Data Science Engineering Methods and Tools – INFO 6107 Lecture #01

Descriptive Statistics

Instructor: Ebrahim Nasrabadi

Note 1. Some wording is taken from Wikipedia, Redfin, Zillow.com

1 Introduction

Data is everywhere in today's business world and data-driven decisions play a critical role to the business growth and success. Some of the typical examples of data are as follows:

- Housing market data: Real estate websites have direct access to public data from local multiple listing services as well as insight from real estate agents across the country. They give buyers, sellers, and homeowners data on the state of the housing market including prices (estimate of a home's value as well as price history), sales (number of homes sold), and inventory (number of homes on the market) for metropolitan areas, cities, neighborhoods and zip codes across the nation.

  In particular, Redfin and Zillow (the two Seattle-based online real estate brokerages) analyze millions of public and user-submitted data points (including price, property type, square footage, bedrooms, bathrooms, lot size, views and location as well as buyer preferences in the area) and provide valuable information to buyers and sellers. Redfin analyzes more than 500 variables of each home as soon as it hits the market to identify Hot Homes that are likely to sell within the first two weeks on the market. Redfin's Hot Home feature is a very useful for buyers in the fast-moving, competitive environment of most major housing markets today. It offers buyers know which homes to go see in person right away. Zillow estimates a home's value (called Zestimate) and forecast the value one year from now, based on current home and market information. Zestimate is calculated for about 100 million homes nationwide and updated for all homes daily.

- Marketing data: Marketing data provides a critically important aspect that any successful business relies on to increase its revenue, strengthen its relationships with customers and continue evolving in a crowded marketplace. Retail companies (such as Wal-Mart, Costco, Amazon) use customer data to understand customers' needs, desires and behaviors, and provide them with the products and services which will best fit their customers' needs. They use marketing data to provide customers with relevant and personalized experiences. The main three sources for marketing data are

  - Data that is inferred by your interests: What you buy, search, save, like, watch, etc
  - Data you provide: marketing forms that you will out such as application for a loan
  - Data from public records: homeownership records, marriage licenses

- Credit report data: When a customer applies for a loan, banks and other credit providers analyze different data points (including annual income, home ownership, years in current job, credit payment history, credit utilization, length of credit history, new credit and credit

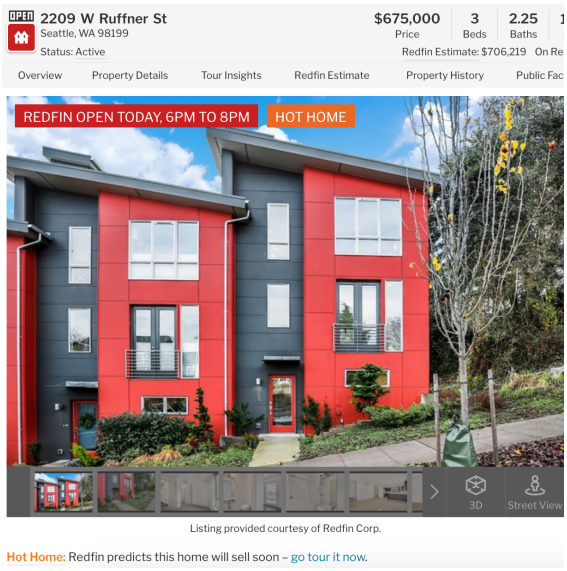

Figure 1: Redfin Hot Home

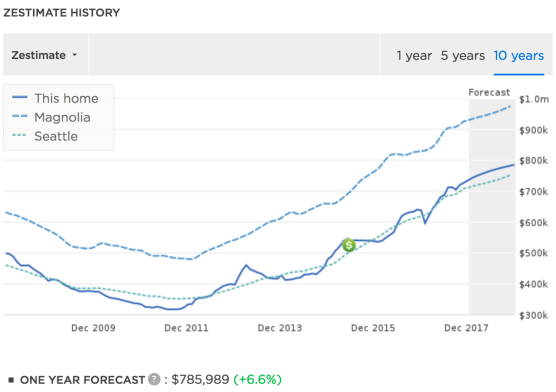

Figure 2: Zillow's best estimate of a home's value

mix) and use statistical models to determine whether or not to grant the loan based on the likelihood of the loan being repaid. The factors involved in determining this likelihood are complex, and extensive statistical analysis and modeling are required to predict the outcome for each individual case.

Vast amounts of data are collected every day, every hour, every minute, and even every second in today's business environment. While raw data does not provide actionable insights and it can be overwhelming, we can use statistical methods to produce insightful information. In addition, many problems in business and management involve decision making under uncertainty?that is, choosing actions based on often imperfect observations, with unknown outcomes. Statistic is a very broad subject with many applications in many different fields. It is a branch of science that deals with data analysis, i.e., with the process of collecting, cleaning, transforming, and modeling data with the goal of highlighting useful information, suggesting conclusions, and supporting decision making under uncertainty.

Statistics can be used to answer the following questions:

- What is the impact of a new feature or change to user experience?
- Is there any relationship between GPA and starting salary after graduation?
- What is the effect of package designs on sales?
- How to interpret polls. How many individuals you need to sample for your inferences to be acceptable? What is meant by the margin of error?
- What is the effect of market strategy on market share?
- How to pick the stocks to invest in?

- What is the price sensitivity of a product or service (e.g., the degree to which price affects the sales)?
- Is there a relationship between advertising budget and sales? If yes, how strong is the relationship?
- How accurately can we predict demand for a product or service?
- What is the default risk (i.e., the chance that companies or individuals will be unable to make the required payments on their debt obligations)?
- What is the chance that a house goes under contract in its first two weeks on the market?

The most basic application of statistics is to summarize and describe a collection of data and find basic quantitative characteristics such as central tendency (average, median, mode) or dispersion (standard deviation or range). This is called *Descriptive Statistics*. While descriptive statistics tell us basic information about the population or data set under study, *inferential statistics* allow us to infer trends about a larger population based on a study of a sample taken from it. We use inferential statistics to examine the relationships between variables within a sample, and then make generalizations or predictions about how those variables will relate within a larger population. Techniques that are used to examine the relationships between variables, and thereby to create inferential statistics, include but are not limited to: linear regression, logistic regression that will be covered in Lectures 5 and 6.

Inferential statistics are valuable when

- it not convenient or possible to examine each member of an entire population (e.g., too expensive to have an a.).
- population is constantly changing and growing (e.g., Facebook's users)
- we want to test a few feature and would like to measure the impact before making a deployment decision (e.g., demonstrate the impact of new features or changes to user experience)

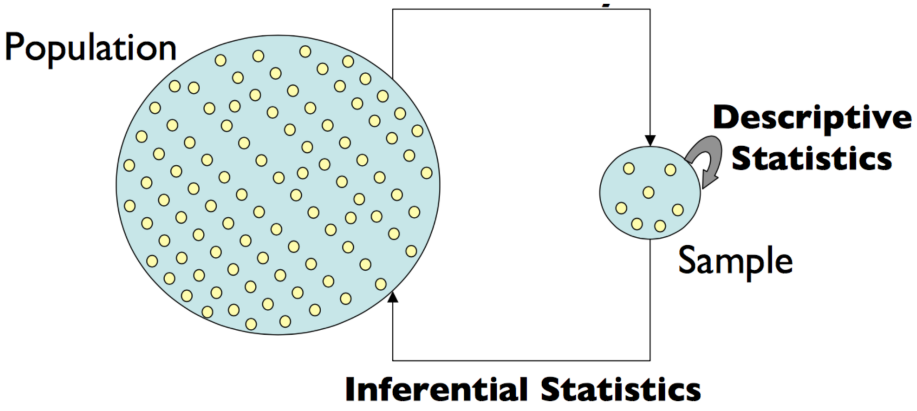

Figure 3: Inferential Statistics: make inference about the population from samples such as point estimate, confidence intervals, hypothesis testing.

The basic elements of an inferential statistical problem are

(i) A clear definition of the population and variable of interest

(ii) A design of the experiment or sampling procedure.

(iii) Collection and analysis of data (gathering and summarizing data).

(iv) Procedure for making inferences (predictions, decisions) about the population based on sample information.

(v) A measure of 'goodness' or reliability for the procedure.

Before making inferences from data, it is essential to examine all the variables and data to

- to catch mistakes and missing values
- to see patterns in the data
- to find violations of statistical assumptions
- to generate hypotheses
- and because if you don't, you will have trouble later

2 Descriptive Statistics

Descriptive statistics are the basic statistics that describe what is going on in a population or data set. There are two types of descriptive statistics: measures of central tendency and measures of spread.

Measures of central tendency capture general trends within the data, and are calculated and expressed as the mean, median, and mode. A mean is the mathematical average of all of our data, like for example average price of house in Seattle; median represents the middle of the data distribution, like the median price of house in Seattle; and mode is the most frequently occurring value, like the most common movie title. The mode is rarely used as a measure of central tendency for quantitative variables. However, for qualitative variables, the mode is more useful because the mean and median do not make sense. The mode can be used with qualitative data, but the mean and median cannot.

Meanwhile, measures of spread describe how the data are distributed and how they relate to each other. Statistical measures that show us this include range (the entire range of values present in a data set), frequency distribution (how many times a particular value occurs within a data set), quartiles (subgroups formed within a data set when all values are divided into four equal parts across the range), mean absolute deviation (the average of how much each value deviates from the mean), variance (illustrates how much of a spread exists in our data), and standard deviation (illustrates the spread of data relative to the mean).

In what follows, we formally define descriptive statistics. We start with some basic definitions and terminology that will be used throughout the course.

2.1 Definitions and Terminology

Data: Any kind of recorded information that can be used for some purpose (e.g., home's value)

Population: set of all subjects of interest in a particular study (e.g., all houses in Seattle)

Sample: A subset of subjects selected from the population of interest

Variable: (also called feature or attribute in Machine Learning): A property of an individual population unit (e.g., price, property type, square footage, bedrooms, bathrooms)

Types of variables:

Qualitative: variables whose values representing counts or measurements. A quantitative variable is either

- Discrete: number of bedrooms
- Continuos: price, lot size

Quantitative: variables whose values can be placed into nonnumeric categories. A qualitative variable is either

- Binary: Hot Home or Not
- Nominal: Property Type (House, Condo, Townhouse)
- Ordinal: Medal Type (Gold, Silver, Bronze)

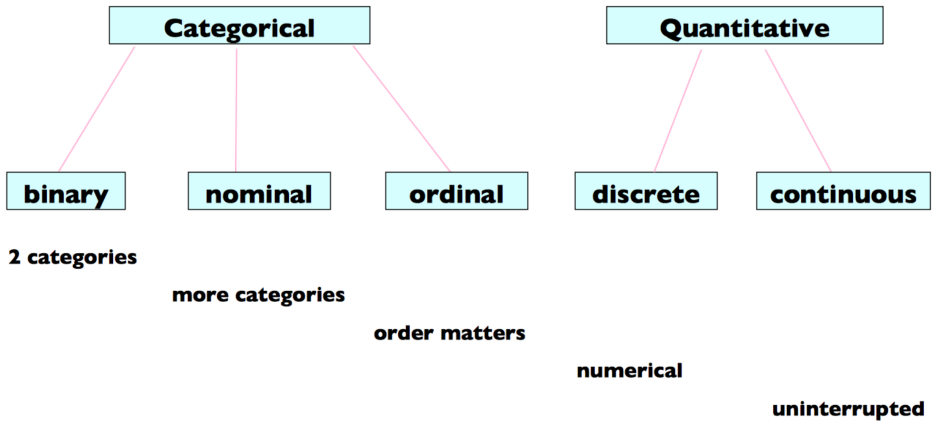

Figure 4: Types of variables

Dimensionality of Data Sets:

Univariate : Measurement made on one feature per subject

- User Activity: number of likes per user
- Purchase History: number of units per customer
- Shipment: weight per package

Bivariate : Measurement made on two features per subject

- User Activity: number of likes and comments per user
- Purchase History: dollar amount and number of units per customer
- Shipment: weight and volume per package

Multivariate : Measurement made on many features per subject

- User Activity: number of likes, comments, views, ... per user
- Purchase History: dollar amount, number of units, number of unique products, ... per customer
- Shipment: weight, volume, and number of units, ... per package

2.2 Central Tendency Measures

Given a sample of measurements $x_1, x_2, \ldots, x_n$ where n = sample size and x_i = value of the i^{th} observation in the sample, the mean sample and median are defined as:

Sample Mean :

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Sample Median : The median of a sample (data set) is the middle number when the measurements are arranged in ascending order. Note that

- If n is odd, the median is the middle number.
- If n is even, the median is the average of the middle two numbers.

Mode: The mode is the value of x (observation) that occurs with the greatest frequency.

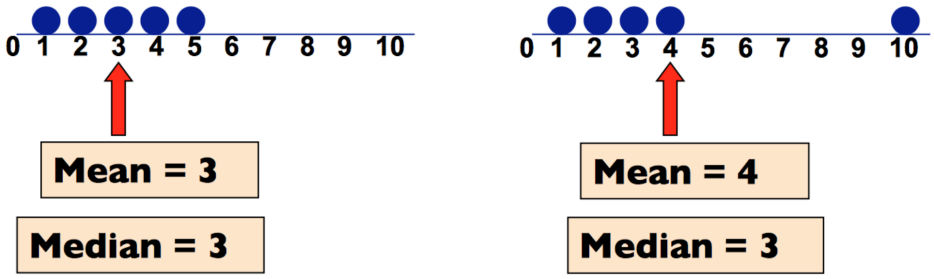

Figure 5: Mean vs Median

Example 2.1. Given a sample of 5 home values

$$\$650{,}000, \$52{,}000, \$745{,}000, \$690{,}000, \$1{,}100{,}000$$

then

$$\bar{x} = \frac{\$650{,}000 + \$52{,}000 + \$745{,}000 + \$690{,}000 + \$1{,}100{,}000}{5} = \frac{\$3{,}237{,}000}{5} = \$647{,}400$$

To find the mode, we first arrange the numbers in ascending order:

$$\$52{,}000, \$650{,}000, \$690{,}000, \$745{,}000, \$1{,}100{,}000$$

The median home value is \$690,000.

Example 2.2. Given a sample of 10 home values

$$\$550{,}000, \$425{,}000, \$645{,}000, \$595{,}000, \$495{,}000, \$550{,}000, \$625{,}000, \$545{,}000, \$60{,}000, \$2{,}000{,}000$$

then

$$\bar{x} = \frac{\$6{,}490{,}000}{10} = \$649{,}000$$

To find the mode, we first arrange the numbers in ascending order:

$$\$425{,}000, \$495{,}000, \$545{,}000, \$550{,}000, \$550{,}000, \$595{,}000, \$60{,}000, \$625{,}000, \$645{,}000, \$2{,}000{,}000$$

The median home value is $\frac{550000+595000}{2} = \$572{,}500$.

Remarks:

(i) the mean is sensitive to extreme values, so it is best for symmetric distributions without outliers

(ii) the median is insensitive to extreme values (because median is a measure of location or position). So it is useful for skewed distributions or data with outliers

(iii) the mode is rarely used as a measure of central tendency for quantitative variables. However, for qualitative variables, the mode is more useful because the mean and median do not make sense. The mode can be used with qualitative data, but the mean and median cannot.

2.3 Measures of Spread

Consider a sample of measurements $x_1, x_2, \ldots, x_n$.

Range : Range = largest value - smallest value or Range = max - min

Mean Absolute Difference (MAD) : Average of absolute deviations of values from the mean

$$MAD = \frac{\sum_{i=1}^{n} |x_i - \bar{x}|}{n}$$

Sample Variance: Average of squared deviations of values from the mean

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

Remarks

(i) Adding deviations will yield a sum of 0

(ii) Absolute values do not have nice mathematical properties

(iii) Squares eliminate the negatives

Sample Standard Deviation: Sample standard deviations are simply the square root of the sample variance

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

Remarks

(i) The standard deviation is in the same unit of measurement as the original data, whereas the variance is expressed in squared units.

Chebyshev's Inequality. (Regardless of how the data are distributed) Given a number $k \geq 1$, and a set of measurements $x_1, x_2, \ldots, x_n$, at least (1-1/k^2) of the observations lie within k standard deviations from the mean. In other words, at least (1-1/k^2) of observations lie in the interval $(\bar{x} - ks, \bar{x} + ks)$

Example 2.3. A sample data set has $\bar{x} = 75$ $s = 6$. Then

(i) ($k = 1$): at least 0% of all observations lie in $(\bar{x} - s, \bar{x} + s) = (69, 81)$

(ii) ($k = 2$): at least 75% of all observations lie in $(\bar{x} - 2s, \bar{x} + 2s) = [63, 87]$

(iii) ($k = 3$): at least 88% of all observations lie in $(\bar{x} - 3s, \bar{x} + 3s) = [57, 93]$

Often we can do better, especially if the frequency distribution is bell shaped. If the frequency distribution is approximately bell shaped, then

(i) approximately 68% of the observations lie within one standard deviation of their sample mean, i.e. $(\bar{x} - s, \bar{x} + s)$

(ii) approximately 95% of the observations lie within two standard deviations of their sample mean, i.e. $(\bar{x} - 2s, \bar{x} + 2s)$

(iii) approximately 99% of the observations lie within two standard deviations of their sample mean, i.e. $(\bar{x} - 2s, \bar{x} + 2s)$

Example 2.4. A sample data set has $\bar{x} = 75$ $s = 6$. The frequency distribution is approximately bell shaped. Then

(i) (69, 81) contains approximately 68% of the observations

(ii) (63, 87) contains approximately 95% of the observations

(iii) (57, 93) contains at least 99% (almost all) of the observations

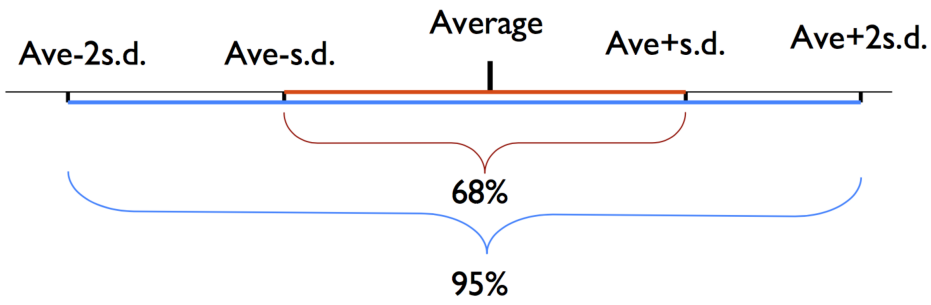

Figure 6: Assuming the frequency distribution is approximately bell shaped