Foundations of Data Science
Lecture 1, Module 2

Rumi Chunara, PhD

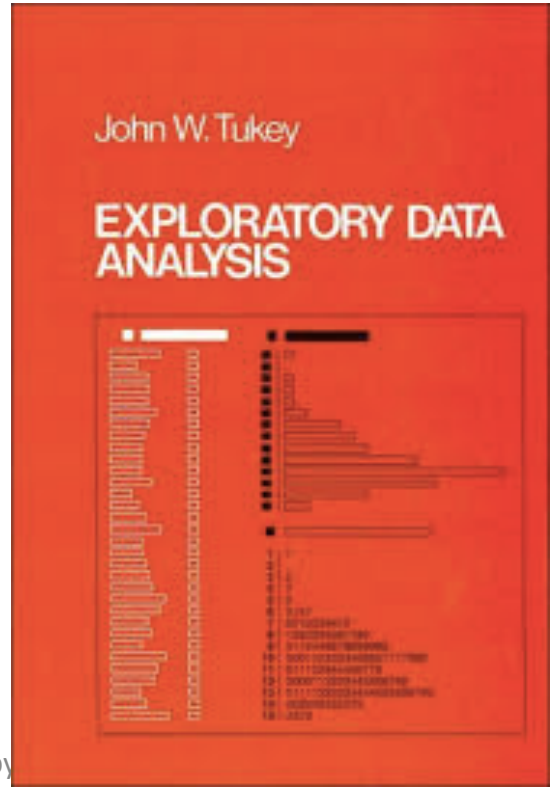
CS6053

Getting to Know Data

- Techniques (we will cover)
- Basic Statistical Descriptions of Data
- Measuring Data Similarity and Dissimilarity
- Summary

Exploratory Data Analysis 1977

- Based on insights developed at Bell Labs in the 60's
- Techniques for visualizing and summarizing data
- What can the data tell us? (in contrast to "confirmatory" data analysis)
- Introduced many basic techniques:
  - 5-number summary, box plots, stem and leaf diagrams,…
- 5 Number summary:
  - extremes (min and max)
  - median & quartiles
  - More robust to skewed & longtailed distributions



Descriptive vs. Inferential Statistics

- **Descriptive:** e.g., Median; describes data you have but can't be generalized beyond that
  - We'll talk about Exploratory Data Analysis
- **Inferential:** e.g., t-test, that enable inferences about the population beyond our data
  - These are the techniques we'll leverage for Machine Learning and Prediction

Applying techniques

- Many questions are causal: what would happen if? (e.g. I show this ad)
- But its easier to ask correlational questions, (what happened in this past when I showed this ad).
- **Supervised Learning:**
  - Classification and Regression
- **Unsupervised Learning:**
  - Clustering and Dimension reduction
- Note: Unsupervised Learning is often used inside a larger Supervised learning problem.
  - E.g. auto-encoders for image recognition neural nets.

Applying techniques

- **Supervised Learning:**
  - kNN (k Nearest Neighbors)
  - Naïve Bayes
  - Logistic Regression
  - Support Vector Machines
  - Random Forests
- **Unsupervised Learning:**
  - Clustering
  - Factor analysis
  - Latent Dirichlet Allocation

Basic Statistical Descriptions of Data

- Motivation
  - To better understand the data: central tendency, variation and spread
- Measures of central tendancy
  - mean, median, mode
- Measures of dispersion
  - max, min, variance, standard deviation
  - Data dispersion

Measuring the Central Tendency

- Mean (algebraic measure) (sample vs. population):

  Note: *n* is sample size and *N* is population size.

  $$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \mu = \frac{\sum x}{N}$$

  - Weighted arithmetic mean:

  $$\bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

  - Trimmed mean: chopping extreme values
- Median:
  - Middle value if odd number of values, or average of the middle two values otherwise
  - Estimated by interpolation (for *grouped data*):

age	*frequency*
1–5	200
6–15	450
16–20	300
21–50	1500
51–80	700
81–110	44

- Mode
  - Value that occurs most frequently in the data
  - Unimodal, bimodal, trimodal
  - Empirical formula: $mean - mode = 3 \times (mean - median)$

Measuring the Dispersion of Data

- Quartiles, outliers and boxplots
  - **Quartiles**: Q_1 (25th percentile), Q_3 (75th percentile)
  - **Inter-quartile range**: IQR = $Q_3 - Q_1$
  - **Five number summary**: min, Q_1, median, Q_3, max
  - **Boxplot**: ends of the box are the quartiles; median is marked; add whiskers, and plot outliers individually
  - **Outlier**: usually, a value higher/lower than 1.5 x IQR
- Variance and standard deviation (*sample: s, population: σ)*
  - **Variance**: (algebraic, scalable computation)

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}[\sum_{i=1}^{n} x_i^2 - \frac{1}{n}(\sum_{i=1}^{n} x_i)^2] \qquad \sigma^2 = \frac{1}{N}\sum_{i=1}^{n}(x_i - \mu)^2 = \frac{1}{N}\sum_{i=1}^{n} x_i^2 - \mu^2$$

  - **Standard deviation** *s (or σ)* is the square root of variance s^2 *(or* σ^2*)*

Graphic Displays of Basic Statistical Descriptions

- **Boxplot**: graphic display of five-number summary
- **Histogram**: x-axis are values, y-axis arefrequencies
- **Quantile plot**: each value x_i is paired with f_i indicating that approximately 100 f_i% of data are $\leq x_i$
- **Quantile-quantile (q-q) plot**: graphs the quantiles of one univariant distribution against the corresponding quantiles of another
- **Scatter plot**: each pair of values is a pair of coordinates and plotted as points in the plane

The Trouble with Summary Stats

Set A		Set B		Set C		Set D	
X	Y	X	Y	X	Y	X	Y
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.1	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.1	4	5.39	19	12.5
12	10.84	12	9.11	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89

Summary Statistics Linear Regression

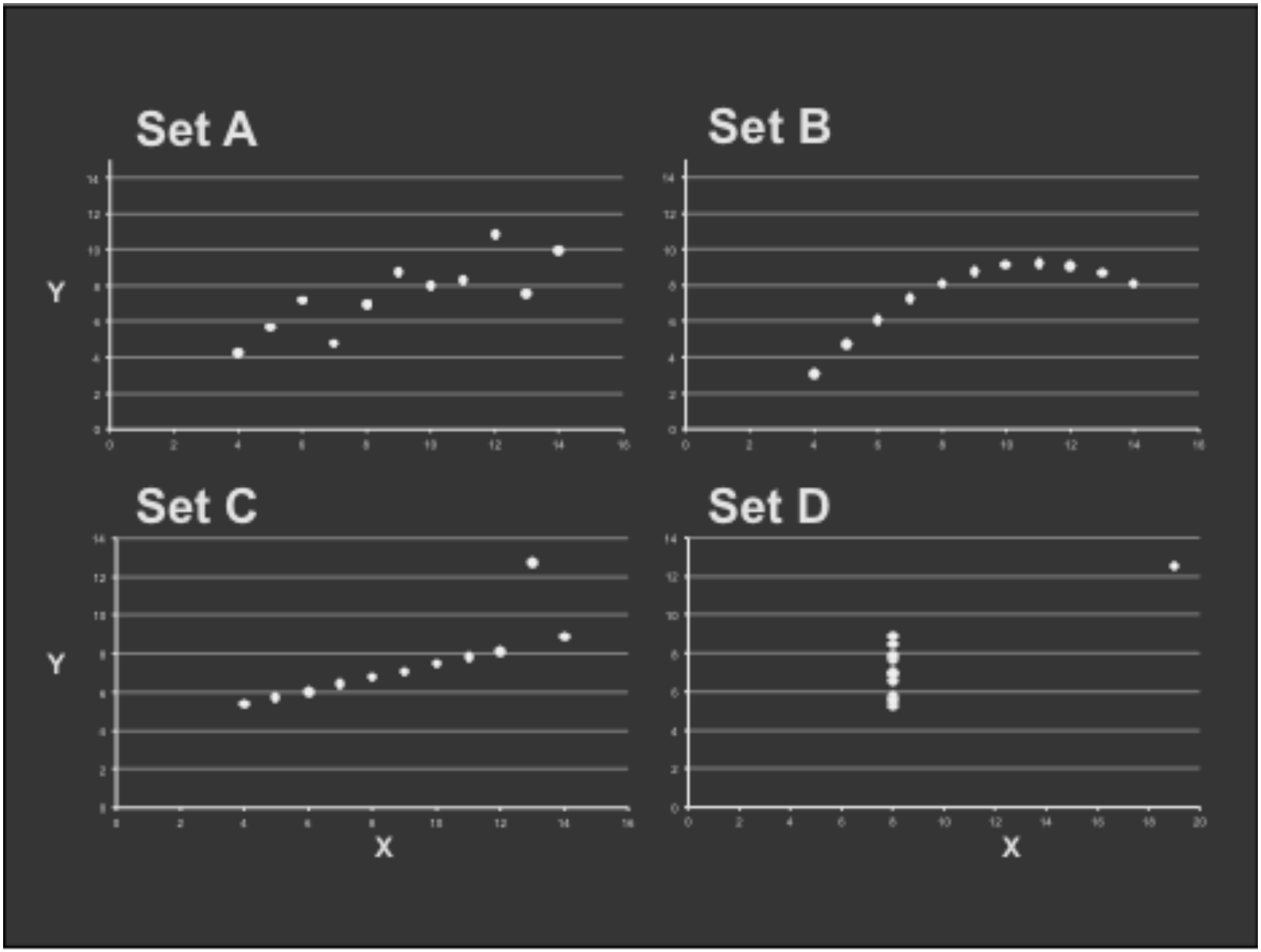
$u_X = 9.0$ $\sigma_X = 3.317$ $Y = 3 + 0.5\,X$

$u_Y = 7.5$ $\sigma_Y = 2.03$ $R^2 = 0.67$

[Anscombe 73]

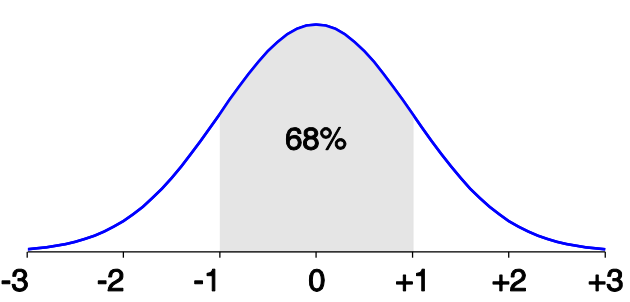
ce
ed

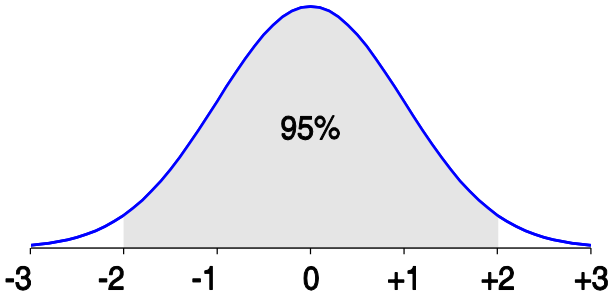
Looking at Data

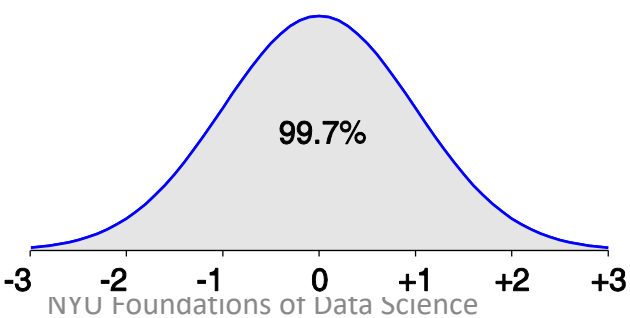


Properties of Normal Distribution Curve

- The normal (distribution) curve
  - mean = median = mode
  - From μ–σ to μ+σ: contains about 68% of the measurements (μ: mean, σ: standard deviation)
  - From μ–2σ to μ+2σ: contains about 95% of it
  - From μ–3σ to μ+3σ: contains about 99.7% of it



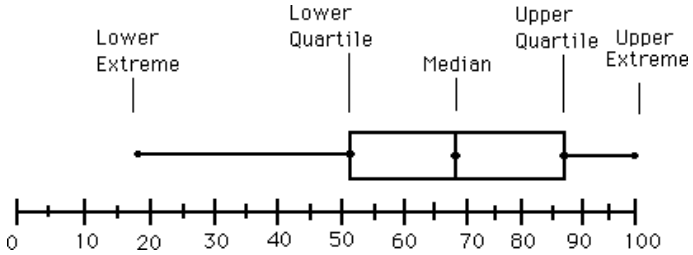




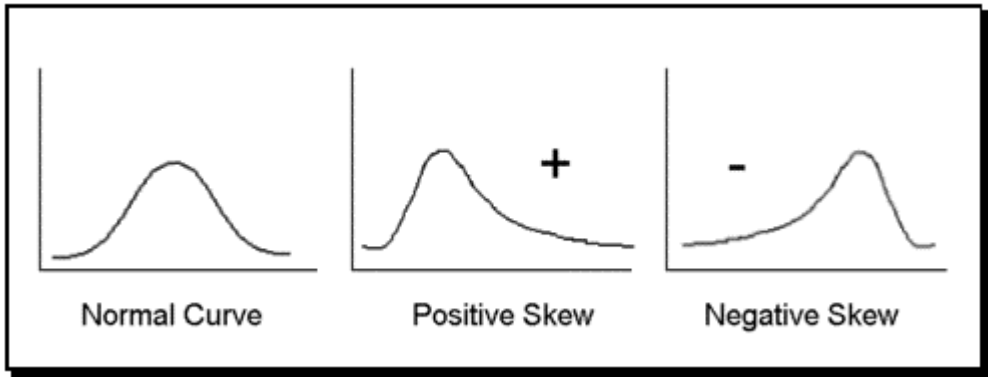
Correcting distributions

Many statistical tools, including mean and variance, t-test, ANOVA etc. **assume data are normally distributed**.

Very often this is not true. The box-and-whisker plot is a good clue



Whenever its asymmetric, the data cannot be normal. The histogram gives even more information

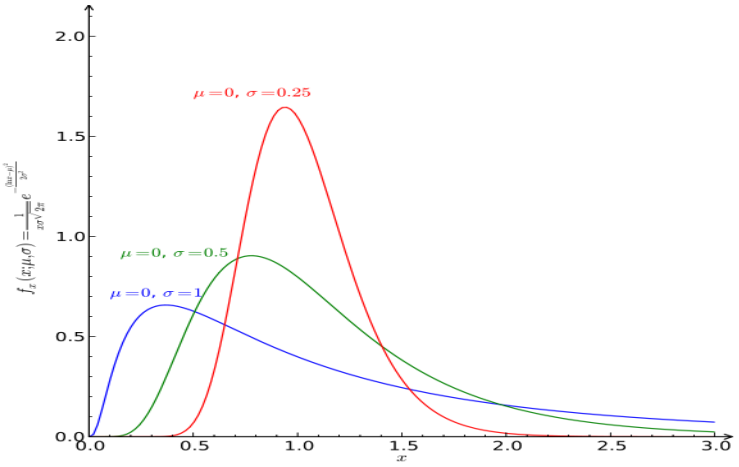


Correcting distributions

In many cases these distribution can be corrected before any other processing.

Examples:

- X satisfies a log-normal distribution, Y=log(X) has a normal dist.



- X poisson with mean k and standard deviation: sqrt(k). Then sqrt(X) is approximately normally distributed with standard deviation = 1

Distributions

Some other important distributions:

- **Poisson:** the distribution of counts that occur at a certain "rate".
  - Observed frequency of a given term in a corpus.
  - Number of visits to a web site in a fixed time interval.
  - Number of web site clicks in an hour.
- **Exponential:** the interval between two such events.
- **Zipf/Pareto/Yule distributions:** govern the frequencies of different terms in a document, or web site visits.
- **Binomial/Multinomial:** The number of counts of events (e.g. die tosses = 6) out of n trials.

- You should understand the distribution of your data before applying any model.

Rhine Paradox*

Joseph Rhine was a parapsychologist in the 1950's (founder of the *Journal of Parapsychology* and the *Parapsychological Society, an affiliate of the AAAS*).

He ran an experiment where subjects had to guess whether 10 hidden cards were red or blue.

He found that about 1 person in 1000 had ESP, i.e. they could guess the color of all 10 cards.

Q: what's wrong with his conclusion?

* Example from Jeff Ullman/Anand Rajaraman

Rhine Paradox

He called back the “psychic” subjects and had them do the same test again. They all failed.

He concluded that the act of telling psychics that they have psychic abilities causes them to lose it…(!)

Today, Recap

- What is Data Science?
- Doing Data Science
- About the course
- Statistics Review
- Intro to Python

Next Time

- Data Handling
- Pandas Intro

- Don’t forget: Homework 1 due next Tuesday 9am