Data Exploration Part 1

Lesson 1

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

Data Exploration

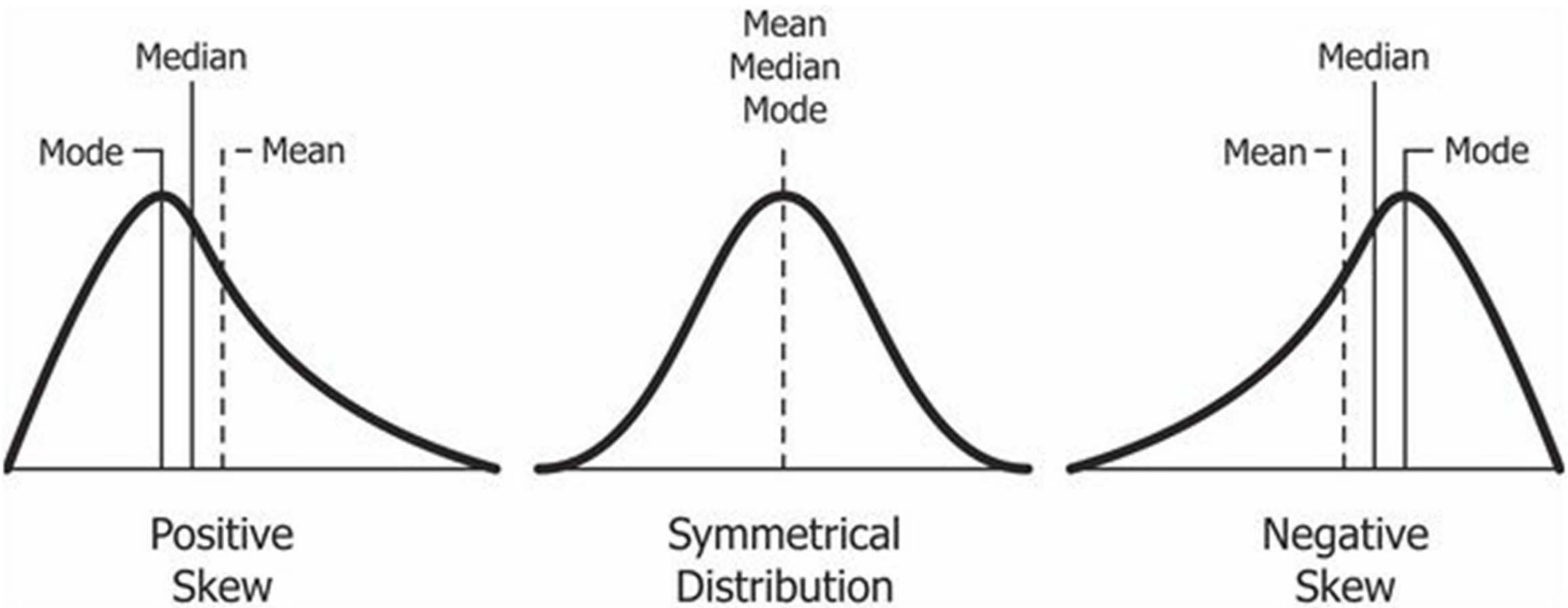
> Why data exploration?
> Need to understand relationships in data
  > How to explain relationships?
  > Which variables are dependent on other variables?
  > Which feature contain information to predict the label?
> Poor understanding of relationships leads to poor models
  > Errors in the data
  > Model based on poor understanding
  > Model based on incorrect predictors

Data Exploration (Descriptive Statistics)

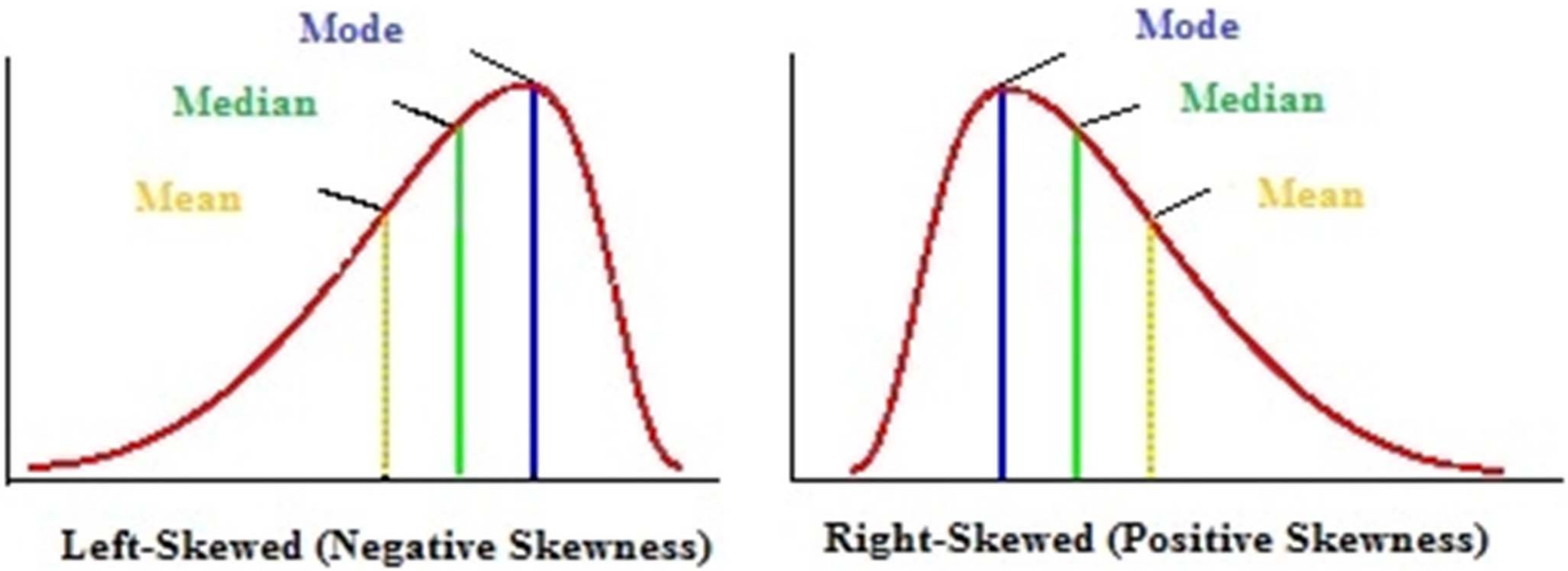
- What is it?
  - First look at your data
  - Summary Statistics
- Purpose: To gain a clear understanding of your data
  - What are the dimensions?
  - What columns are of interest?
  - Missing data?
  - Outliers?
  - Patterns?
  - Need to reformat?
  - Data types

Summary Statistics

Skew


Median
Mode
Mean
Positive
Skew
Mean
Median
Mode
Symmetrical
Distribution
Median
Mean
Mode
Negative
Skew

Skew

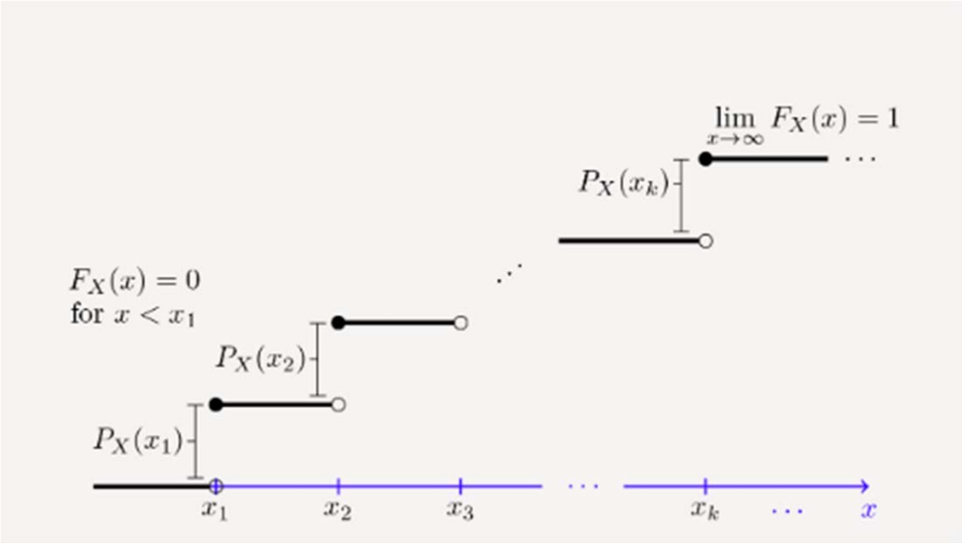

Mode
Median
Mean
Left-Skewed (Negative Skewness)
Mode
Median
Mean
Right-Skewed (Positive Skewness)

Cumulative Distribution Function

Probability that some random variable X will be less than or equal to a certain value

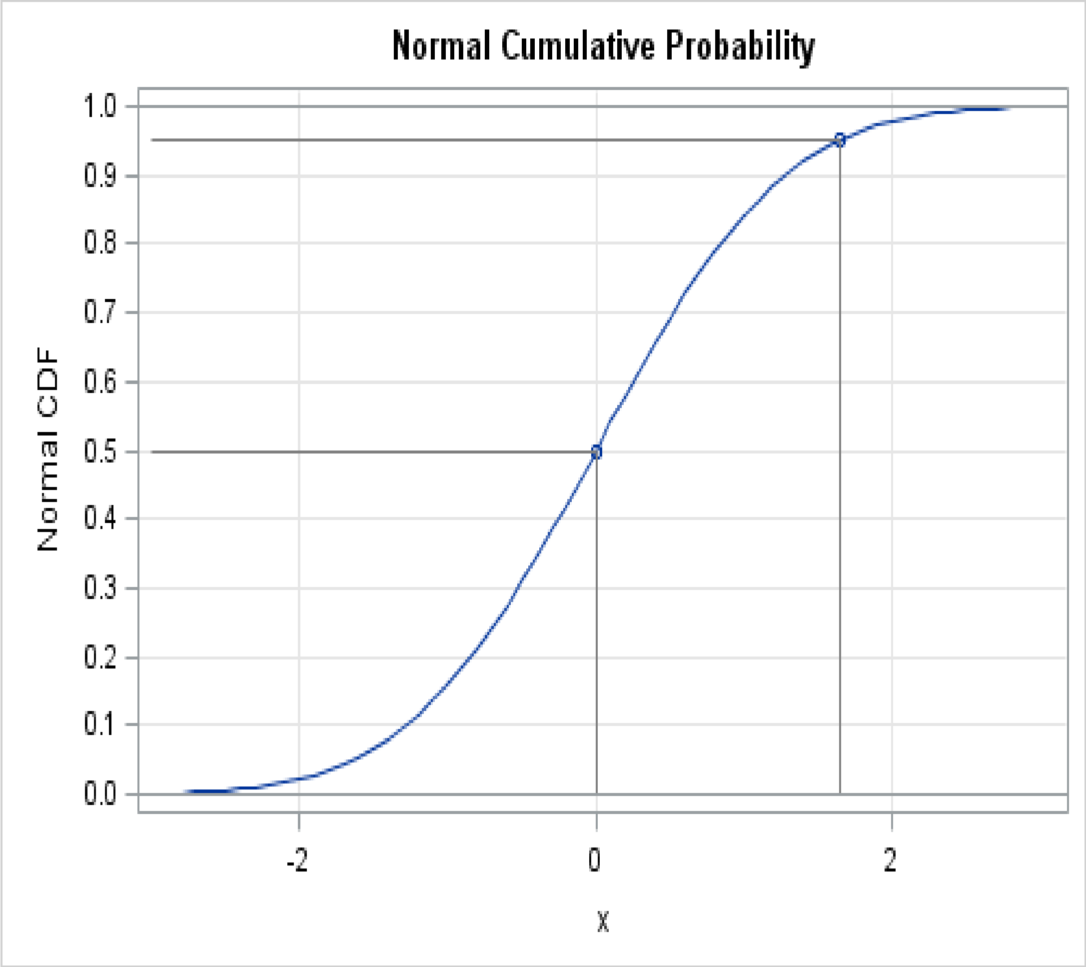
$$F(x)=P(X\leq x)$$

- > Probability, so 0 < x < 1
- > Continuous and discrete variables
- > PMF can only be used on discrete
- > Takes as input x, returns vector from [0,1] of probabilities "p"
- > Form of a staircase
- > Jumps at each x(k)



Quantiles of numerical variables

- Quantiles are inverse values of the CDF (cumulative distribution function).
- Inverse tells you what value of x would make F(x) return a value "p"
- Standard Normal: (shown in figure)
  - Quantile(0.5) = 0, means at x=0, 50% of the distribution lies to the left. (This is also the median)
  - Quantile(0.95) = 1.65



Frequency

Frequency: Counts

- Numerical and categorical variables
- Number of occurrences for an event in a fixed period
  - Ex. Number of times a gene is expressed after a medical treatment
- Modeled using Poisson distribution
  - Assume events are random and univormly distributed

Poisson Distribution Formula

$$P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

where

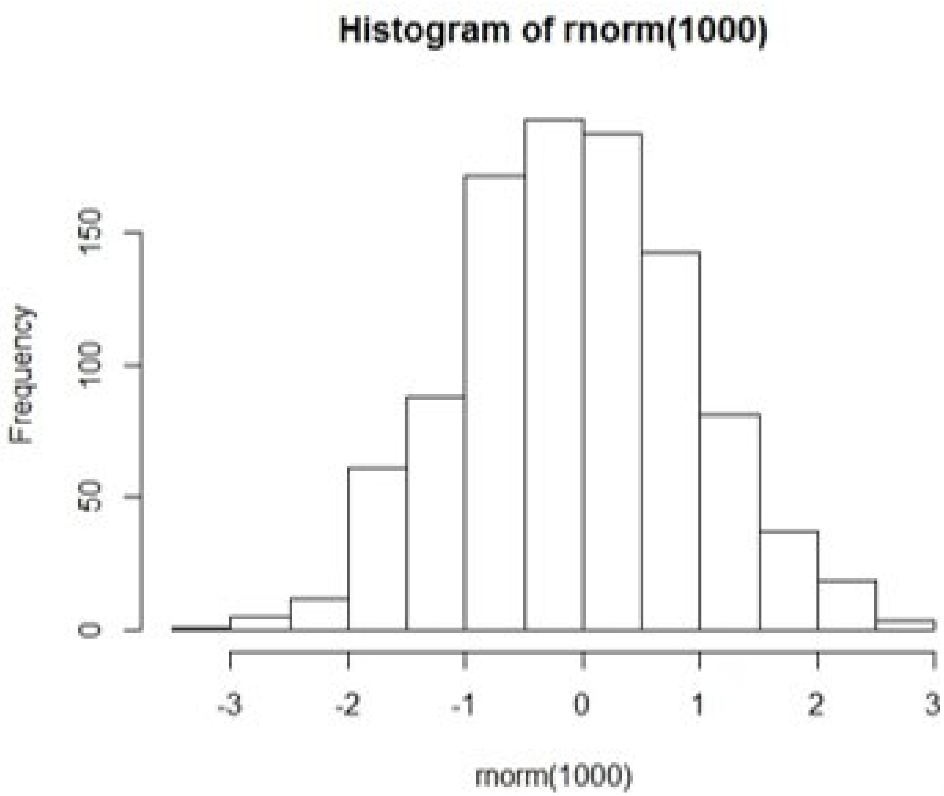
x = 0, 1, 2, 3, …

λ = mean number of occurrences in the interval

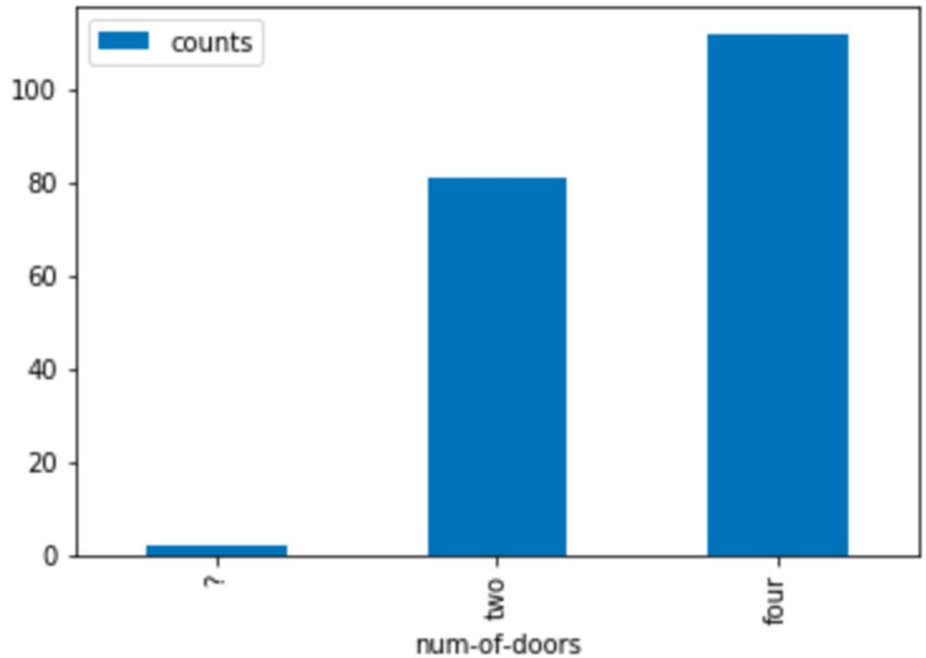
e = Euler's constant $\approx$ 2.71828

Visualizing Counts

Histogram: Number of values in bin

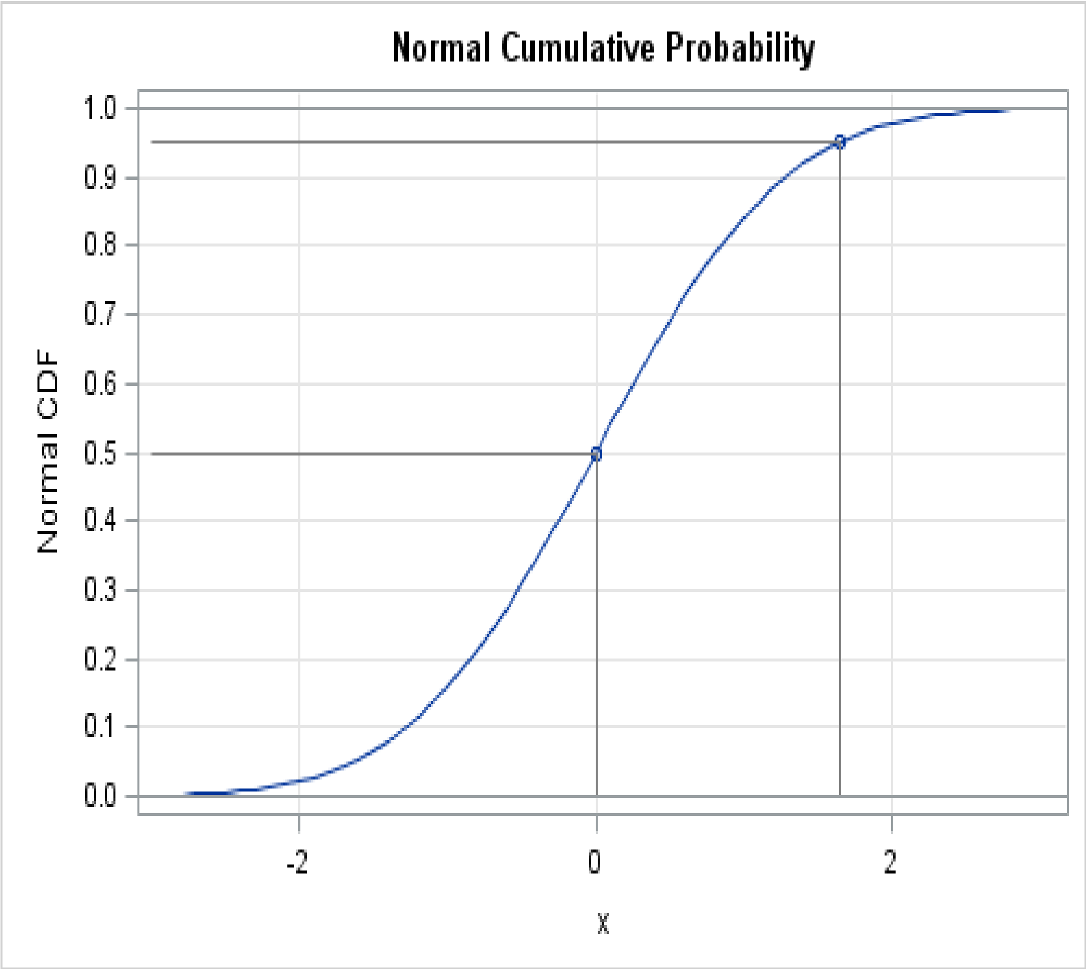


Bar Plot: Count of Categorical Variables



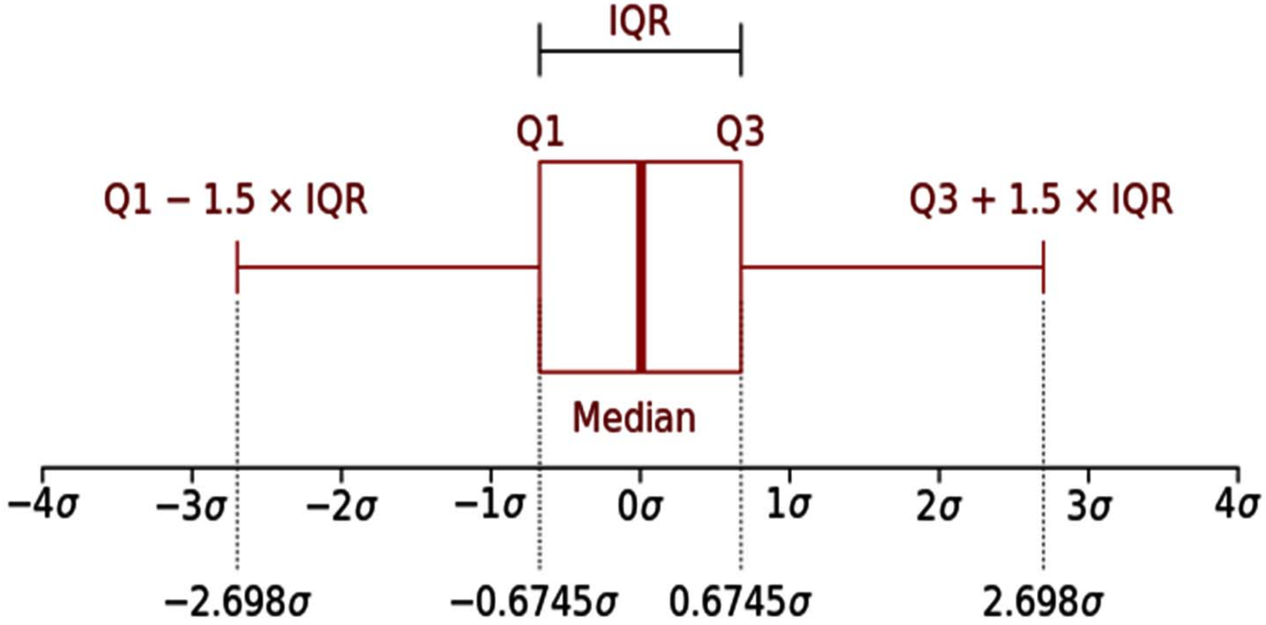
Quantiles of numerical vectors

- Quantiles are inverse values of the CDF (cumulative distribution function).
- Inverse tells you what value of x would make F(x) return a value "p"
- Standard Normal: (shown in figure)
  - Quantile(0.5) = 0, means at x=0, 50% of the distribution lies to the left. (This is also the median)
  - Quantile(0.95) = 1.65

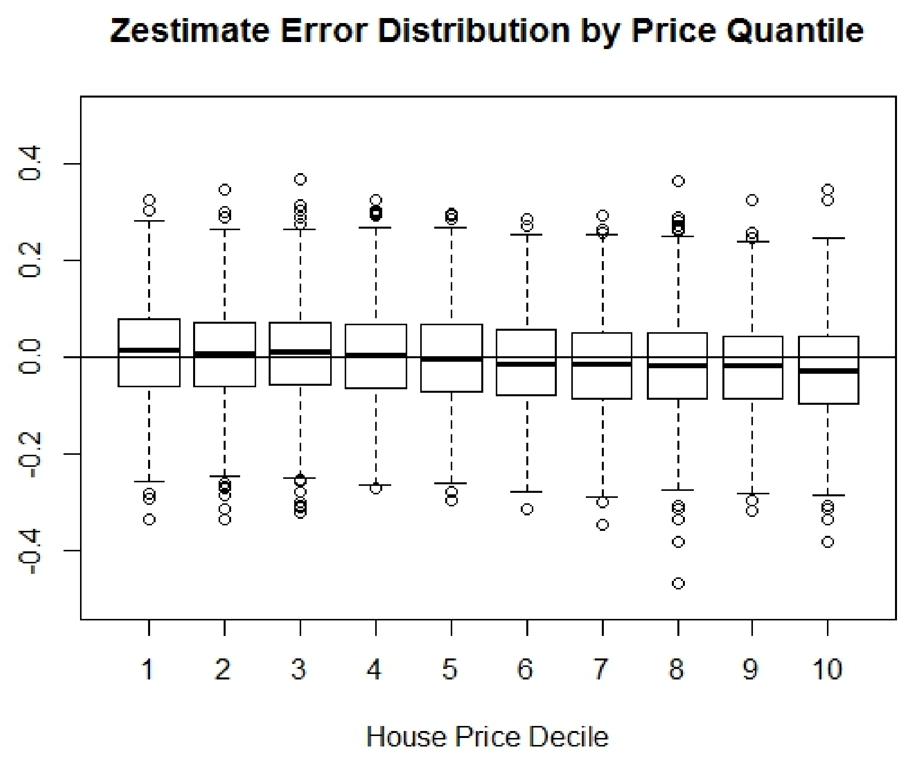


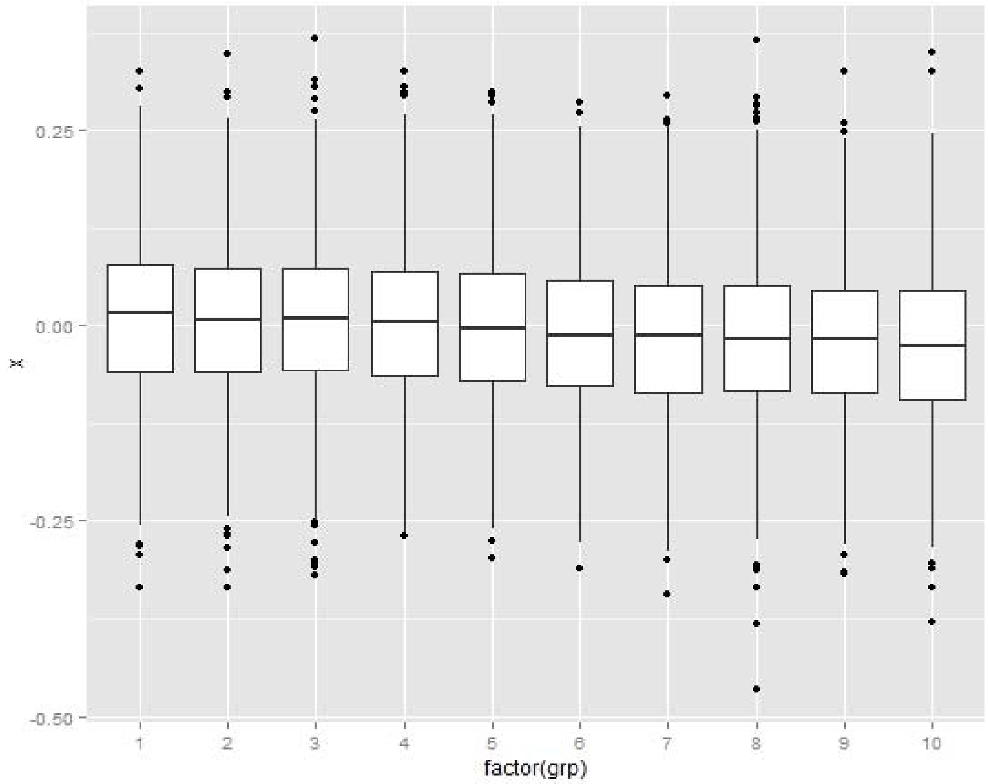
Inter Quartile Range (Q3 – Q1)

- "Middle 50%" = 75% - 25th percentile
- Measures variability
- Identifies outliers
  - below Q1 − 1.5 IQR or above Q3 + 1.5 IQR

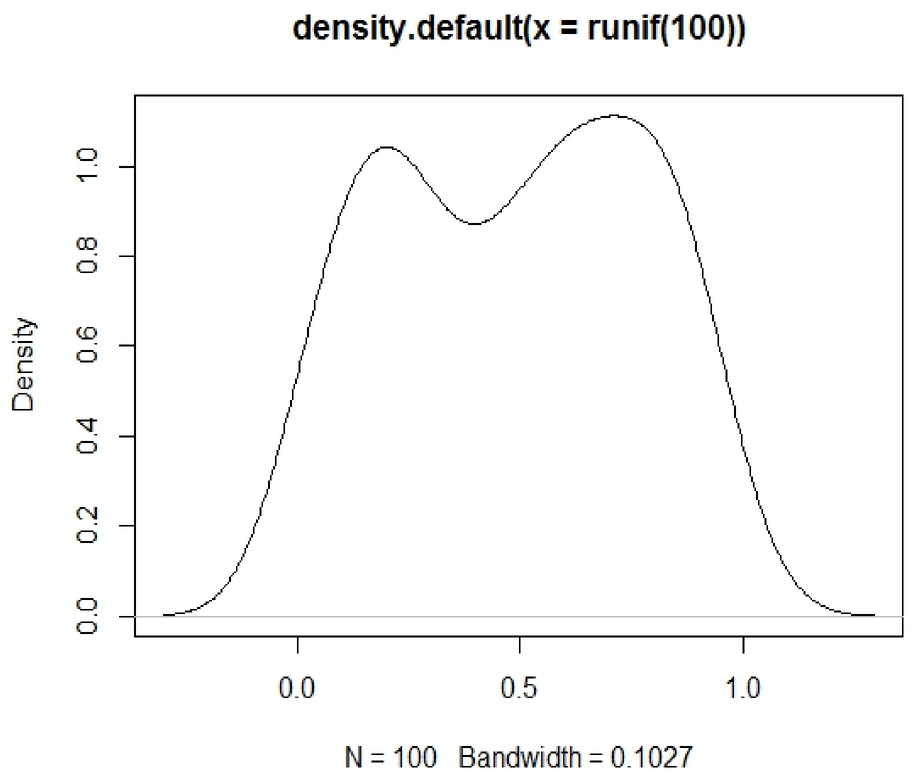


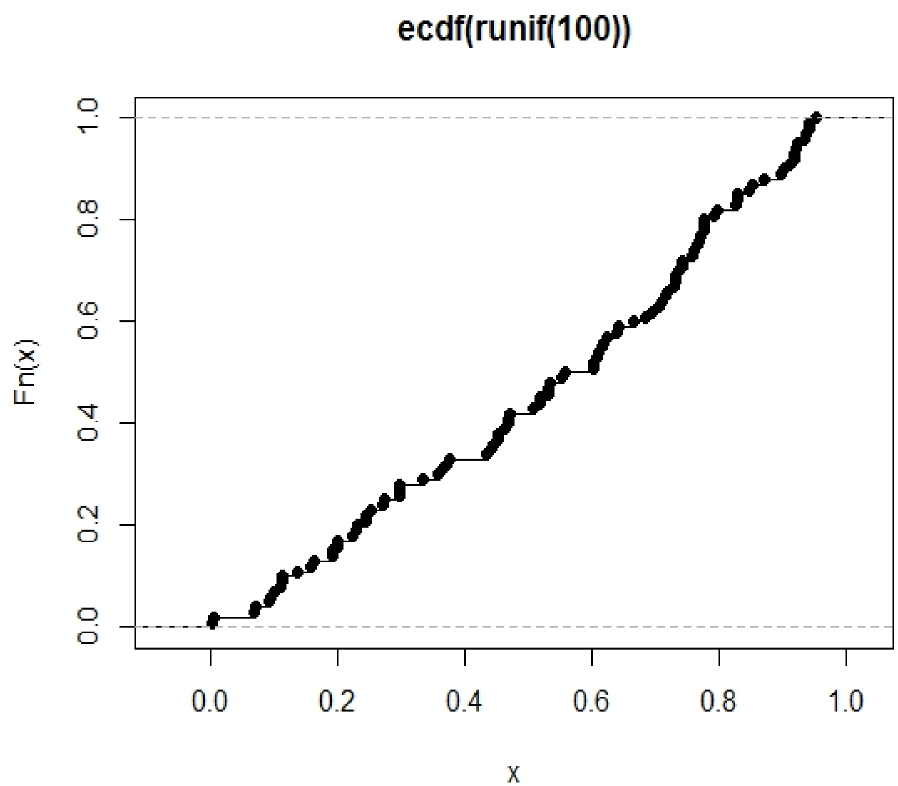
Visualizing IQR: Boxplots





Visualizing Densities/CDFs





Relationships between variables

Covariance

- Expected value of the differences between x and y and their corresponding mean.
- E.g. if x is above it's mean when y is also above it's mean, then they will have a high covariance.
- Highly interpretable, but not bounded.
- Measures strength and direction of relationship

$$Cov(X,Y) = \frac{\sum (X_i - \overline{X}) * (Y_i - \overline{Y})}{n}$$

Xi = some elemet in the sample X
Xbar = sample mean for x
N = number of elements in both samples

Correlation

> Correlations (Pearson's) = scaled covariance
  - Bounded between 0 and 1.
  - Can be easier to interpret

$$r = r_{xy} = \frac{\mathrm{Cov}(x, y)}{S_x \times S_y}$$

Sx = std dev

Visualizing Relationships: Scatterplots