Data Analysis Course

Descriptive Statistics(Version-1)

Venkat Reddy

Data Analysis Course

- Data analysis design document
- Introduction to statistical data analysis
- **Descriptive statistics**
- Data exploration, validation & sanitization
- Probability distributions examples and applications
- Simple correlation and regression analysis
- Multiple liner regression analysis
- Logistic regression analysis
- Testing of hypothesis
- Clustering and decision trees
- Time series analysis and forecasting
- Credit Risk Model building-1
- Credit Risk Model building-2

Note

- This presentation is just class notes. The course notes for Data Analysis Training is by written by me, as an aid for myself.
- The best way to treat this is as a high-level summary; the actual session went more in depth and contained other information.
- Most of this material was written as informal notes, not intended for publication
- Please send questions/comments/corrections to venkat@trenwiseanalytics.com or 21.venkat@gmail.com
- Please check my website for latest version of this document

-Venkat Reddy

Contents

- What are Descriptive statistics
- Frequency tables and graphs, Histograms
- Central Tendency
- Mean, Median, Mode
- Dispersion
- Range, variance, standard deviation
- Quartiles, Percentiles
- Box Plots
- Bivariate Descriptive Statistics
  - Contingency Tables
  - Correlation
  - Regression

Why Descriptive statistics?

- Who is a better ODI batsmen - Sachin or Muralidharan?
  - Batting average?
- Who is the reliable- Dhoni or Afridi?
  - Score variance
- A triangular series among Aus, Eng & Newziland ; Who will win?
  - Most number of wins - Mode
- I am going to buy shoes. Which brand has verity- Power or Adidas?
  - Price range - Range

- We used Average, Variance, Mode, Range to make some inferences. These are nothing but descriptive statistics
- Descriptive statistics tell us what happened in the past.
- Descriptive statistics avoid inferences but, they help us to get a feel of the data.
- Some times they are good enough to make an inference.

Descriptive Statistics

- A statistic or a measure that describes the data
  - Average salary of employees
- Describing data with tables and graphs (quantitative or categorical variables)
- Numerical descriptions
  - Center – Give some example measures of center of the data
  - Variability– Give some example measures of variability of the data
- Bivariate descriptions (In practice, most studies have several variables)
  - Dependency measures(Correlation)

Simple Descriptive Statistics

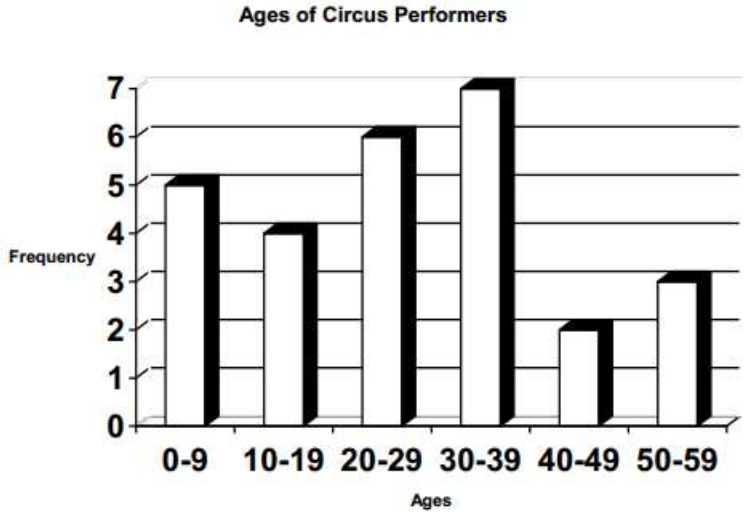
- N
- Sum
- Min
- Max
- Average
- Frequency of each level
- Variance
- Standard deviation

These simple descriptive statistics will be use in inferential statistics later.

Frequency tables & Histograms

- Frequency distribution: Lists possible values of variable and number of times each occurs

Ages of Circus Performers		
Intervals	Tally Marks	Frequency
0 – 9	𝍸	5
10 – 19	||||	4
20 – 29	𝍸 |	6
30 – 39	𝍸 ||	7
40 – 49	||	2
50 – 59	|||	3



Shapes of histograms

- Bell-shaped (IQ, SAT, political ideology in all U.S.)
- Skewed right
  - Example Annual income
  - No. times arrested
- Skewed left
  - Score on easy exam
  - Daily level if excitement in office
- Bimodal
  - Hardworking days in a year (Peaks near Mid year & year end Appraisal)

Lab : Histogram

- Create a histogram on variable ‘actual’ in prdsale data
  - How many modes?
  - What is the skewness?
  - What is its kurtosis?
- Create a histogram on variable ‘msrp’ in cars data
  - How many modes?
  - What is the skewness?
  - What is its kurtosis?
- Create a histogram on variable ‘weight’ in cars data
  - How many modes?
  - What is the skewness?
  - What is its kurtosis?

Compare the above three histograms.

Central tendency

- What is the flight fare from Bangalore to Delhi? 3500–Exact or average?
- What is central tendency? - Average
- Three types of Averages
  - Mean
  - Median
  - Mode

Mean

- Center of gravity
- Evenly partitions the sum of all measurement among all cases; average of all measures

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

- Crucial for inferential statistics
- Mean is not very resistant to outliers –See in Median

Median

- What is the mean of [0.1 0.8 0.4 0.3 0.1 0.4 9.0 0.1 0.9 0.3 1.0 0.3 0.1]
- Guess without calculation – Around **0.5**?
- Now calculate the mean

- Median is exactly in the middle. Isn't mean exactly in the middle
- Order the observations in ascending or descending order and pick the middle observation
- less useful for inferential purposes
- More resistant to effects of outliers…

Calculation of Median

rim diameter (cm)

unit 1	unit 2	
9.7	9.0	
11.5	11.2	
11.6	11.3	
12.1	11.7	
12.4	12.2	
12.6	12.5	
12.9 <--	13.2	13.2
13.1	13.8	
13.5	14.0	
13.6	15.5	
14.8	15.6	
16.3	16.2	
26.9	16.4	

Mode

- How do you express average size of the shoes ?
  - 6.567 or 6?

- Mode is the most numerous category
- Can be more or less created by the grouping procedure
- For theoretical distributions—simply the location of the peak on the frequency distribution

Lab

- Run Proc means data product data
- What is the mean of ‘msrp’ in cars data?
- Is it reflecting the average value of price?
- What is median of ‘msrp’ in cars data?
- Is it reflecting the average value of price?
- Run Proc Univariate on weight varaibale in cars data. Find mean, Median & Mode.

Dispersion

Person1: What is the average depth of this river? 5 feet

Person2: I am 5.5 I can easily cross it(and starts crossing it)

Person 2: Help….help.

Person 1: Some times just knowing the central tendency is not sufficient

- Measures of dispersion summarize the degree of clustering/spread of cases, esp. with respect to central tendency…
  - range
  - variance
  - standard deviation

Range

- Max –Min

R: range(x)

unit 1	unit 2
9.7	9.0
11.5	11.2
11.6	11.3
12.1	11.7
12.4	12.2
12.6	12.5
13.1	13.2
13.5	13.8
13.6	14.0
14.8	15.5
16.3	15.6
26.9	16.2
	16.4

Variance

- Take deviation from Mean- It can be zero some times
- Hence take square of deviation from mean → Take average of that
- Average mean squared distance is **variance**

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

- Units of variance are squared… this makes variance hard to interpret
- Eg : Mean length = 22.6 mm variance = 38 mm^2
- What does this mean??? –I don't Know

Standard Deviation

- Square root of variance

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

- Units are in same units as base measurements
- Mean = 22.6 mm standard deviation = 6.2 mm
- Mean +/- sd (16.4—28.8 mm)
  - should give at least some intuitive sense of where most of the cases lie, barring major effects of outliers

Quartiles & Percentiles

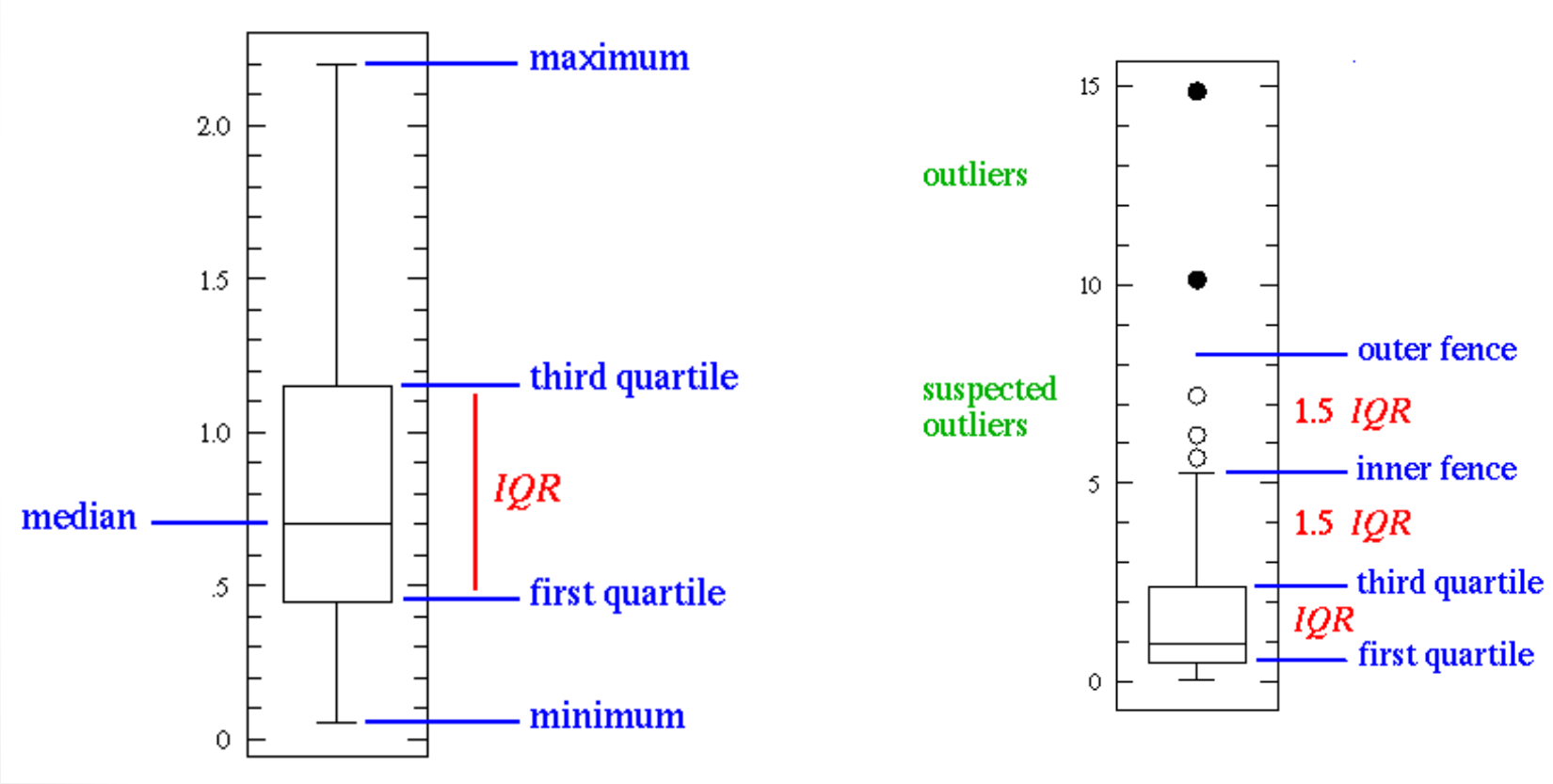
- pth percentile: p percent of observations below it, (100 - p)% above it.
- Like 95% of CAT percentile means→ 5% are above & 95% are below
- 1,2,3,4,5,6,7,8,9,10 - What is 25th percentile?
- 1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20 - What is 25th percentile? What is 80th percentile?

  - p = 50: median
  - p = 25: lower quartile (LQ)
  - p = 75: upper quartile (UQ)

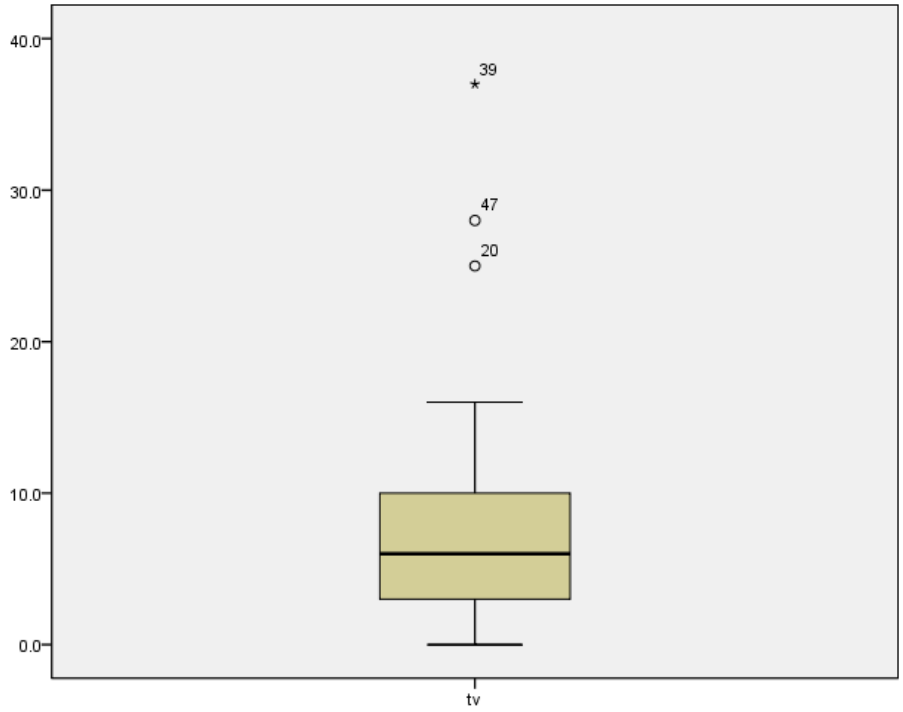
- Interquartile range IQR = UQ - LQ

Box Plots

- Quartiles portrayed graphically by box plots



Box Plots



Example: weekly TV watching for *n*=60, 3 outliers

Box Plots Interpretation

- Box plots have box from LQ to UQ, with median marked. They portray a five-number summary of the data: Minimum, LQ, Median, UQ, Maximum
- Except for outliers identified separately
- **Outlier** = observation falling

  below LQ – 1.5(IQR) or above UQ + 1.5(IQR)
- Ex. If LQ = 2, UQ = 10, then IQR = 8 and outliers above 10 + 1.5(8) = 22

Lab

- Run proc univariate on a variable from sample data in sas default library(prd sale / cars)
- Run proc means on actual & predicted variables from product sales data
- What are the values of Range, Variance, SD
- What are 1,2,3 & 4 quartile values
- What is 95^{th} percentile?
- Use “all” option to display the box plots

Contingency Tables

- Cross classifications of categorical variables in which rows (typically) represent categories of explanatory variable and columns represent categories of response variable.
- Counts in "cells" of the table give the numbers of individuals at the corresponding combination of levels of the two variables

Example: Happiness and Family Income of 1993 families (GSS 2008 data: "happy," "finrela")

	Happiness			
Income	Very	Pretty	Nottoo	Total
Above Aver.	164	233	26	423
Average	293	473	117	883
Below Aver.	132	383	172	687
Total	589	1089	315	1993

Contingency tables

- Example: Percentage “very happy” is
  - 39% for above average income (164/423 = 0.39)
  - 33% for average income (293/883 = 0.33)
  - What percent for below average income?

	Happiness			
Income	Very	Pretty	Not oo	Total
Above	164 (39%)	233 (55%)	26 (6%)	423
Average	293 (33%)	473 (54%)	117 (13%)	883
Below	132 (19%)	383 (56%)	172 (25%)	687

- What can we conclude? Is happiness depending on Income? Or Happiness is independent of Income?
- Inference questions for later chapters?

Correlation

- **Correlation** describes strength of association between two variables
- Falls between -1 and +1, with sign indicating direction of association (formula & other details later)
- The larger the correlation in absolute value, the stronger the association (in terms of a straight line trend)
- **Examples**: (positive or negative, how strong?)
  - Mental impairment and life events, correlation =
  - GDP and fertility, correlation =
  - GDP and percent using Internet, correlation =

Strength of Association

- Correlation 0 →No linear association
- Correlation 0 to 0.25 →Negligible positive association
- Correlation 0.25-0.5 → Weak positive association
- Correlation 0.5-0.75 →Moderate positive association
- Correlation >0.75 →Very Strong positive association
- What are the limits for negative correlation


Correlation $r = 0$


Correlation $r = -0.3$


Correlation $r = 0.5$


Correlation $r = -0.7$


Correlation $r = 0.9$


Correlation $r = -0.99$

Regression

- **Regression analysis** gives line predicting *y* using *x*(algorithm & other details later)

- *y* = college GPA, *x* = high school GPA
- Predicted y = 0.234 + 1.002(x)

Lab

- Create a contingency table for product sales data
- Find contingency tables for
  - Region by product type
  - Division by Product type
- Find the correlation between actual sales and predicted sales.
- Find the correlation between weight & msrp in cars data

Venkat Reddy Konasani

Manager at Trendwise Analytics

venkat@TrendwiseAnalytics.com

21.venkat@gmail.com

+91 9886 768879

www.TrendwiseAnalytics.com/venkat