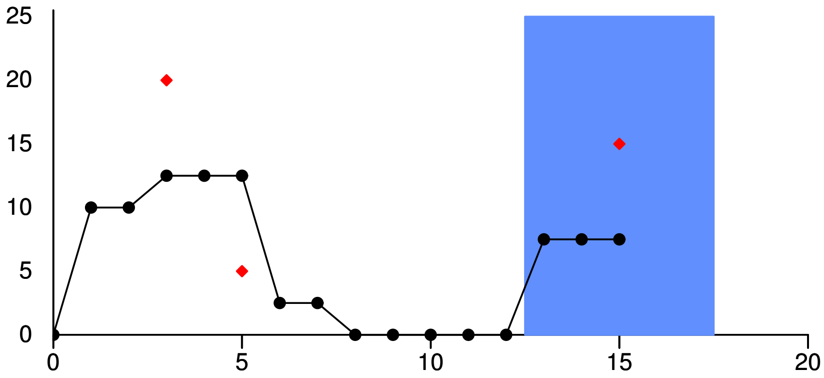



2015-06-15

CS147
└Identifying Distributions
└Kernel Density Estimation
└KDE Intuition (Rectangular)

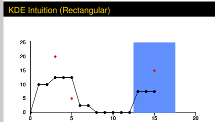


# KDE Intuition (Rectangular)

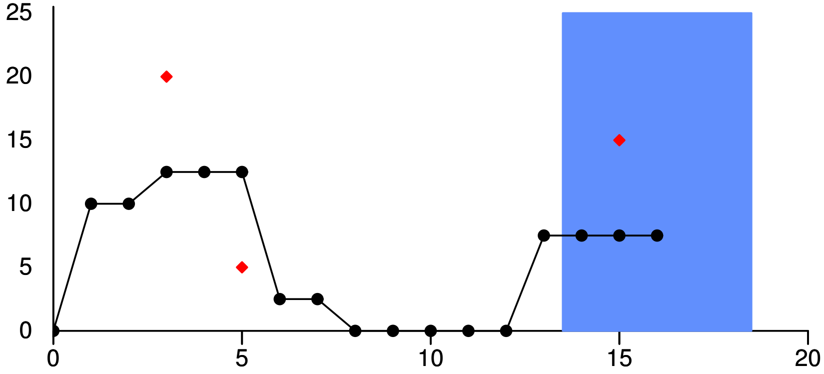






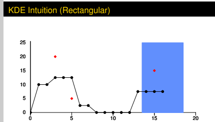
# KDE Intuition (Rectangular)



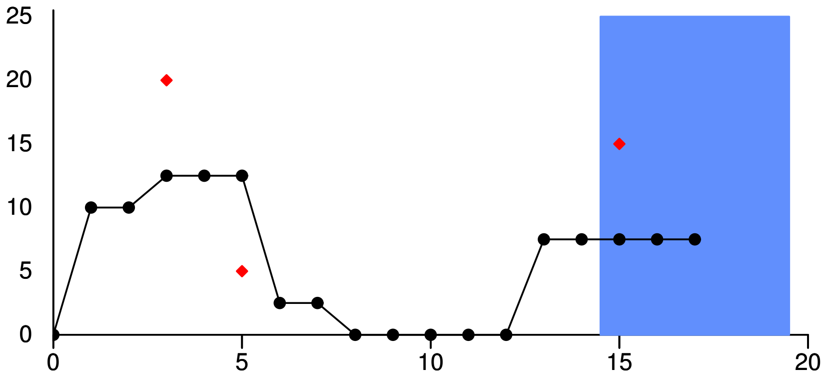



2015-06-15

CS147
└Identifying Distributions
 └Kernel Density Estimation
  └KDE Intuition (Rectangular)

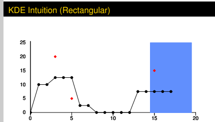


# KDE Intuition (Rectangular)

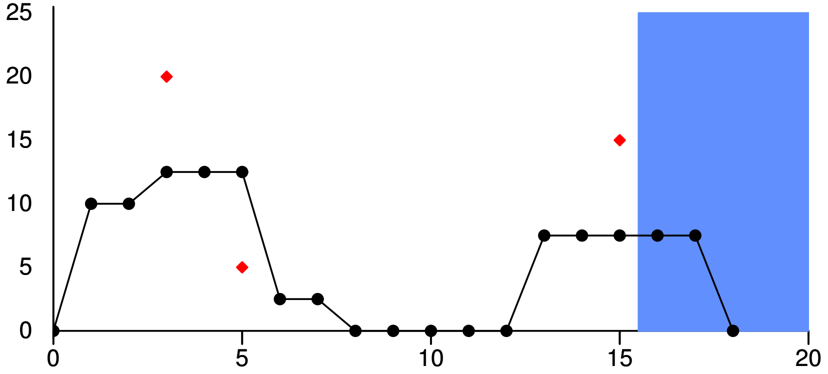






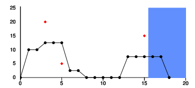
# KDE Intuition (Rectangular)



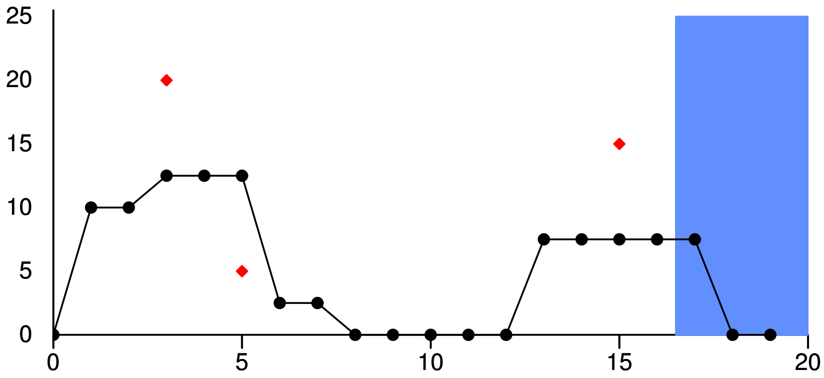





KDE Intuition (Rectangular)

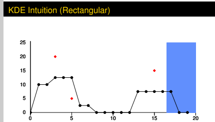


# KDE Intuition (Rectangular)

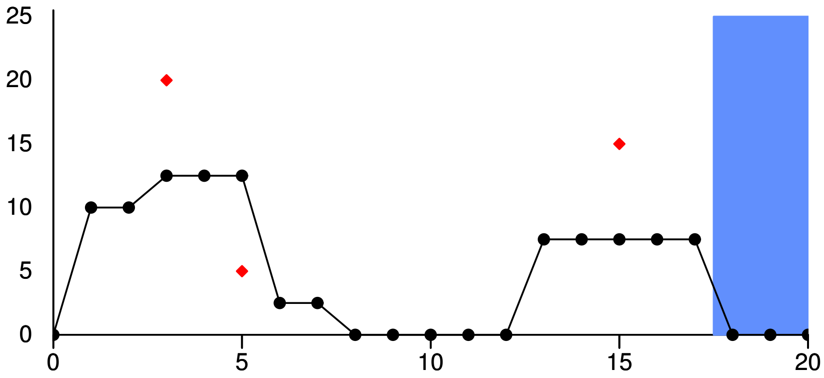






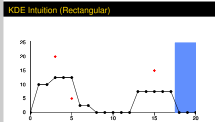
# KDE Intuition (Rectangular)



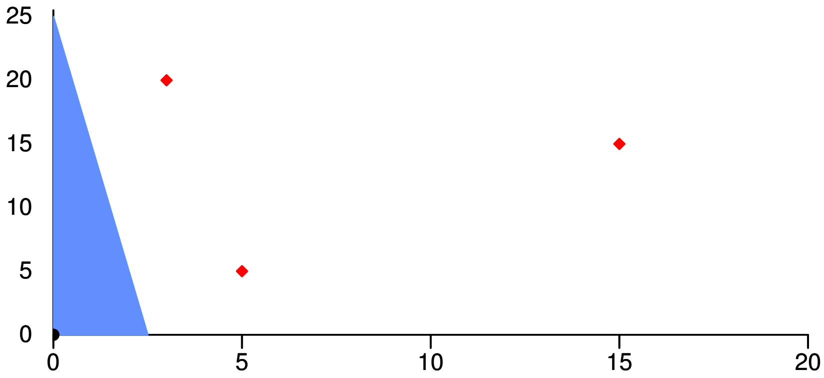



2015-06-15

CS147
└Identifying Distributions
 └Kernel Density Estimation
  └KDE Intuition (Rectangular)

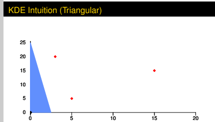


# KDE Intuition (Triangular)

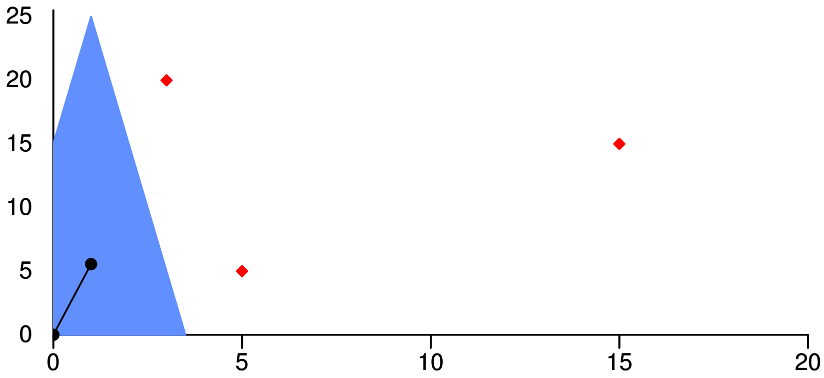






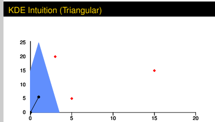
# KDE Intuition (Triangular)



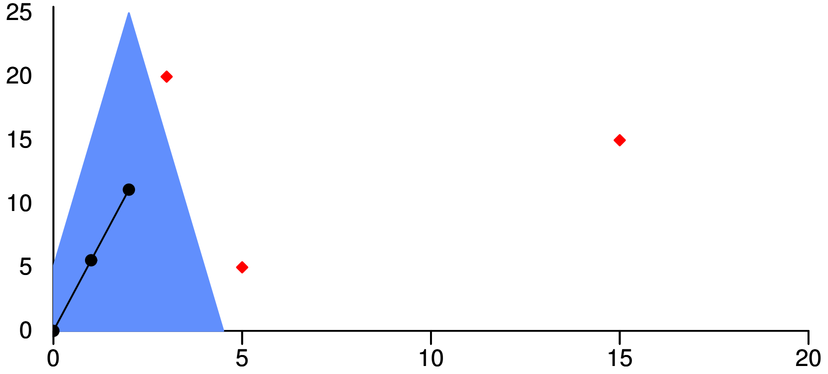



2015-06-15

CS147
└Identifying Distributions
└Kernel Density Estimation
└KDE Intuition (Triangular)

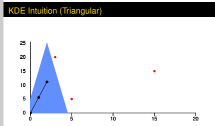


# KDE Intuition (Triangular)

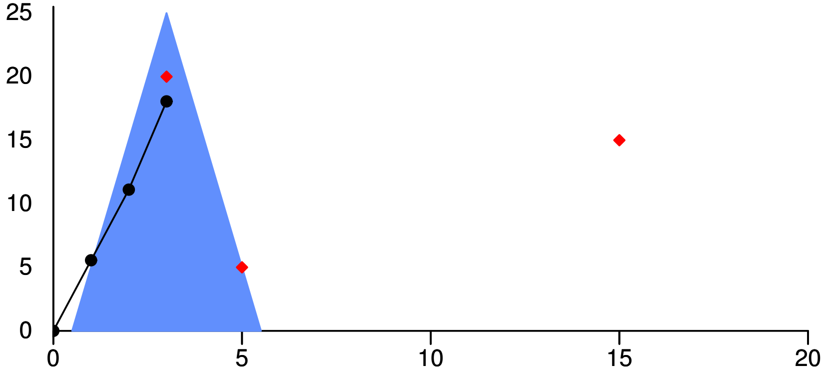






# KDE Intuition (Triangular)

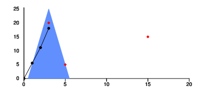




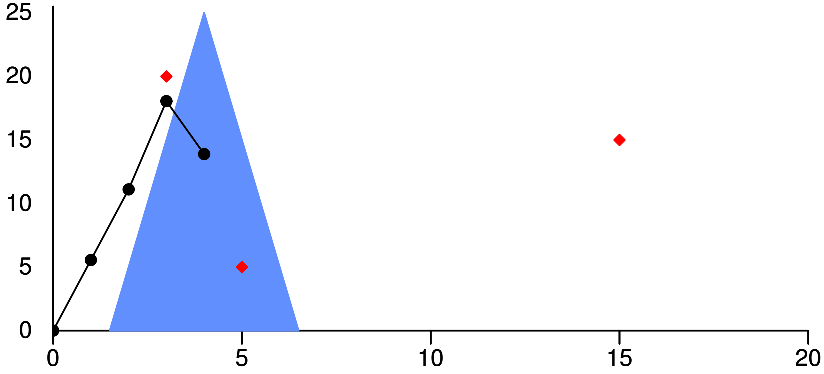

2015-06-15

CS147
└Identifying Distributions
  └Kernel Density Estimation
    └KDE Intuition (Triangular)

KDE Intuition (Triangular)

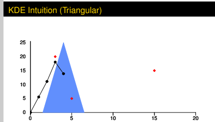


# KDE Intuition (Triangular)

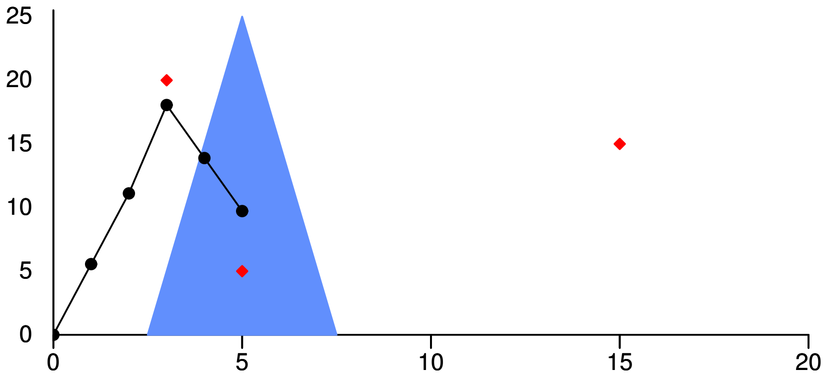






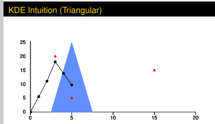
# KDE Intuition (Triangular)



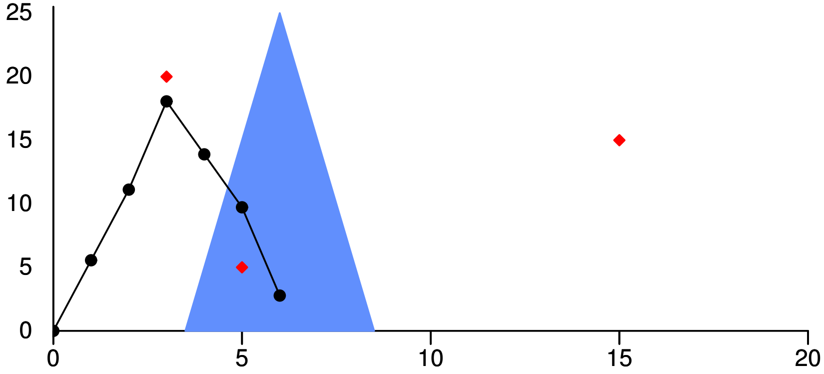



2015-06-15

CS147
└Identifying Distributions
└Kernel Density Estimation
└KDE Intuition (Triangular)



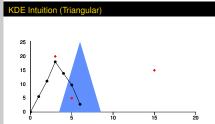
# KDE Intuition (Triangular)



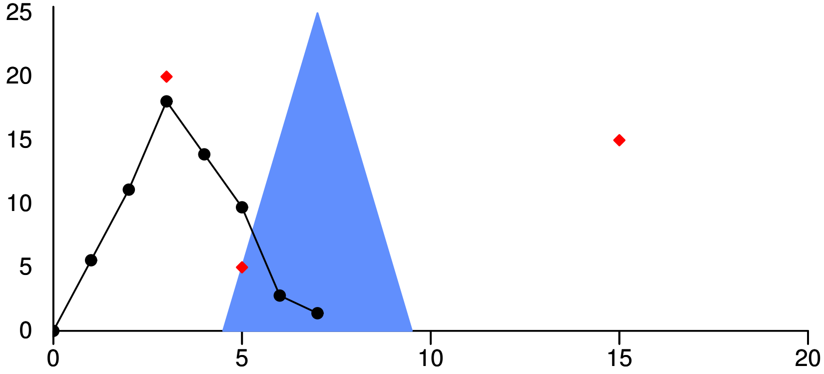



2015-06-15

CS147
└Identifying Distributions
└Kernel Density Estimation
└KDE Intuition (Triangular)

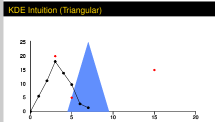


# KDE Intuition (Triangular)

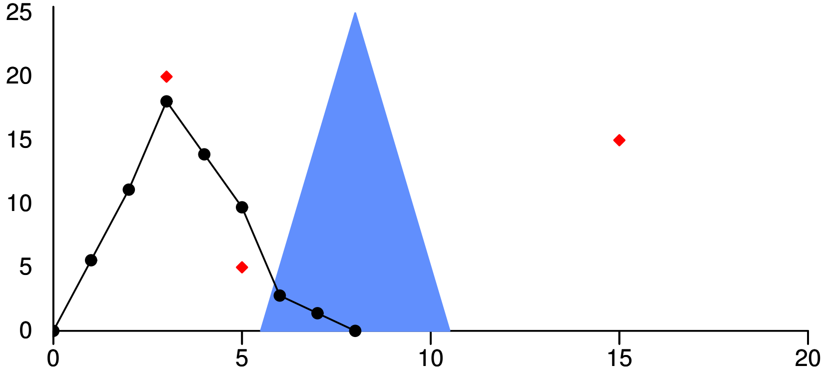






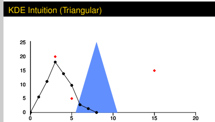



# KDE Intuition (Triangular)

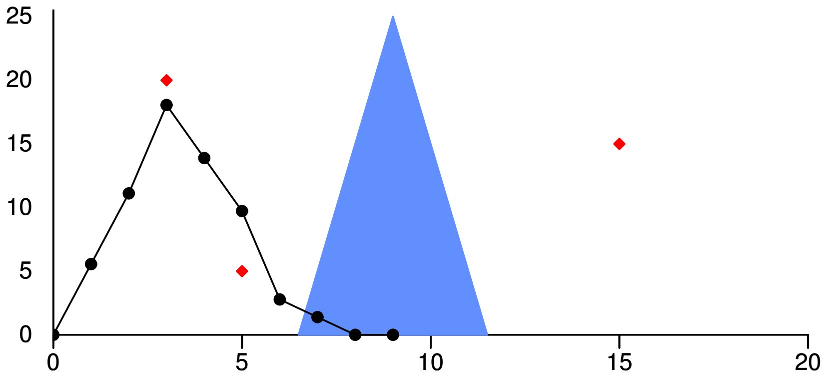






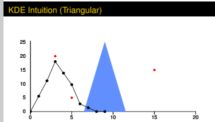



# KDE Intuition (Triangular)

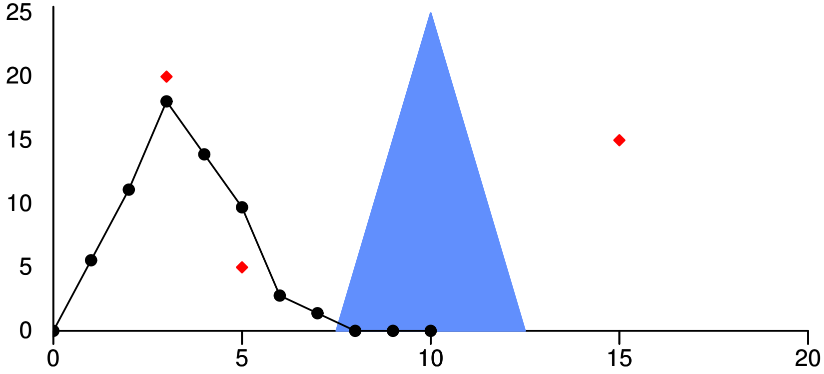






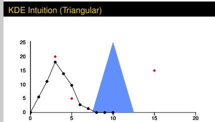



# KDE Intuition (Triangular)

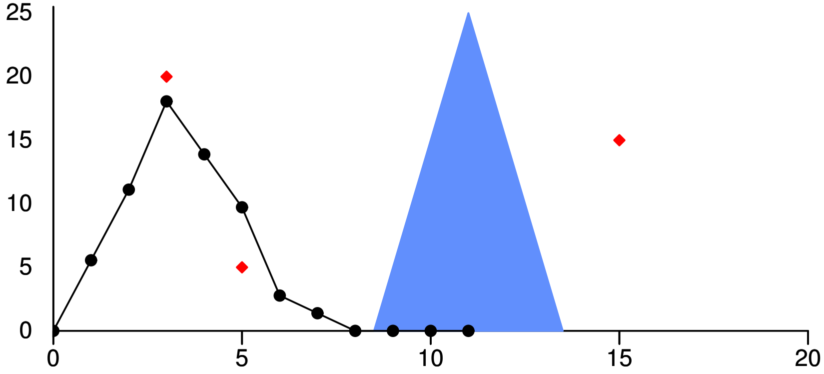






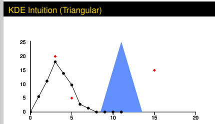



# KDE Intuition (Triangular)

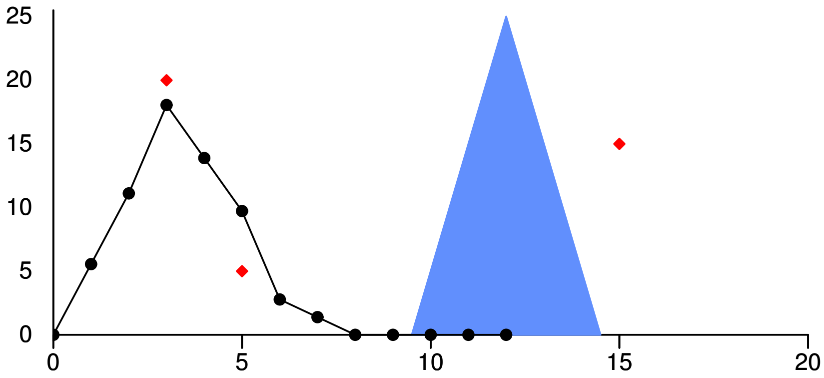






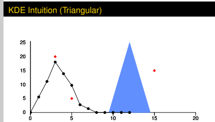



# KDE Intuition (Triangular)

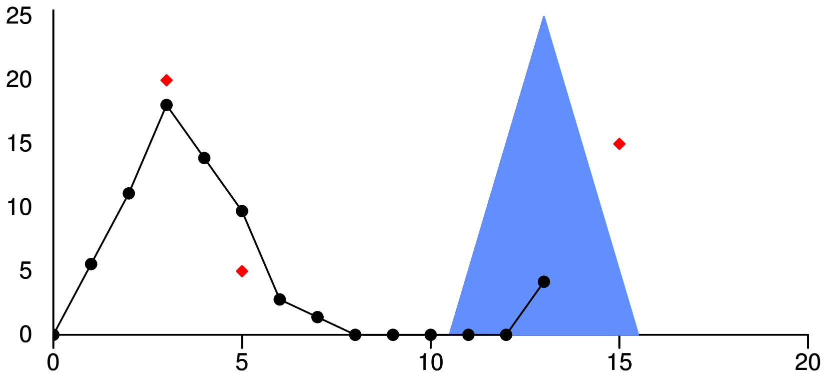






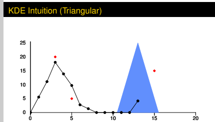



# KDE Intuition (Triangular)

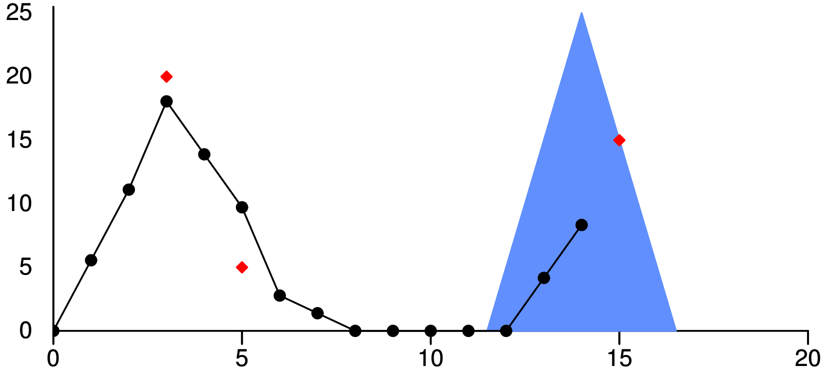






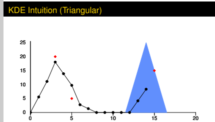



# KDE Intuition (Triangular)

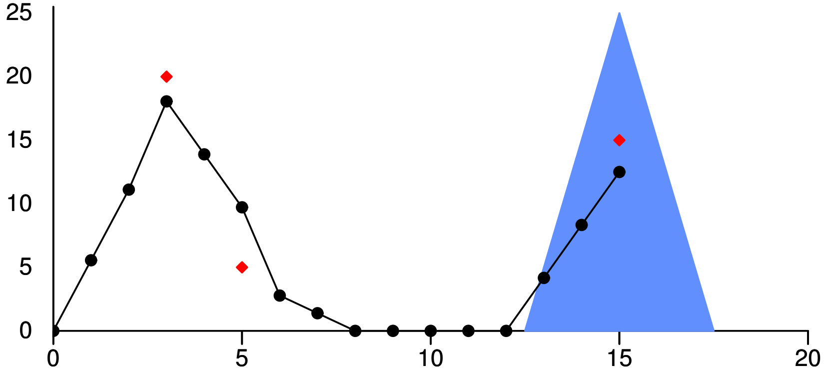






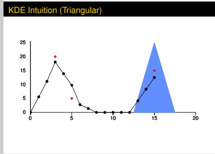
# KDE Intuition (Triangular)



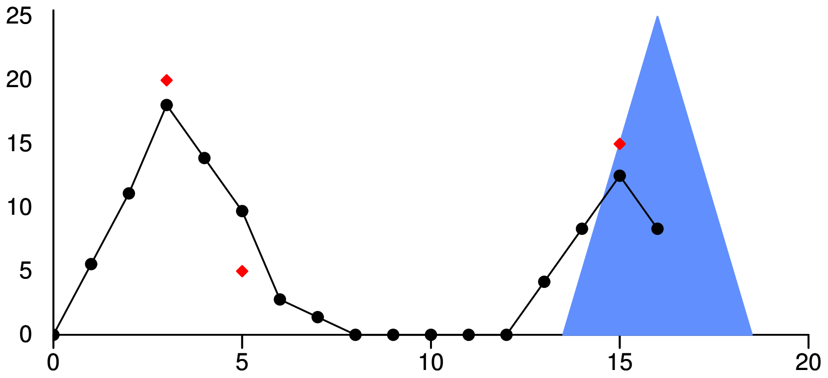



2015-06-15

CS147
└Identifying Distributions
 └Kernel Density Estimation
  └KDE Intuition (Triangular)

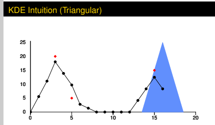


# KDE Intuition (Triangular)

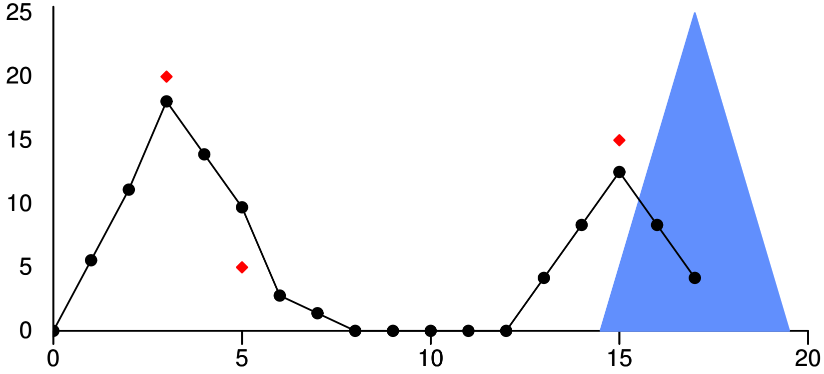






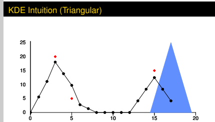
# KDE Intuition (Triangular)



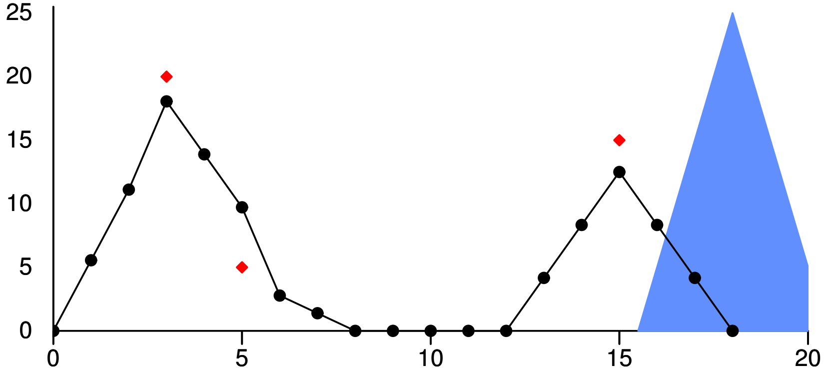



2015-06-15

CS147
└Identifying Distributions
└Kernel Density Estimation
└KDE Intuition (Triangular)

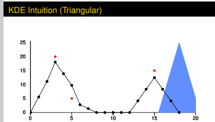


# KDE Intuition (Triangular)

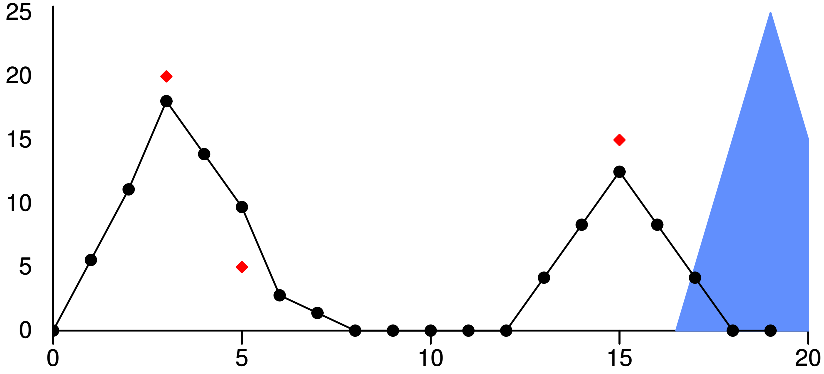






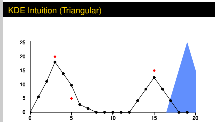



# KDE Intuition (Triangular)

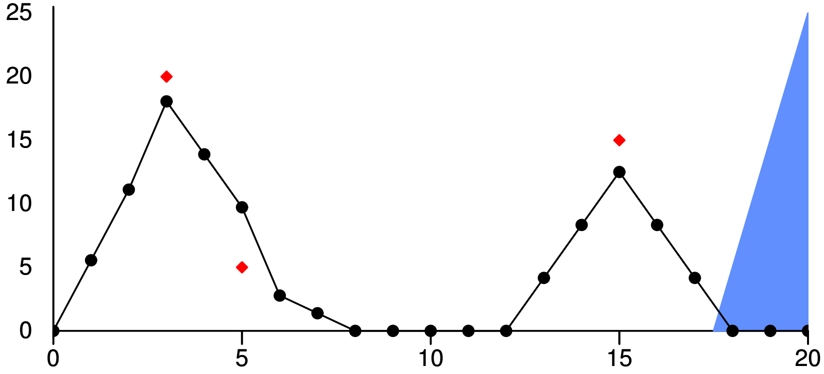






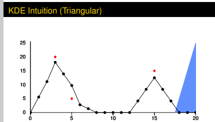



# KDE Intuition (Triangular)



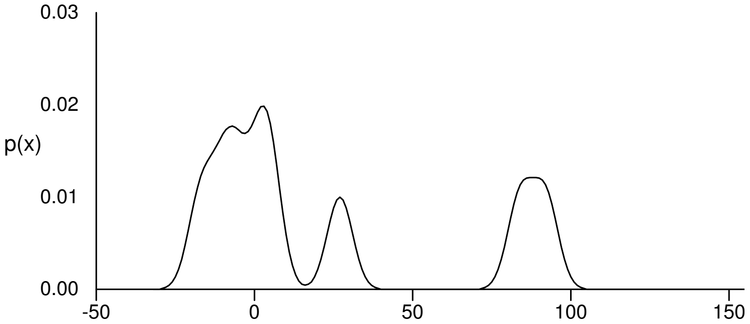



2015-06-15
CS147
Identifying Distributions
Kernel Density Estimation
KDE Intuition (Triangular)



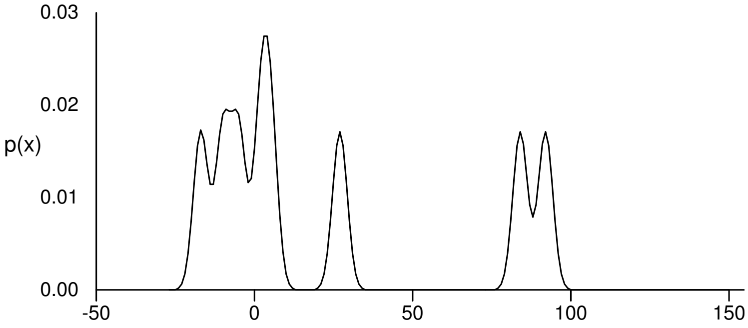
# KDE Example

- Sample data set: -17, -10, -4.8, 2, 5.4, 27, 84.3, 92, 445, 2056
- One observation per sample
- KDE with Gaussian window (RHS dropped):



# KDE Example #2

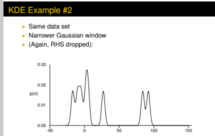
- Same data set
- Narrower Gaussian window
- (Again, RHS dropped):






2015-06-15

CS147
└Identifying Distributions
└Kernel Density Estimation
└KDE Example #2



# Quantile-Quantile Plots

- More suitable than KDE for small data sets
- Basically, guess a distribution
- Plot where quantiles of data should fall in that distribution
  - Against where they actually fall
- If plot is close to linear, data closely matches guessed distribution





Quantile-Quantile Plots

- More suitable than KDE for small data sets
- Basically, guess a distribution
- Plot where quantiles of data should fall in that distribution
  - Against where they actually fall
- If plot is close to linear, data closely matches guessed distribution

# Obtaining Theoretical Quantiles

- Need to determine where quantiles should fall for a particular distribution
- Requires inverting CDF for that distribution
  - Then determining quantiles for observed points
  - Then plugging quantiles into inverted CDF





Obtaining Theoretical Quantiles
- Need to determine where quantiles should fall for a particular distribution
- Requires inverting CDF for that distribution
  - Then determining quantiles for observed points
  - Then plugging quantiles into inverted CDF

# Inverting a Distribution

- Many common distributions have already been inverted (how convenient...)
- For others that are hard to invert, tables and approximations often available (nearly as convenient)





Inverting a Distribution

- Many common distributions have already been inverted (how convenient...)
- For others that are hard to invert, tables and approximations often available (nearly as convenient)

# Is Our Sample Data Set Normally Distributed?

- Our data set was -17, -10, -4.8, 2, 5.4, 27, 84.3, 92, 445, 2056
- Does this match normal distribution?
- Normal distribution doesn't invert nicely
  - But there is an approximation:

$$x_i = 4.91 \left( q_i^{0.14} - (1 - q_i)^{0.14} \right)$$

  - Or invert numerically





Is Our Sample Data Set Normally Distributed?
- Our data set was -17, -10, -4.8, 2, 5.4, 27, 84.3, 92, 445, 2056
- Does this match normal distribution?
- Normal distribution doesn't invert nicely
  - But there is an approximation:
$$x_i = 4.91 \left( q_i^{0.14} - (1 - q_i)^{0.14} \right)$$
  - Or invert numerically

# Data For Example Normal Quantile-Quantile Plot

| $i$ | $q_i$ | $y_i$ | $x_i$ |
|---|---|---|---|
| 1 | 0.05 | -17.0 | -1.64684 |
| 2 | 0.15 | -10.0 | -1.03481 |
| 3 | 0.25 | -4.8 | -0.67234 |
| 4 | 0.35 | 2.0 | -0.38375 |
| 5 | 0.45 | 5.4 | -0.12510 |
| 6 | 0.55 | 27.0 | 0.12510 |
| 7 | 0.65 | 84.3 | 0.38375 |
| 8 | 0.75 | 92.0 | 0.67234 |
| 9 | 0.85 | 445.0 | 1.03481 |
| 10 | 0.95 | 2056.0 | 1.64684 |



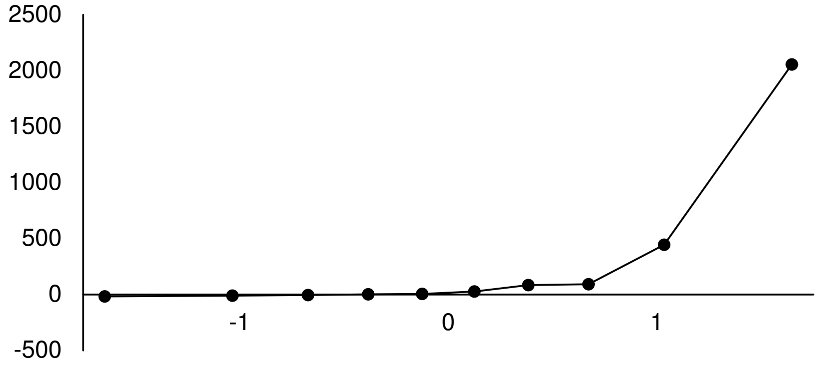
2015-06-15



Data For Example Normal Quantile-Quantile Plot

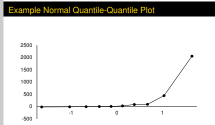
| $i$ | $q_i$ | $y_i$ | $x_i$ |
|---|---|---|---|
| 1 | 0.05 | -17.0 | -1.64684 |
| 2 | 0.15 | -10.0 | -1.03481 |
| 3 | 0.25 | -4.8 | -0.67234 |
| 4 | 0.35 | 2.0 | -0.38375 |
| 5 | 0.45 | 5.4 | -0.12510 |
| 6 | 0.55 | 27.0 | 0.12510 |
| 7 | 0.65 | 84.3 | 0.38375 |
| 8 | 0.75 | 92.0 | 0.67234 |
| 9 | 0.85 | 445.0 | 1.03481 |
| 10 | 0.95 | 2056.0 | 1.64684 |

# Example Normal Quantile-Quantile Plot







# Analysis

- Definitely not normal
  - Because it isn't linear
  - Tail at high end is too long for normal
- But perhaps the lower part of graph is normal?

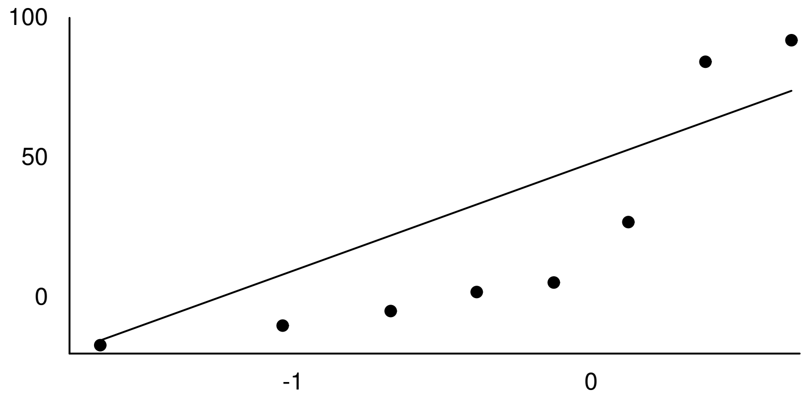




Analysis

- Definitely not normal
  - Because it isn't linear
  - Tail at high end is too long for normal
- But perhaps the lower part of graph is normal?

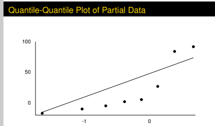
# Quantile-Quantile Plot of Partial Data






2015-06-15

CS147
└Identifying Distributions
  └Quantile-Quantile Plots
    └Quantile-Quantile Plot of Partial Data



# Analysis of Partial Data Plot

- Again, at highest points it doesn't fit normal distribution
- But at lower points it fits somewhat well
- So, again, this distribution looks like normal with longer tail to right





Analysis of Partial Data Plot

- Again, at highest points it doesn't fit normal distribution
- But at lower points it fits somewhat well
- So, again, this distribution looks like normal with longer tail to right

# Analysis of Partial Data Plot

- Again, at highest points it doesn't fit normal distribution
- But at lower points it fits somewhat well
- So, again, this distribution looks like normal with longer tail to right
- (Really need more data points)





Analysis of Partial Data Plot

- Again, at highest points it doesn't fit normal distribution
- But at lower points it fits somewhat well
- So, again, this distribution looks like normal with longer tail to right
- (Really need more data points)

# Analysis of Partial Data Plot

- Again, at highest points it doesn't fit normal distribution
- But at lower points it fits somewhat well
- So, again, this distribution looks like normal with longer tail to right
- (Really need more data points)
- You can keep this up for a good, long time



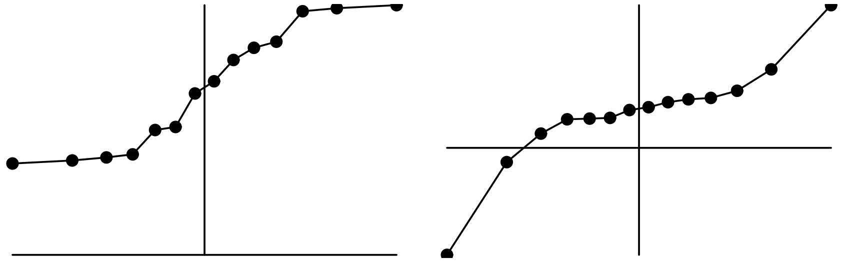


Analysis of Partial Data Plot

- Again, at highest points it doesn't fit normal distribution
- But at lower points it fits somewhat well
- So, again, this distribution looks like normal with longer tail to right
- (Really need more data points)
- You can keep this up for a good, long time

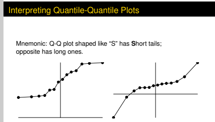
# Interpreting Quantile-Quantile Plots

Mnemonic: Q-Q plot shaped like “S” has **S**hort tails; opposite has long ones.









# What is a Sample?

- How tall is a human?
  - Could measure every person in the world
  - Or could measure everyone in this room
- Population has *parameters*
  - Real and meaningful
- Sample has *statistics*
  - Drawn from population
  - Inherently erroneous





What is a Sample?
How tall is a human?
Could measure every person in the world
Or could measure everyone in this room
Population has *parameters*
Real and meaningful
Sample has *statistics*
Drawn from population
Inherently erroneous

# Sample Statistics

- How tall is a human?
  - People in B126 have a mean height
  - People in Edwards have a different mean
- Sample mean is *itself* a random variable
  - Has own distribution





Sample Statistics

- How tall is a human?
  - People in B126 have a mean height
  - People in Edwards have a different mean
- Sample mean is *itself* a random variable
  - Has own distribution

# Estimating Population from Samples

- How tall is a human?
  - Measure everybody in this room
  - Calculate sample mean $\overline{x}$
  - Assume population mean $\mu$ equals $\overline{x}$
- What is the error in our estimate?





Estimating Population from Samples

- How tall is a human?
  - Measure everybody in this room
  - Calculate sample mean $\overline{x}$
  - Assume population mean $\mu$ equals $\overline{x}$
- What is the error in our estimate?

# Estimating Error

- Sample mean is a random variable
  - $\Rightarrow$ Mean has some distribution
  - $\therefore$ Multiple sample means have “mean of means”
- Knowing distribution of means, we can estimate error





Estimating Error

- Sample mean is a random variable
  - $\Rightarrow$ Mean has some distribution
  - $\therefore$ Multiple sample means have “mean of means”
- Knowing distribution of means, we can estimate error

# Estimating the Value of a Random Variable

- How tall is Fred?





Estimating the Value of a Random Variable

- How tall is Fred?

# Estimating the Value of a Random Variable

- How tall is Fred?
- Suppose average human height is 170 cm





Estimating the Value of a Random Variable

- How tall is Fred?
- Suppose average human height is 170 cm

# Estimating the Value of a Random Variable

- How tall is Fred?
- Suppose average human height is 170 cm
  - $\therefore$ Fred is 170 cm tall





Estimating the Value of a Random Variable

- How tall is Fred?
- Suppose average human height is 170 cm
  - ∴ Fred is 170 cm tall

# Estimating the Value of a Random Variable

- How tall is Fred?
- Suppose average human height is 170 cm
  - $\therefore$ Fred is 170 cm tall
  - Yeah, right





Estimating the Value of a Random Variable

- How tall is Fred?
- Suppose average human height is 170 cm
  - ∴ Fred is 170 cm tall
  - Yeah, right

# Estimating the Value of a Random Variable

- How tall is Fred?
- Suppose average human height is 170 cm
  - ∴ Fred is 170 cm tall
  - Yeah, right
- Safer to assume a range





Estimating the Value of a Random Variable
- How tall is Fred?
- Suppose average human height is 170 cm
  - ∴ Fred is 170 cm tall
  - Yeah, right
- Safer to assume a range

# Confidence Intervals

- How tall is Fred?





Confidence Intervals

- How tall is Fred?

# Confidence Intervals

- How tall is Fred?
  - Suppose 90% of humans are between 155 and 190 cm





Confidence Intervals

- How tall is Fred?
  - Suppose 90% of humans are between 155 and 190 cm

# Confidence Intervals

- How tall is Fred?
  - Suppose 90% of humans are between 155 and 190 cm

  ∴ Fred is between 155 and 190 cm





Confidence Intervals
- How tall is Fred?
  - Suppose 90% of humans are between 155 and 190 cm
  ∴ Fred is between 155 and 190 cm

# Confidence Intervals

- How tall is Fred?
  - Suppose 90% of humans are between 155 and 190 cm
  - ∴ Fred is between 155 and 190 cm
- We are 90% *confident* that Fred is between 155 and 190 cm





Confidence Intervals
- How tall is Fred?
  - Suppose 90% of humans are between 155 and 190 cm
  - ∴ Fred is between 155 and 190 cm
- We are 90% *confident* that Fred is between 155 and 190 cm