EXPLORATORY DATA ANALYSIS

Where all starts

Northwestern | McCORMICK SCHOOL OF ENGINEERING

Introduction

- Walk before run
- Analysis of data
  - First understand the data
- Basically getting a feel for your numbers
  - Easier to find mistakes
  - Easier to guess what actually happened
  - Easier to find odd values
- Exploratory Data Analysis
  - Summary of the data
  - Accidental and unexpected patterns

Introduction

- Exploratory Data Analysis or EDA
- Get a feel for the data
- Starting point
  - Understand data attributes
  - Data gathering process
    - Helps in data cleansing
- Data Screening
  - Check for statistical hiccups
  - Compare statistical results with (human) expectations

Statistical Concepts

- Mean - arithmetic average
- Median - middle value
- Mode - most frequent value
- Standard Deviation - variation around the mean
- Interquartile Range - range encompasses 50% of the values
- Kurtosis - tails of the data distribution
- Skewness - symmetry of the data distribution

Data Visualization Concepts

- Histogram - a bar plot where each bar represents the frequency of observations for a given range of values
- Density estimation - an estimation of the frequency distribution based on the sample data
- Quantile-quantile plot - a plot of the actual data values against a normal distribution
- Box plots - a visual representation of median, quartiles, symmetry, skewness, and outliers
- Scatter plots - a graphical display of one variable plotted on the x axis and another on the y axis
- Radial plots - plots formatted for the representation of circular data

HISTOGRAMS

Categorical Data

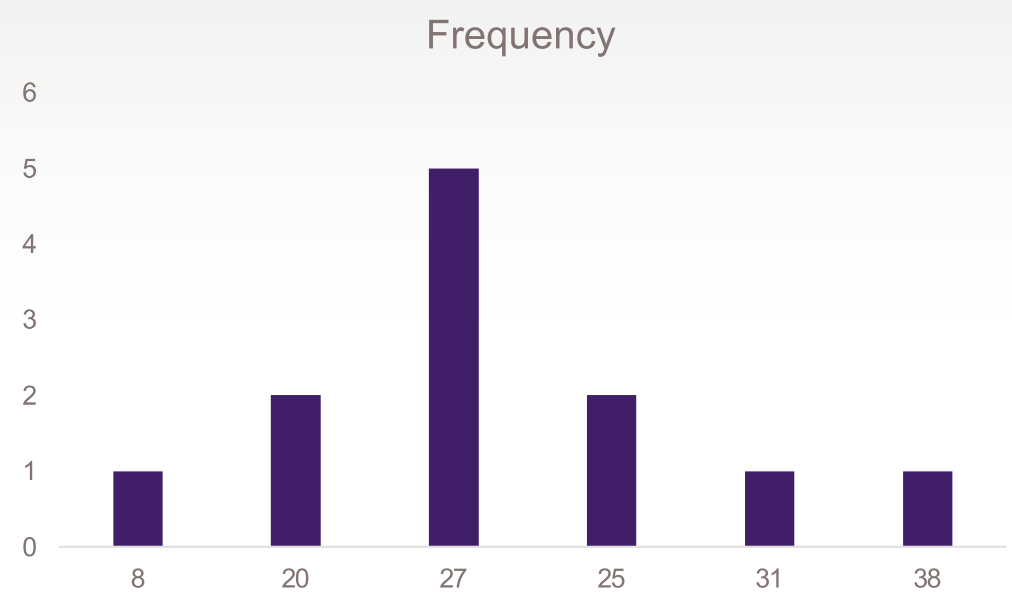
Value	Frequency
8	1
20	2
27	5
25	2
31	1
38	1

8 + 2 * 20 + 5 * 27 + 2 * 25 + 1 * 31 + 1 * 38 = 302

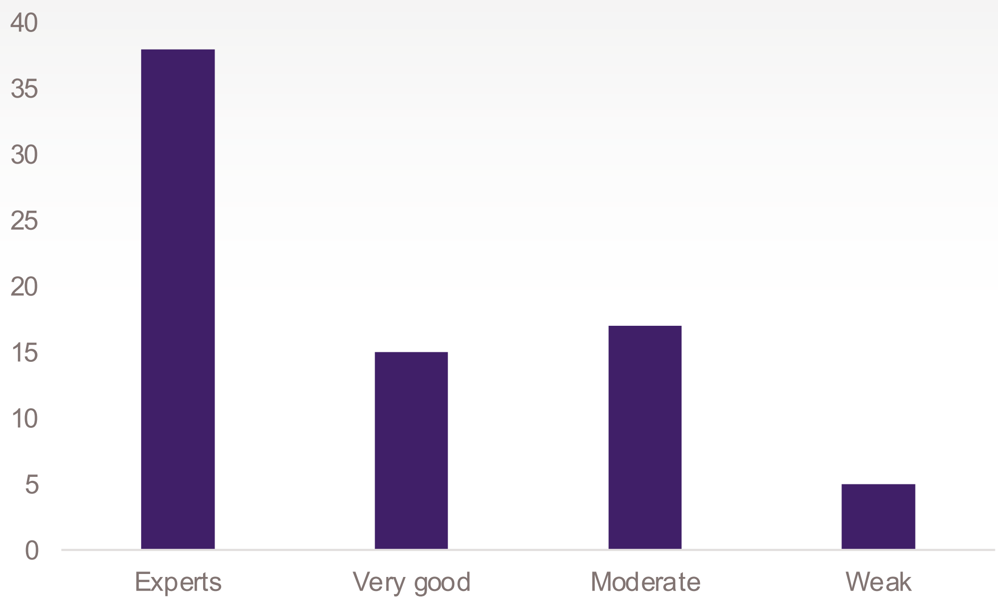
Average = 302/12=25

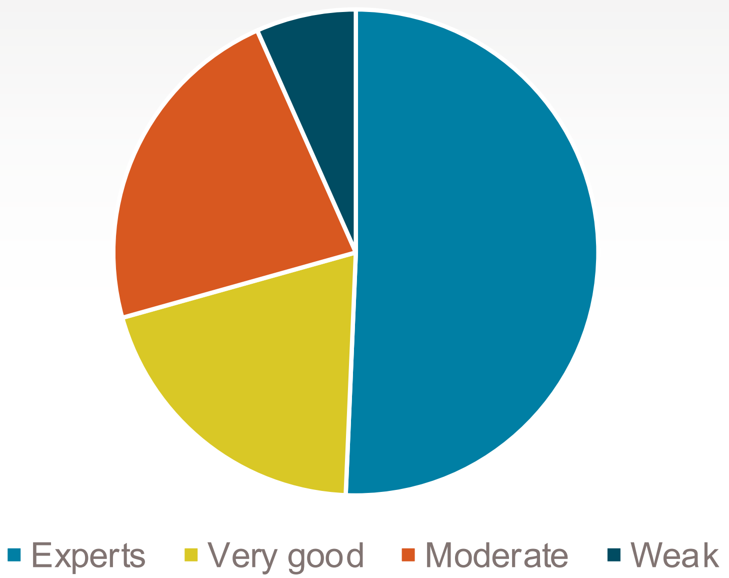
Relative Frequency Histogram

- Make a relative frequency histogram
- Lose no richness in the data
- Allows you to spot oddities



Pie Charts


40
35
30
25
20
15
10
5
0
Experts
Very good
Moderate
Weak

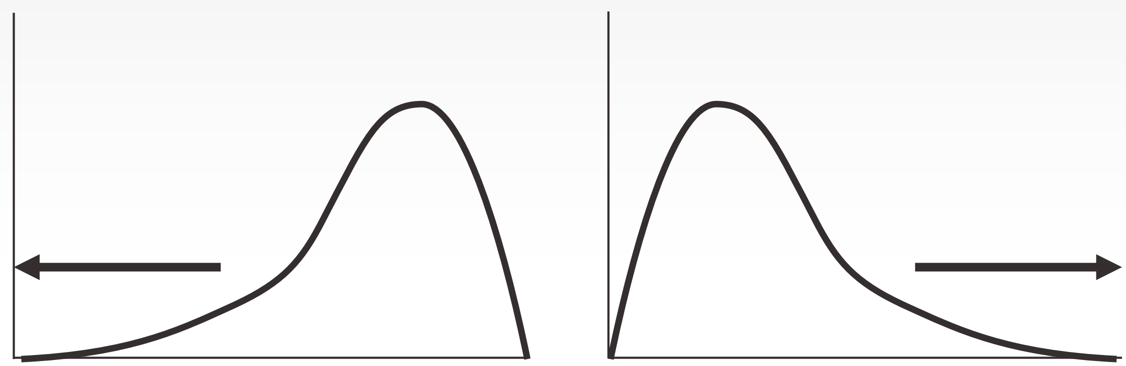

Experts
Very good
Moderate
Weak

“Continuous” Data

- Categorical data yield bar graph
- Continuous data
  - On computer everything is discrete
  - Create buckets
  - Calculate frequency of each bucket
  - Display histogram based on buckets and frequencies
- Quantitative variables

Histogram of Quantitative Variables

- Central tendency
- Spread
- Shape
  - Symmetrical or asymmetrical
- Unimodal or bimodal
- Uniform

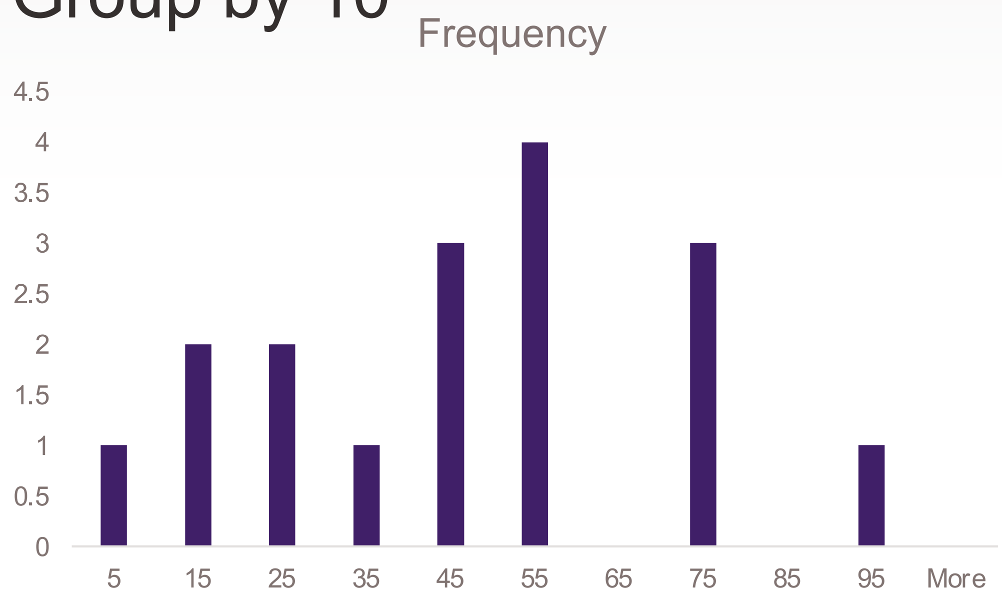


Negative Skew
Elongated tail at the **left**

Positive Skew
Elongated tail at the **right**

Example

- Number of goals scored per year by Mario Lemieux
  - 43 48 54 70 85 45 19 44 69 17 69 50 35 6 28 1 7
- Group by 10



Caveat

- Make sure the scale makes sense
- Especially the y axis
- Beware of too coarse bucketing
  - Distorts the picture
  - Special care with small amount of data
  - Histograms lose some richness of data
- Lesser problem when a lot of data available

Stemplot: Stem-and-Leaf

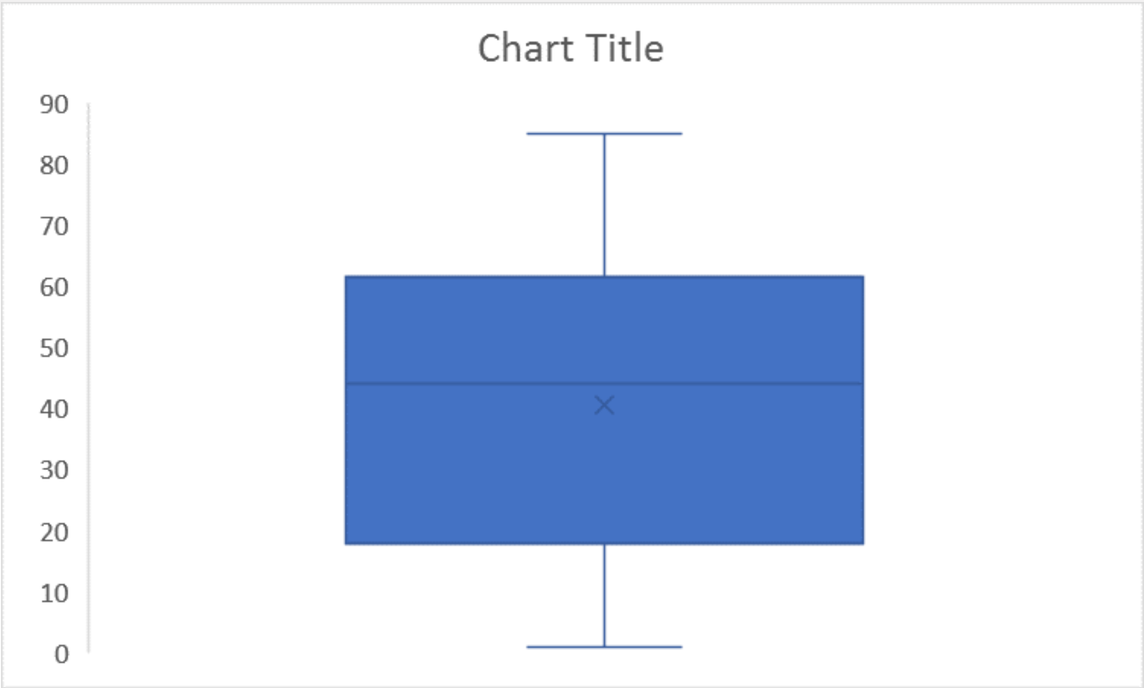
0	1	6	7	
1	7	9		
2	8			
3	5			
4	3	4	5	8
5	0	4		
6	9	9		
7	0			
8	5			

- Interpret as histogram
- Easy to spot outliers
- Preserves data
- Easy to get the middle or 50th percentile
  - 44 in this case
  - Median

The Five Number Summary

- Median
- Quartiles are the 25th and 75th percentiles
  - First quartile
    - 17 in the example
  - Third quartile
    - 54 in the example
  - Halfway between the minimum and the median
  - Halfway between the median and the maximum
- Maximum and minimum

Box Plot



Maximum

Upper hinge

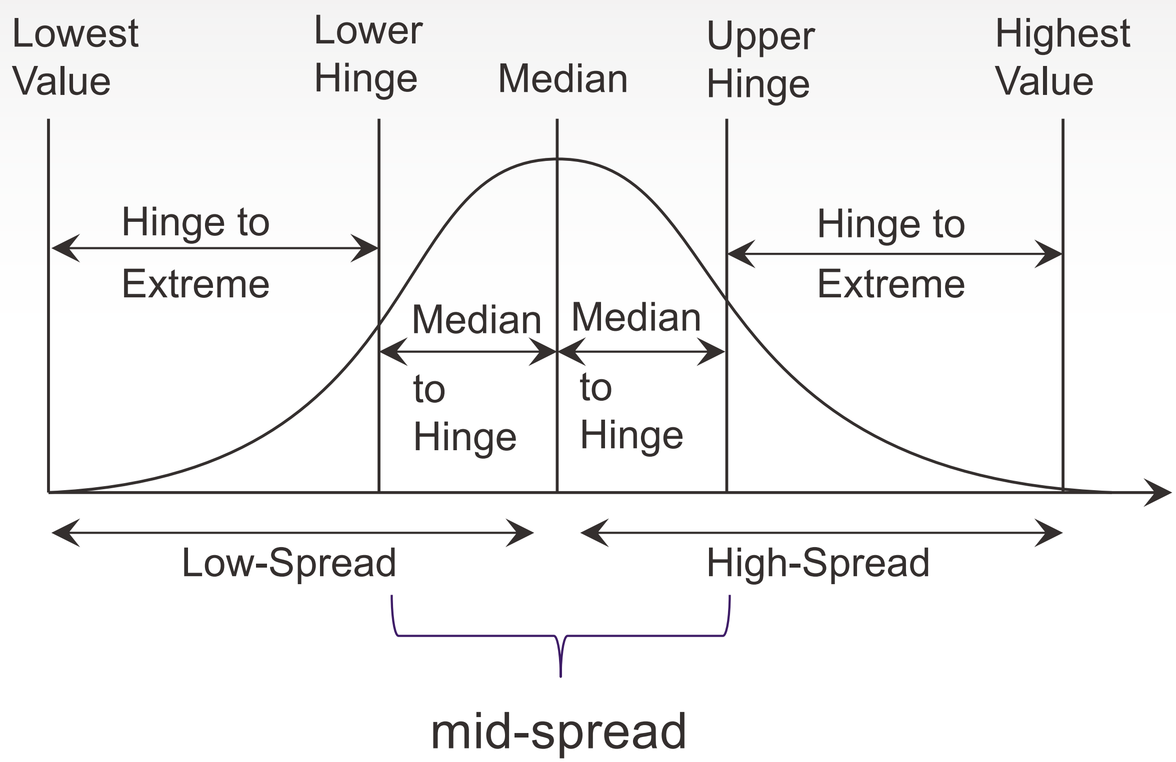
Median

Lower hinge

Minimum

Basic Concepts

- Spreads
- Hinge
  - 25% and 75% percentile
- Half of the data within mid-spread



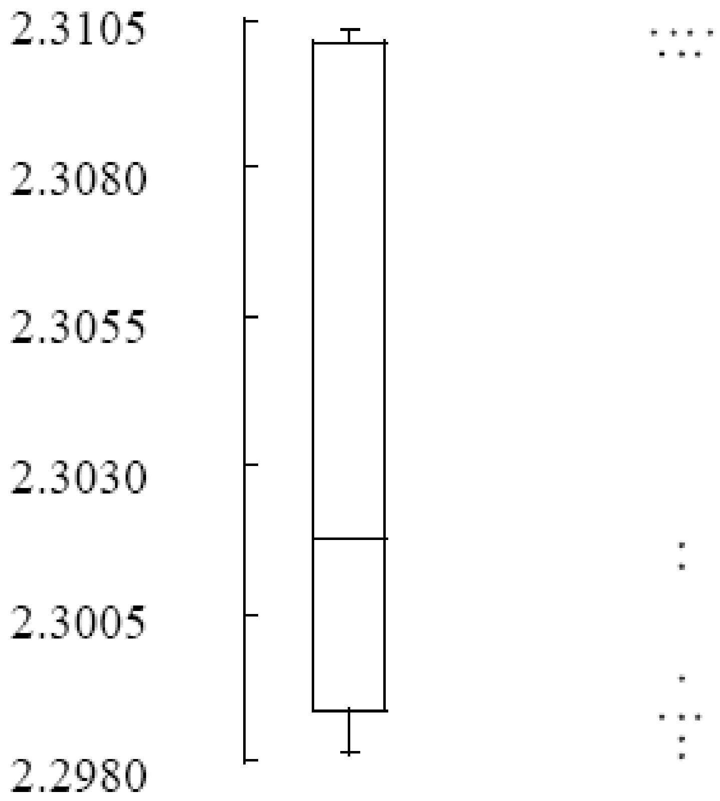
Box-and-Whisker Plot

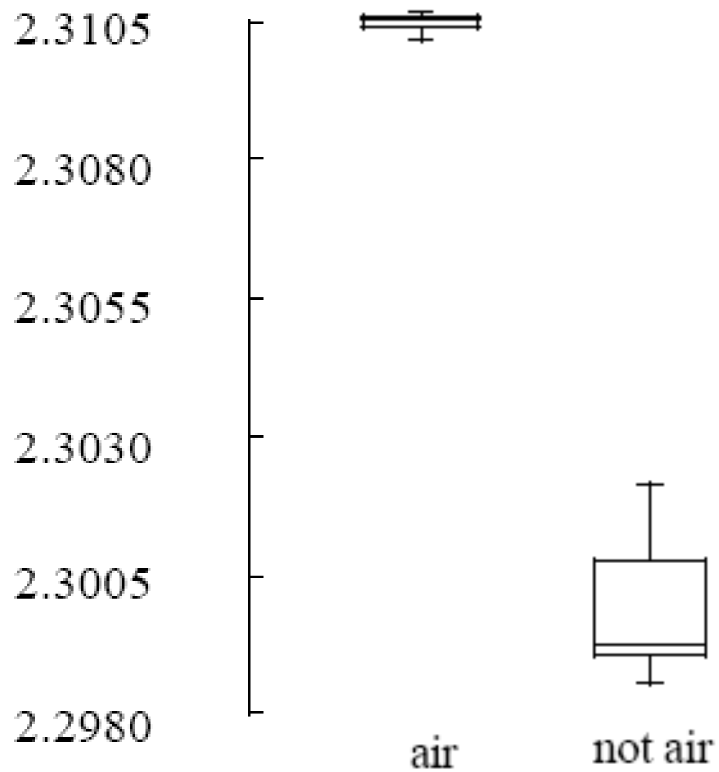
- Box always the same
- Common options for whiskers
  - Min and max
  - 2%, 98% percentiles
  - 9%, 91% percentiles
    - If data normal, equal difference
  - Lower hinge – 1.5 mid spread, upper hinge + 1.5 mid spread
    - Outside of this range usually consider as outliers
    - Sometime mark those outside of this range (outliers)

Weight of nitrogen

- Lord Rayleigh's research on the weight of nitrogen (1893)
- Two ways
  - Used a chemical compound to isolate a fixed amount of nitrogen
  - Remove oxygen from air
- Repeated this experiment 15 times

Date	*Source compound*	*Extraction method*	*Weight observed*
29.11.93	NO	hot iron	2.30143
5.12.93	NO	hot iron	2.29816
6.12.93	NO	hot iron	2.30182
8.12.93	NO	hot iron	2.29890
12.12.93	Air	hot iron	2.31017
14.12.93	Air	hot iron	2.30986
19.12.93	Air	hot iron	2.31010
22.12.93	Air	hot iron	2.31001
26.12.93	N_2O	hot iron	2.29889
28.12.93	N_2O	hot iron	2.29940
9.1.94	NH_4NO_2	hot iron	2.29849
13.1.94	NH_4NO_2	hot iron	2.29889
27.1.94	Air	ferrous hydrate	2.31024
30.1.94	Air	ferrous hydrate	2.31030
1.2.94	Air	ferrous hydrate	2.31028





- Some element present in air
  - Discovery of argon
- Graph values
  - Shows data can be divided

OUTLIERS

Outliers

- Extreme values
- Can be erroneous
- Correct
  - Investigate why
- Either case do not simply discard

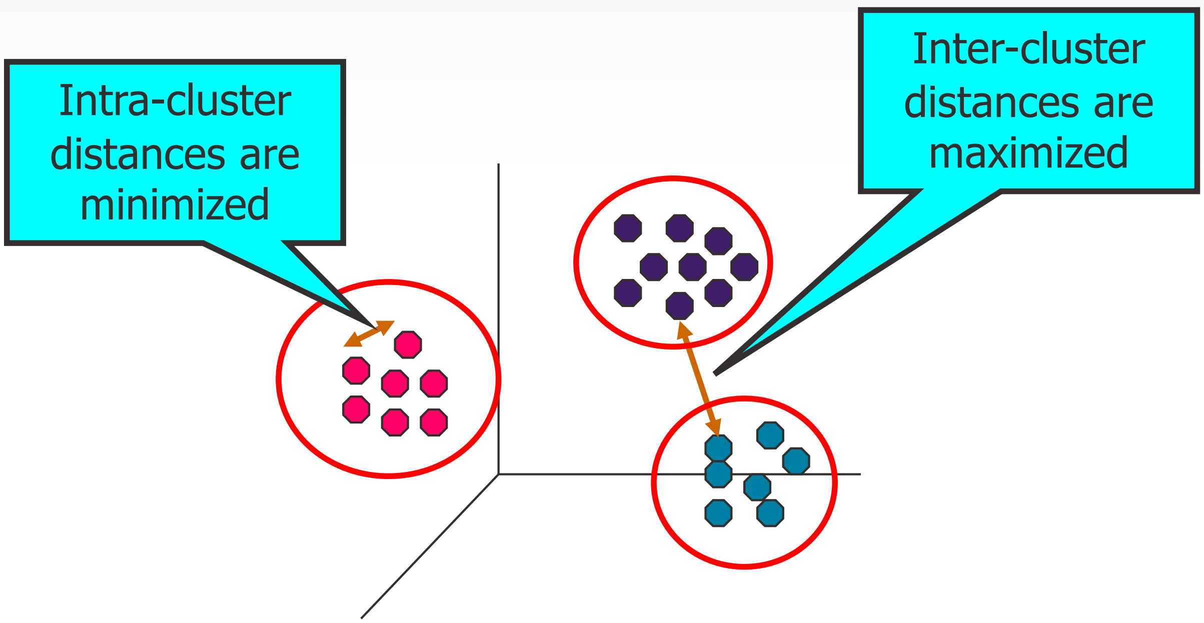
- Distort descriptive statistics
  - Mean, standard deviation
  - False skewness
    - US income

Criteria

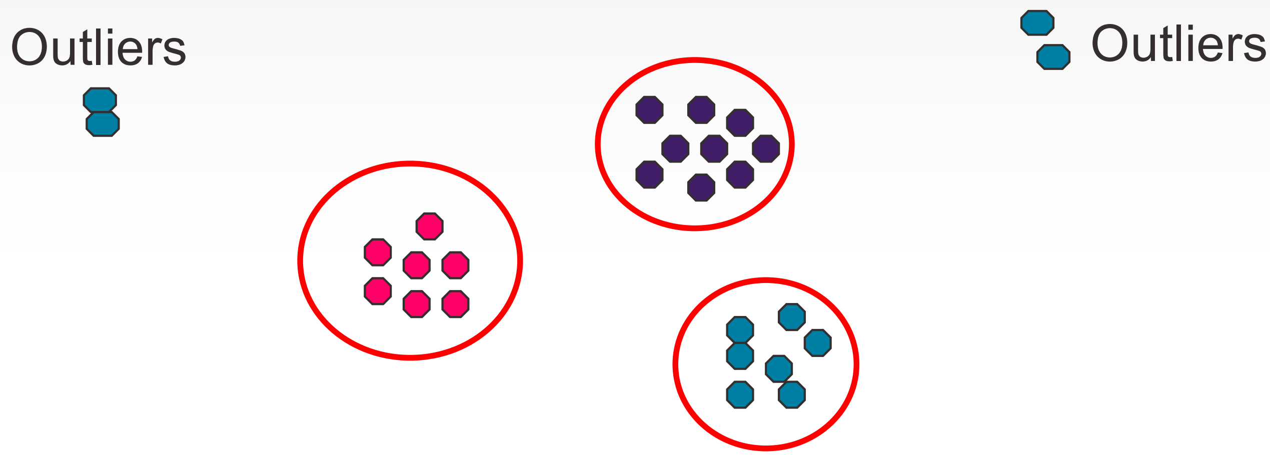
- Standard deviation
  - Outside of range [mean – k std, mean + k std]
  - k = 2, 3
  - Using standard normal same as 99.7%, 99.8%
- Percentiles
  - Lower hinge – 1.5 mid spread, upper hinge + 1.5 mid spread
  - Box-and-whisker plot

Clustering

- Finding groups of objects such that the objects in a group will be similar (or related) to one another and different from (or unrelated to) the objects in other groups

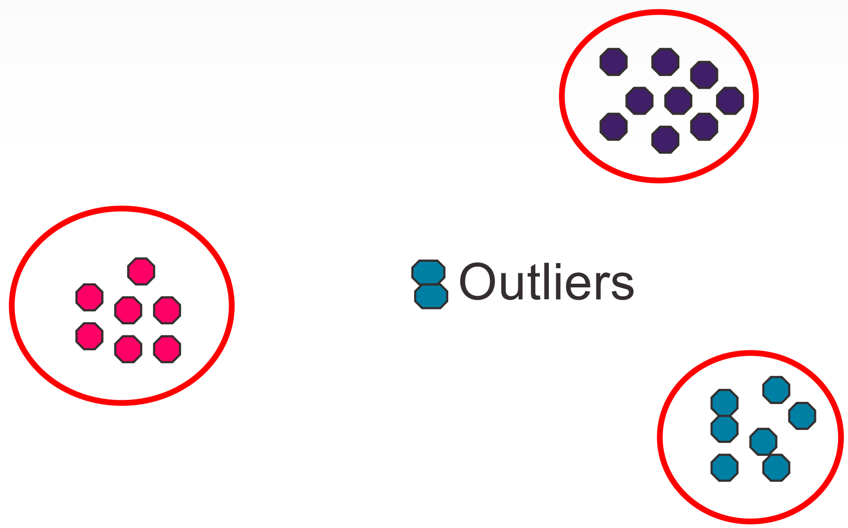


Clustering for Outlier Detection



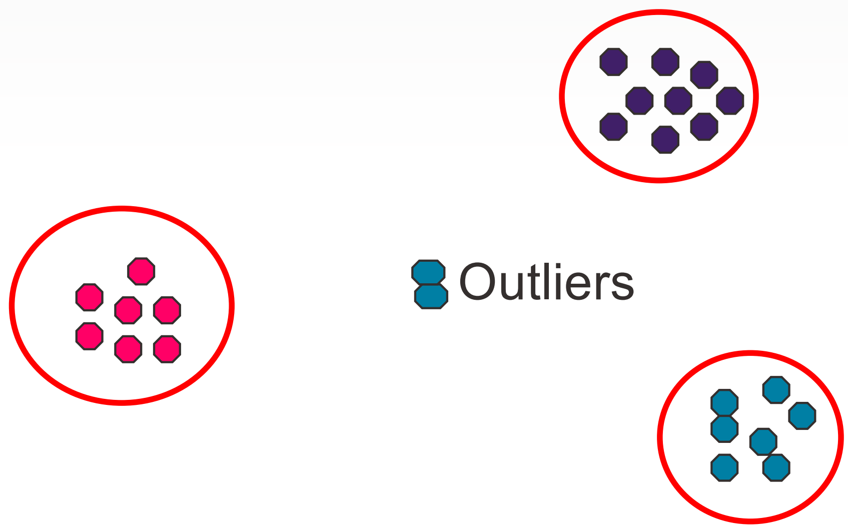
Clustering for Outlier Detection

- Why this could be better than by-feature outlier detection?
  - Outliers can be ‘in the middle of data’
  - Not detected by-feature



Clustering for Outlier Detection

- Why this could be better than by-feature outlier detection?
  - Outliers can be 'in the middle of data'
  - Not detected by-feature



Outlier Detection by Clustering

- k number of clusters
  - Challenging how to set k
- Solve clustering with requirement of k clusters
- Outliers are clusters j with
  - Number of samples in cluster $j \leq \alpha$
  - $\mathrm{Min}_{i:i \neq j}\, d(s_i,\ s_j) \geq \tau$
- Hyper parameters α, τ
  - Hard to set them
- s_k = centroid of cluster k (average of all samples in cluster k)

Algorithm

- Parameters k and l
- Loop (for all subsets of cardinality l of all samples)
  - Remove l samples (set S)
  - Find clustering with k clusters on the data set without the l samples
  - Record the total inter distance d(S)
    - Among centroids
- The l samples with the highest
  - U = inter distance all samples – inter distance without l samples
  - U = inter distance all samples – d(S)
  - Candidates for outliers
- In practice too costly to perform so many clustering steps
  - Perform only a few iterations of iterative clustering
  - Carefully select the l samples to remove

Outlier Detection on Steroids

- Start with k clusters
- Loop
  - For each sample i compute $h_i = \sum_j d(s_j, x_i)/k$
  - Sort based on h_i in non-increasing order
  - Let S be the top l samples in the order
  - Cluster to k clusters all samples without S
  - Remove S if U above threshold *h*
- Stop when some stopping criteria met
  - Running time
  - Top h_i does not change substantially

Enhancement

- Replace Euclidian distance
- Mahalanobis distance

$$d(x,y)^2 = (x-y)M(x-y)$$

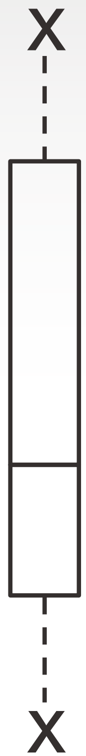
- Challenge
  - How to find M?
  - Tune on some kind of a training data set

Other Methods

- Principle component analysis
  - Variance with and without potential outliers
- Self organizing maps
  - Project multidimensional data to 2D or 3D
  - Preserve distance
  - Visually find outliers
    - Or use an automated algorithm in 2D
- t-SNE
  - Similar to self organizing maps
- One-class SVM

TAIL AND SKEWNESS

Literacy Data from 1972



0	2 3 3 3 5 5 5 5 8 8 8 8 8 8 8 8
1	0 0 0 0 0 0 0 3 3 5 5 5 8
2	0 0 0 0 0 3 3 3 5 8
3	0 0 0 2 3 5 5 5 5 5 5 8
4	1 1 3 3 5 5 6 7 9
5	0 0
6	0 0 0 1 1 2 5 5 5 5 7 8
7	0 1 1 2 2 2 3 5 5 5 6 7
8	0 0 0 1 2 4 4 5 5 6 6 7
9	0 0 1 2 5 6 8 8 8 8 8 9 9 9 9 9 9 9 9 9 9 9 9 9
10	0 0 0 0 0 0

Analysis

- Two separate distributions
  - Lower spread
    - Right skew
  - Upper spread
    - Left skew
- Skew refers to tail
- Box plot does not show the split

Kurtosis

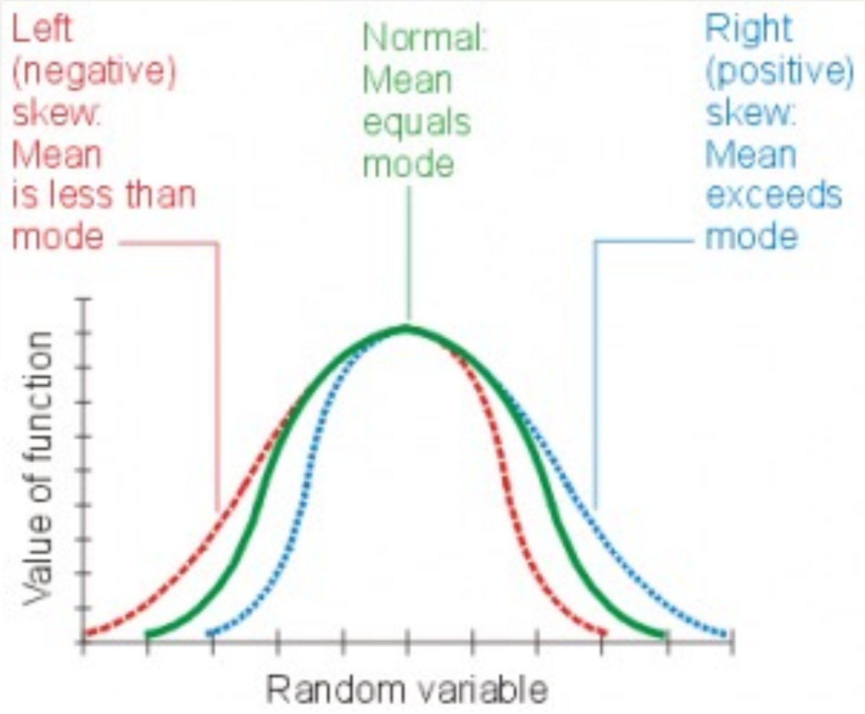
- $K(X) = E\left[\left(\frac{X-\mu}{\sigma}\right)^4\right]$
- Average of standardized data raised to the power of 4
- Standardized data in range [-1,1]
  - One standard deviation from the mean
  - Contribute almost nothing
    - Raised to power of 4
- Those outside of range contribute to Kurtosis
  - Tail extremity
- High Kurtosis implies mass concentrated in tails

Kurtosis

- Typically
  - Values larger than 5 considered high
  - Data usually not normally distributed
    - Not a test if data normally distributed
  - Can also take value 3
    - Kurtosis of standard normal
    - Sometimes value subtracted by 3

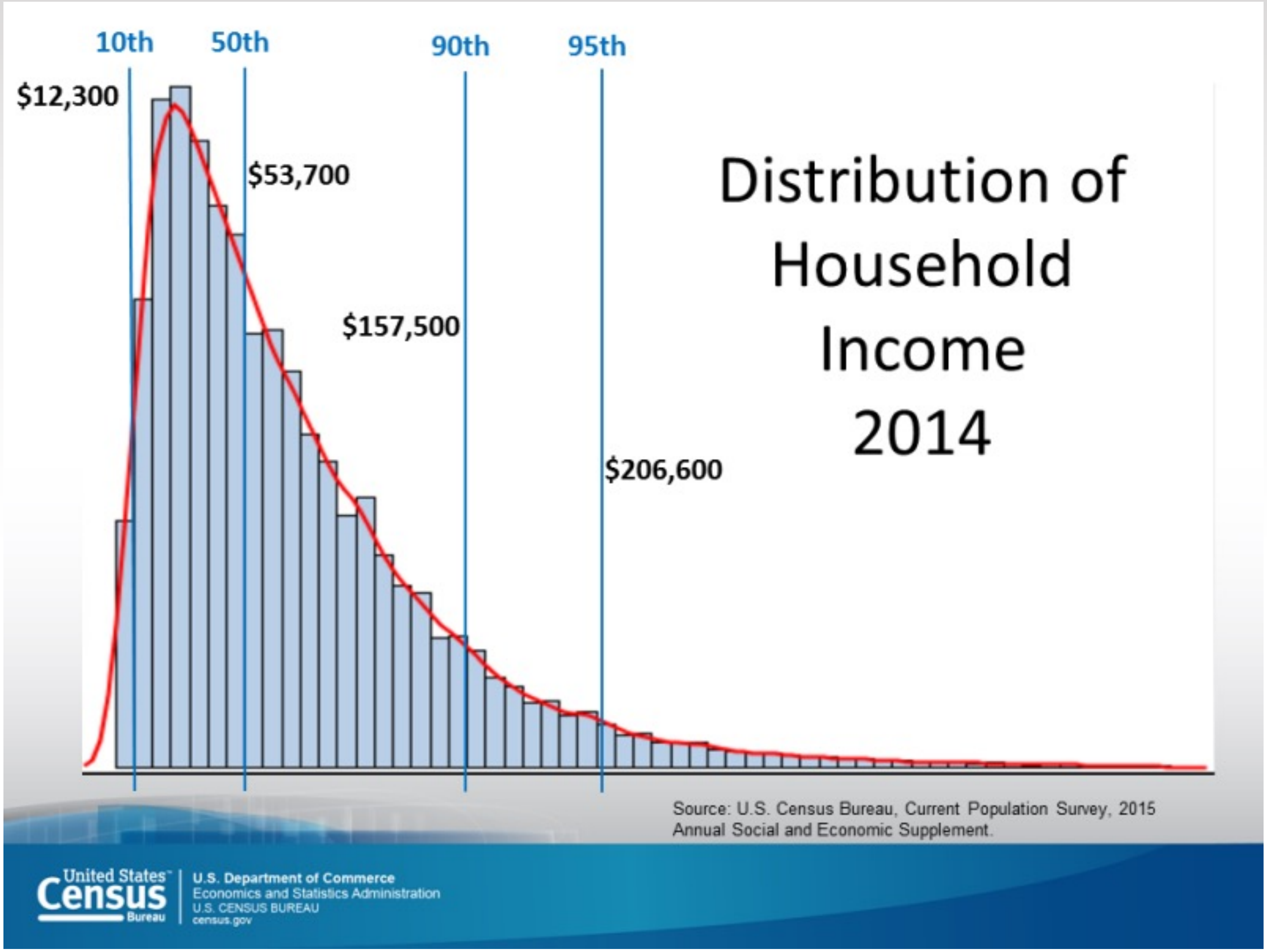
Skewness

- $\gamma_1(X) = E\left[\left(\frac{X-\mu}{\sigma}\right)^3\right]$
- Average of standardized data raised to the power of 3
- Most values >= -1
  - Values larger than 1 make it large
- Most values <= 1
  - Values larger than -1 make it large



Skewness

- Atypical values
  - Value larger than 1
  - Value smaller than -1
- Normal distribution has value 0
- Exponential has value 2
- Kurtosis and skewness can be used to measure normality
  - Other tests more comprehensive
    - Kolmogorov-Smirnov



10th
50th
90th
95th
$12,300
$53,700
$157,500
$206,600
Distribution of Household Income 2014
Source: U.S. Census Bureau, Current Population Survey, 2015 Annual Social and Economic Supplement.
United States Census Bureau
U.S. Department of Commerce
Economics and Statistics Administration
U.S. CENSUS BUREAU
census.gov

Q-Q PLOT

Equality of Distributions

- Q-Q plot
  - Quantile = percentile
  - Quantile-to-quantile
- Idea
  - Two distributions are similar if ‘all’ quantiles are similar
- For given fraction $0 \leq r \leq 1$
  - r-quantile is the value where r% of the values are below this value and 1-r% are above
  - Not a unique value for discrete distributions

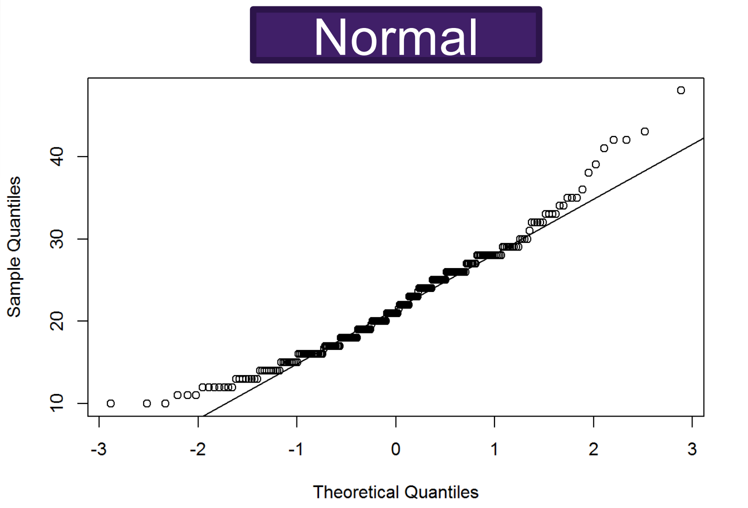
Q-Q Plot

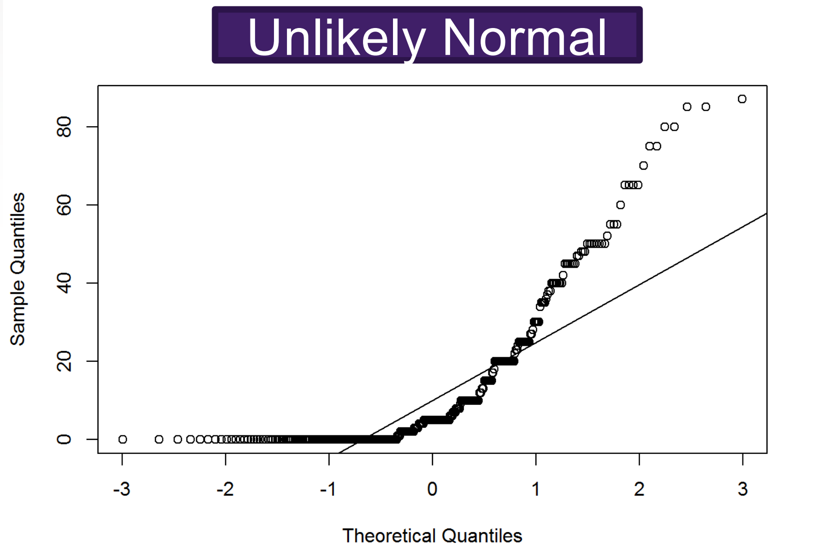
- Two distributions are the same if many quantiles are the same
- Given two distributions or data sets
  - For r = 0, .1, .2, .3, …, 1
    - Compute r-quantile based on each distribution
    - Gives point (x_r, y_r)
- Plot all points
- If distributions the same
  - The plot should be line x=y (45-degree line)

Q-Q plot

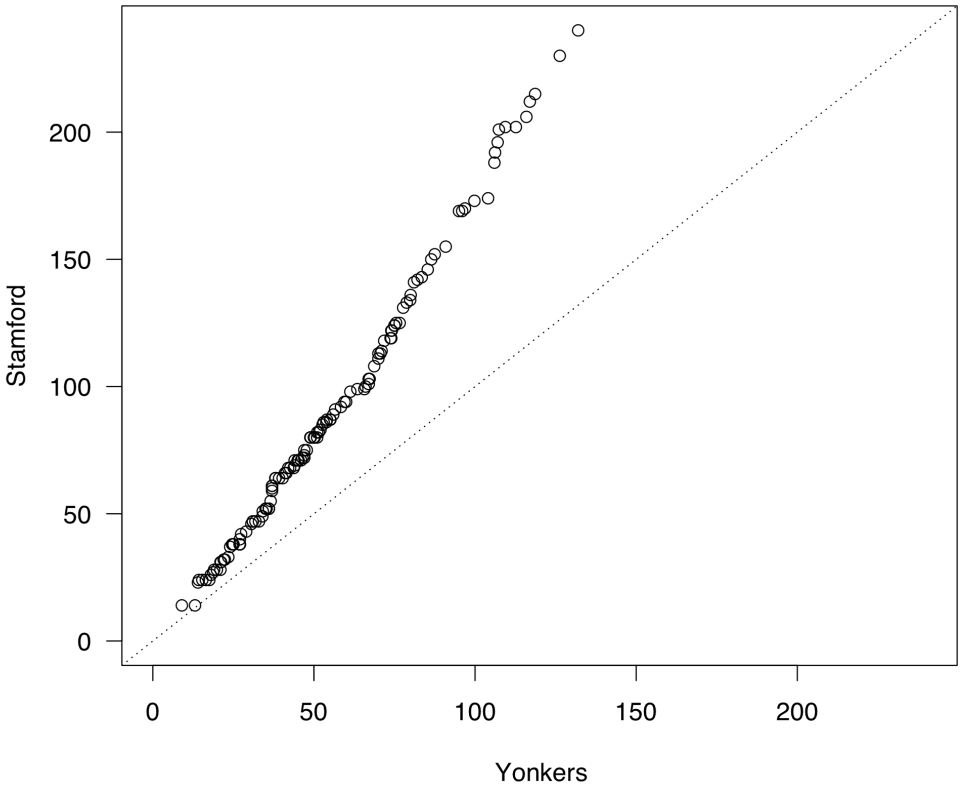
- Two equal size datasets
  - Pick all $r = \frac{i+0.5}{n}$
- Two different size datasets
  - Set r based on the larger dataset
- One data set and one theoretical distribution
  - Pick all $r = \frac{i+0.5}{n}$
  - For theoretical distribution compute the quantile
    - Invert CDF

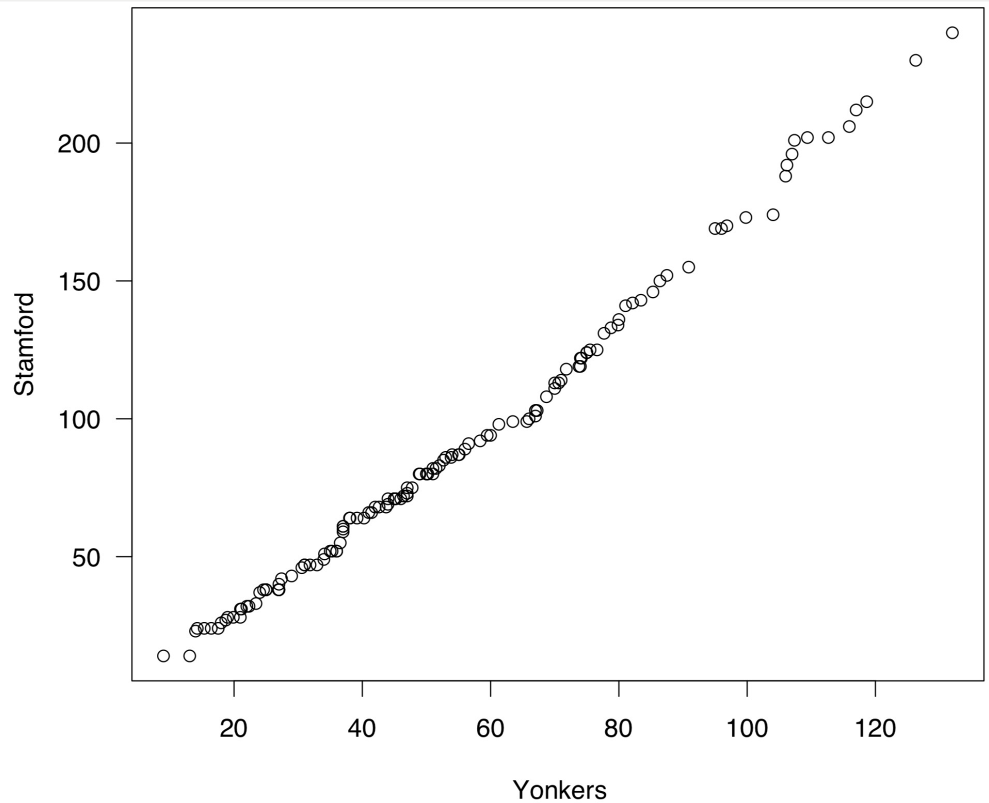
Q-Q Plot


Normal
Sample Quantiles
Theoretical Quantiles
10
20
30
40
-3
-2
-1
0
1
2
3


Unlikely Normal
Sample Quantiles
Theoretical Quantiles
0
20
40
60
80
-3
-2
-1
0
1
2
3

Q-Q plot


200
150
100
50
0
Stamford
0
50
100
150
200
Yonkers


200
150
100
50
Stamford
20
40
60
80
100
120
Yonkers