Statistics in Business Analytics

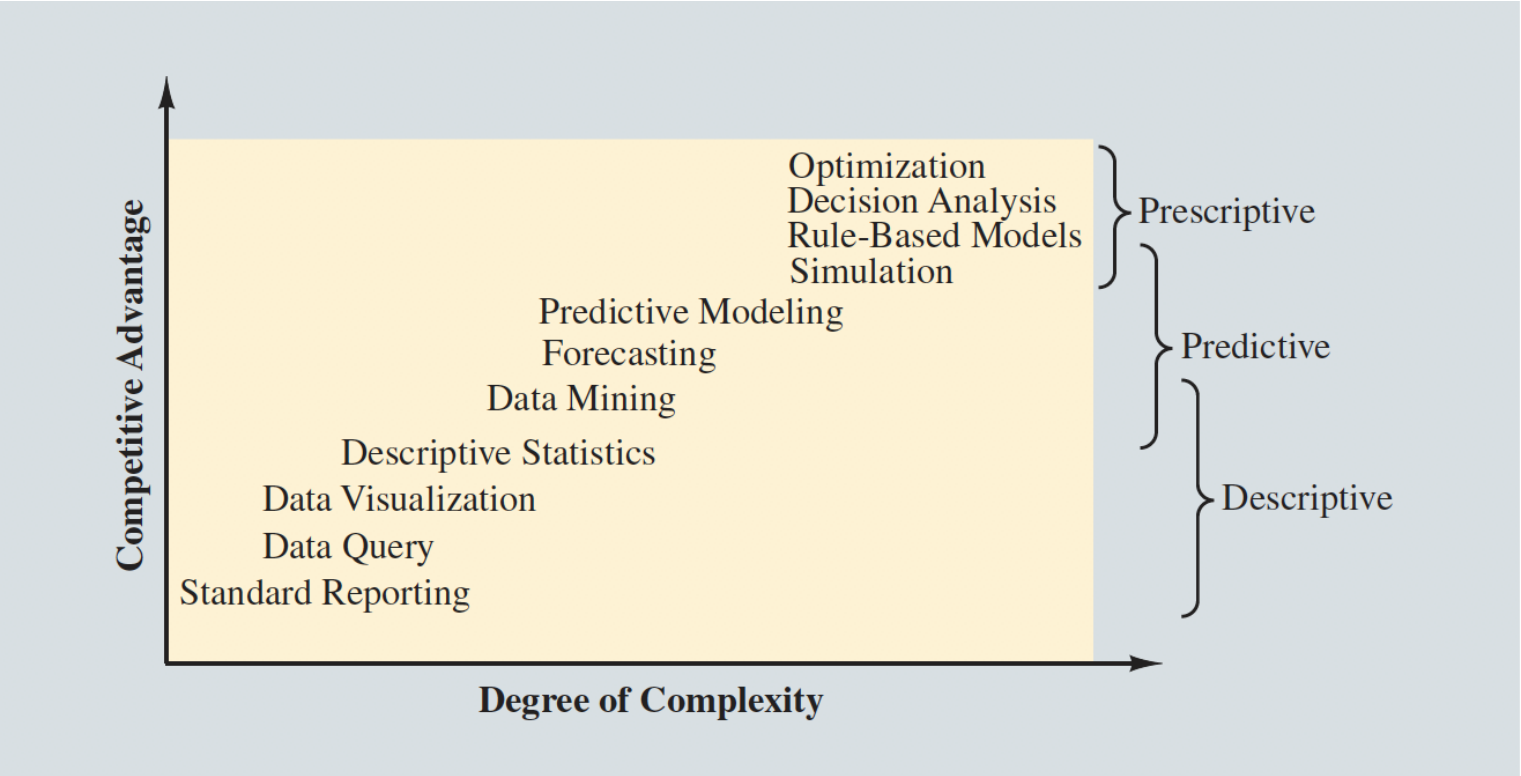
Liu Xuefeng

May 22, 2019

Basic Concepts

- Business analytics is the scientific process of transforming data into insight for making better decisions.
- Statistics is a branch of mathematics working with data collection, organization, analysis, interpretation and presentation.
- Some examples of business analytics by area: financial analytics, human resource analytis, market analytics, supply chain analytics, sports analytics, web analytics.

Tasks in Business Analytics



Tasks in Statistics

- Descriptive Statistics: Analytical tools that describe what has happened.
- Data visualization.
- Statistical inference: Techniques that use models constructed from past data to ascertain the impact of one variable on another.
- Predictive analysis: Techniques that use models constructed from past data to predict the future.
- Prescriptive analysis: Techniques that analyze input data and yield a best course of action.

Descriptive Statistics

- The role of descriptive analytics is to collect and analyze data to gain a better understanding of variation and its impact on the business setting.
- Example of descriptive statistics, mean, standard deviation, TWAP, VWAP.

Type of Data

- Quantitative data. Data are considered quantitative data if numeric and arithmetic operations, such as addition, subtraction, multiplication, and division, can be performed on them.
- Categorical data. We can summarize categorical data by counting the number of observations or computing the proportions of observations in each category.

Measures of Location

- Mean. Excel formula: AVERAGE

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

- Median: the value in the middle when the data are arranged in ascending order (smallest to largest value). Excel formula: MEDIAN
- Mode: the value that occurs most frequently in a data set. Excel formula: MODE.MULT
- Geometric mean.Excel formula: GEOMEAN

$$\bar{x}_g = \sqrt[n]{(x_1)(x_2)\cdots(x_n)}$$

Measures of Variability

- Range. The range can be found by subtracting the smallest value from the largest value in a data set.Excel formula: MAX - MIN
- Variance. Excel formula: VAR.S

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

- Standard deviation. Excel formula: STDEV.S

$$s = \sqrt{s^2}$$

Measures of Association between Two Variables

- Covariance is a descriptive measure of the linear association between two variables. Excel formula: COVARIANCE.S

$$s_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n}$$

- Correlation coefficient measures the relationship between two variables. Excel formula: CORREL.

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

Data Visualization

- Tables
- Charts: scatter plot, line plot, bubble plot, heat map, bar chart, pie chart.

Hypothesis Tests

- Hypothesis test is the process of making a conjecture about the value of a population parameter, collecting sample data that can be used to assess this conjecture, measuring the strength of the evidence against the conjecture that is provided by the sample, and using these results to draw a conclusion about the conjecture.
- Null hypothesis: The hypothesis tentatively assumed to be true in the hypothesis testing procedure. Usually denoted by H_0.
- Alternative hypothesis: The hypothesis concluded to be true if the null hypothesis is rejected. Usually denoted by H_a.

Hypothesis Tests

- Type I error: The error of rejecting H_0 when it is true.
- Type II error: The error of accepting H_1 when it is false.
- Principal: protect null hypothesis, put what you want to prove as alternative hypothesis.

Two-Sample Student's T Test

- Two-sample Student's t test is used to check whether the mean of two populations are the same.
- $H_0: \ \mu_1 = \mu 2$ v.s. $H_a: \ \mu_1 \neq \mu 2$
- Test statistics:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}},$$

where

$$s_p = \sqrt{\frac{(n_1 - 1)s_{X_1}^2 + (n_2 - 1)s_{X_2}^2}{n_1 + n_2 - 2}}$$

- t follows a T distribution with degree of freedom $n_1 + n_2 - 2$ if H_0 holds.
- Excel formula: T.TEST.

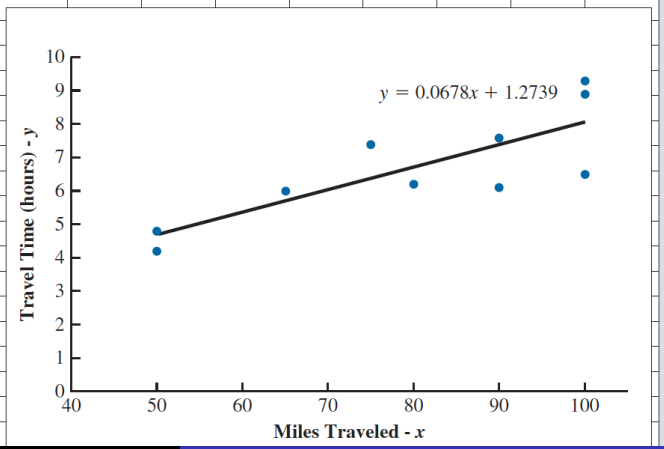
Simple Linear Regression

- $y = \beta_0 + \beta_1 x + \epsilon$, where y is response variable, x is explanatory variable, β_0 is intercept, β_1 is slope, $\epsilon \sim N(0, \sigma^2)$ is error term.

Butler Trucking Company example

Driving Assignment i	x = Miles Traveled	y = Travel Time (hours)
1	100	9.3
2	50	4.8
3	50	8.9
4	100	6.5
5	50	4.2
6	80	6.2
7	75	7.4
8	65	6.0
9	90	7.6
10	90	6.1

	A	B	C
1	Assignment	Miles	Time
2	1	100	9.3
3	2	50	4.8
4	3	100	8.9
5	4	100	6.5
6	5	50	4.2
7	6	80	6.2
8	7	75	7.4
9	8	65	6.0
10	9	90	7.6
11	10	90	6.1



Multiple Linear Regression

- $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_q x_q + \epsilon$.
- Multiple linear regression is an extension to simple linear regression to multiple explanatory variable cases.

Generalized Linear Regression

- Generalized linear regression is an extension to multiple linear regression that allows the response variable have distribution other than normal.
- $\eta = \beta_0 + \beta_1 x_1 + \cdots + \beta_q x_q$, where η known as link function is an monotonic function of the mean of response variable.
- Logistic regression is the most popular generalized linear regression which assumes the response variable follows binary distribution.

Logistic Regression Example

pass	hour
0	0.5
0	0.75
0	1
0	1.25
0	1.5
0	1.75
1	2
0	2.25
1	2.5
0	2.75
1	3
0	3.25
1	3.5
0	4
1	4.25
1	4.5
1	4.75
1	5
1	5.5

- In the example, response y is whether a student pass a exam, explanatory variable x is how many hours a student spend on exam preparation.
- $$\mu = \frac{1}{1 + \exp(-(\beta_0 + \beta_1 x))}$$
  , where $\mu = E(y)$ is the probability for a student to pass the exam.
- Excel add-in RealStatistics support logistic regression.