

# 定量分析：数据思维与 商业统计

陈文波

18986118886

[cwb@whu.edu.cn](mailto:cwb@whu.edu.cn)

2021年10月

# 班级群

- 请大家下载“学习通”App
- 用学习通扫码，加入班级群
- 课程组织
  - Ppt，数据
  - 作业
  - 互动与点名、签到



# 旧闻两则：其一

## “大跌眼镜”！击败奥巴马，特朗普获评“最受尊敬男性”

2020年12月31日 04:14 环球时报



原标题：“大跌眼镜”！击败奥巴马，特朗普获评“最受尊敬男性”

[环球时报记者 刘浩然]美国民调机构盖洛普公司网站29日公布的一项民调结果让很多人“大跌眼镜”。民调显示，美国总统特朗普被评选为2020年“最受尊敬男性”，击败此前连续12年获得该头衔的美国前总统奥巴马。


美国《纽约邮报》30日报道称，这项民调结果是盖洛普通过电话形式对美国民众随机采访得出的。特朗普以18%的支持率位列榜首，奥巴马以15%的支持率位居第二，美国当选总统拜登则以6%的支持率排在第三。报道称，从党派类别来看，48%的共和党支持者选择了特朗普，而在另一边的民主党阵营里，32%的支持者选择奥巴马，13%的人选了拜登。而独立派人士中，特朗普和奥巴马“平分秋色”，各占11%。

### 新浪热榜

- 1 航天员带的过年饺子有三种馅 721.0万
- 2 #神舟十三号# 635.6万
- 3 #神十三发射圆满成功# 523.6万
- 4 #中国空间站第二批客人来了# 504.5万
- 5 #女航天员在太空来例假了怎么… 417.2万
- 6 小伙用5000张扑克摆出王亚平… 388.0万
- 7 神十三飞天与中国原子弹首爆同日 336.3万
- 8 #20岁女子被喊大姐与保安起争… 330.1万
- 9 #我国7次载人航天飞行高光时刻# 327.6万


# 旧闻两则：其一

在关于“最受尊敬女性”评选结果中，美国前第一夫人米歇尔以10%的支持率夺得桂冠，这是她连续第三年获此荣誉。紧随其后的是美国当选副总统哈里斯、美国现任第一夫人梅拉尼娅、“脱口秀女王”奥普拉以及德国总理默克尔。

这项民调结果一经发布就在社交媒体“推特”上引发广泛争论，不少网民更是对特朗普当选表示质疑。有网民在“推特”评论区打出10个问号：“最受尊敬的人是谁？”“他们只采访了3.3亿美国人中的1018人。这不是民意调查，这就是个毫无价值的笑话。”有网民甚至表示：“尊敬他什么？他搞砸了防疫、经济和民主制度，离受尊敬还差得远呢。”也有网民称特朗普在新冠疫情期间“毫无作为”，这恰好说明“美国人有多盲目”，“他们尊敬一个撒谎、创下新冠死亡病例纪录的人，他不会在人们需要他的时候站出来去做一些抗疫的事情。人们怎么会相信他？”  [返回搜狐，查看更多](#)

# 旧闻两则：其一

他们只采访了3.3亿美国人中的1018人。这不是名义调查，这就是个毫无价值的笑话

表示质疑。有网民在“推特”评论区打出10个问号：“最受尊敬的人是谁？”“他们只采访了3.3亿美国人中的1018人。这不是民意调查，这就是个毫无价值的笑话。”有网民甚至表示：“尊敬他什么？他搞砸了防疫、经济和民主制度，离受尊敬还差得远呢。”也有网民称特朗普在新冠疫情期间“毫无作为”，这恰好说明“美国人有多盲目”，“他们尊敬一个撒谎、创下新冠死亡病例纪录的人，他不会在人们需要他的时候站出来去做一些抗疫的事情。人们怎么会相信他？”  [返回搜狐，查看更多](#)



# 旧闻两则：其二

赌注中位数下跌平均数提高 本地更多赌客越赌越大 - Mozilla Firefox

文件(F) 编辑(E) 查看(V) 历史(S) 书签(B) 工具(T) 帮助(H)

Google Gmail - 收件箱 (11,631) - jjcheer@... 联合早报网 zaobao.com 赌注中位数下跌平均数提高 本地更多... 广东新周刊杂志社

www.zaobao.com/sp/sp120224\_008.shtml

Delicious Google W 维基 CNKI W-Lib Gmail G-Scholar Vsharing 计世网 AIS 論文 G-Book W-VPN CISR Evernote Web E-Jou G-R NetLibrary EBSCO ZHI?

**联合早报网** **zaobao.com** 即时 文萃 新加坡 中国 国际 东南亚 体育 副刊 全检索 观点 早点 专评 名采 图片 漫画 来信 投票 专题 人物 特写 企业 财经 中国 全球 狮城 综述 人物 股市 投资 重庆 成都 济南 常州 体坛 星闻 丽人 教育 房产 健康 旅游 汽车 时尚 科技 汇点 读书 论坛 语录 出国 微博 Twitter 语录 逗号 大拇指 English 热点 译名 新南洋 RSS

首页 >> 新闻 >> 新加坡 [buy English translation]

## 赌注中位数下跌平均数提高 本地更多赌客越赌越大

(2012-02-24)

**早报导读**

- [名采文评] 毛尖：给姐烧个哥
- [名家专评] 谁来领导世界银行？
- [时事漫画] 吉拉德宣布党内投票选党魁
- [中国早点] 丢失的脚踏车

李太里 制图  
陈斌勤 摄影

这项在去年5月至8月间进行的调查显示，赌客每月下注的中位数从2008年的100元下降到去年的40元，但平均赌注却从176元增加到212元。潜在病态赌徒去年每月的平均下注额高达1713元，比2008年的619元增加将近两倍。他们下注的中位数也从2008年的450元增加到581元。

全国预防嗜赌理事会（NCPG）的调查显示，大部分受访者每月花费的下注额低于100元，但由于下注额较大的赌客比例有所增加，特别是低收入者，进而推高了去年的平均下注额。

**即时报道** 更多>>>

- (1330)[新加坡]GIC买入邦基贸易5%股权
- (1320)[新加坡]1月制造业产值同比跌8.8%
- (1310)[港澳]曾荫权深圳租下顶层复式大宅
- (1300)[国际]华盛顿大教堂耗2000万美元复修
- (1255)[国际]美国冻结日本“山口组”首领资产
- (1245)[国际]戴维斯：正在讨论扩大对朝“营养援助”问题
- (1230)[新加坡]王瑞杰：教育基金是全方位的

**体坛风云** 拓荒者狂胜马刺40分

**环球星闻** 刘德华新书透露时刻发着导演梦

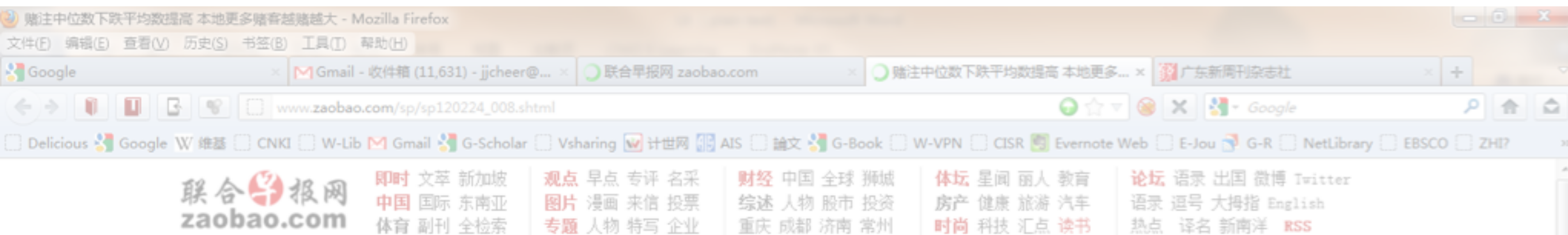
**热门排行** **早报读书** 更多>>>

[分类阅读] 言情 | 玄幻 | 武侠 | 历史 | 游戏 | 科幻

- 大明官途之窃国 [历史]
- 权柄之皇图霸业 [历史]
- 邪少的亿万女人 [都市/言情]
- 大争之世(大结局) [历史]

正在从 s.sdl.sg 传送数据...

# 旧闻两则：其二



这项在去年5月至8月间进行的调查显示，赌客每月下注的中位数额从2008年的100元下降到去年的40元，但平均赌注却从176元增加到212元...



“is the discipline that concerns the collection, organization, analysis, interpretation, and presentation of **data**.”

*–Wikipedia*



# Statistics

---

- Statistics originated from the Latin word “Status” meaning “State”.
- Solely with the displays of data and charts pertaining to the economic, demographic, and political situations prevailing in a country.
- ...

# Applications

---

## Accounting

- Public accounting firms use statistical sampling procedures when conducting audits for their clients.

## Economics

- Economists use statistical information in making forecasts about the future of the economy or some aspect of it.

## Finance

- Financial advisors use price-earnings ratios and dividend yields to guide their investment advice.

# Applications

---

## Marketing

- Electronic point-of-sale scanners at retail checkout counters are used to collect data for a variety of marketing research applications.

## Production

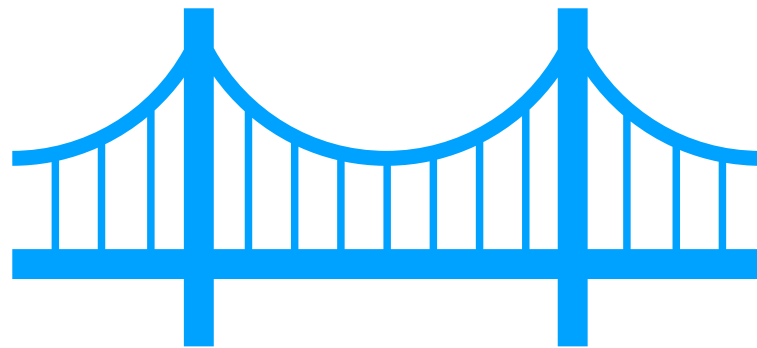
- A variety of statistical quality control charts are used to monitor the output of a production process.

## Information Systems

- A variety of statistical information helps administrators assess the performance of computer networks.

# 课程梗概

描述性分  
析与EDA

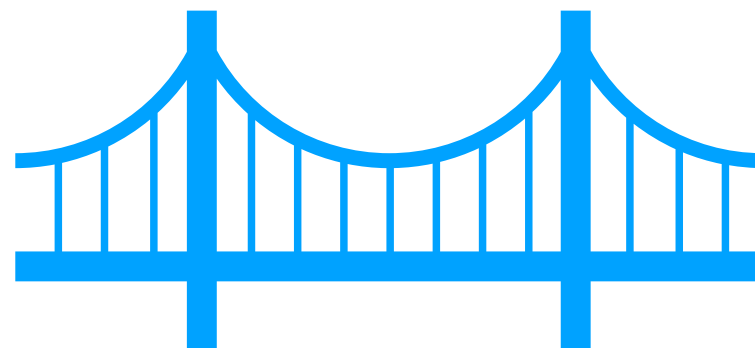


推断性  
分析

# 课程梗概

## 描述性分析与EDA

- 数值方法：三类指标
- 图形方法：五种图形



- 抽样分布、CLT
- 正态分布
- T分布
- 卡方分布
- F分布

## 推断性分析

- 区间估计
- 假设检验
- 应用：
  - 分类 vs. 分类
  - 分类 vs. 数值
  - 数值 vs. 数值
  - 数值 vs. 分类

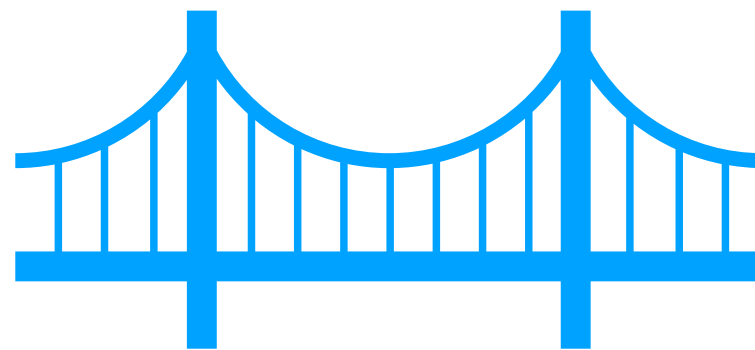


# 课程梗概

## 描述性分析与EDA

- 数值方法：三类指标
- 图形方法：五种图形

**R**



- 抽样分布、CLT
- 正态分布
- T分布
- 卡方分布
- F分布

## 推断性分析

- 区间估计
- 假设检验
- 应用：
  - 分类 vs. 分类
  - 分类 vs. 数值
  - 数值 vs. 数值

# 课程梗概

---

- Session 1 : 内容简介、基本概念、描述性统计1
- Session 2: 描述性统计2
- Session 3: 抽样、中心极限定理、正态分布
- Session 4: 其他三个分布, 区间估计
- Session 5: 假设检验, 两个总体均值和比例的推断
- Session 6: 总体方差和方差比的推断、分类数据 Vs. 数值型数据: 卡方检验与列联表分析
- Session 7: 分类数据 Vs. 数值型数据: 方差分析; 数值型数据: 回归分析
- Session 8: 课程总结

# 参考教材

---

戴维 R.安德森 (David R.Anderson) , & 丹尼斯 J.斯威尼. (2017). 商务与经济统计 (张建华, 王健, & 聂巧平, Trans.; 13版). 机械工业出版社. <https://item.jd.com/12091283.html>

# Grading Policy

---

- 课堂参与： 20%
- 课后作业： 30%
- 最终考试： 50%

Q & A



# Data and Data Sets

---

- Data are the facts and figures collected, analyzed, and summarized for presentation and interpretation.
- All the data collected in a particular study are referred to as the data set for the study.

# Elements, Variables, and Observations

- Elements are the entities on which data are collected.
- A variable is a characteristic of interest for the elements.
- The set of measurements obtained for a particular element is called an observation.
- A data set with  $n$  elements contains  $n$  observations.
- The total number of data values in a complete data set is the number of elements multiplied by the number of variables.

# Data, Data Sets, Elements, Variables, and Observations

The diagram illustrates the relationship between data elements, variables, observations, and a data set. A table contains five rows of data. A bracket on the left labeled 'Element Names' spans the 'Company' column. A bracket above the table labeled 'Variables' spans the 'Stock Exchange', 'Annual Sales (\$M)', and 'Earnings per share (\$)' columns. A bracket on the right labeled 'Observation' spans the entire row for 'Dataram'. A bracket on the right labeled 'Data Set' spans the entire table. A line connects the 'Annual Sales (\$M)' value for 'Keystone' to the 'Data Set' label.

Company	Variables		
	Stock Exchange	Annual Sales (\$M)	Earnings per share (\$)
Dataram	NQ	73.10	0.86
EnergySouth	N	74.00	1.67
Keystone	N	365.70	0.86
LandCare	NQ	111.40	0.33
Psychemedics	N	17.60	0.13

# Scales of Measurement

- Scales of measurement include
  - Nominal
  - Ordinal
  - Interval
  - Ratio
- The scale determines the amount of information contained in the data.
- The scale indicates the data summarization and statistical analyses that are most appropriate.

# Scales of Measurement

## Nominal scale

- Data are labels or names used to identify an attribute of the element.
- A nonnumeric label or numeric code may be used.

## Example

Students of a university are classified by the school in which they are enrolled using a nonnumeric label such as Business, Humanities, Education, and so on.

Alternatively, a numeric code could be used for the school variable (e.g. 1 denotes Business, 2 denotes Humanities, 3 denotes Education, and so on).



# Scales of Measurement

## Ordinal scale

- The data have the properties of nominal data and the order or rank of the data is meaningful.
- A nonnumeric label or numeric code may be used.

## Example

Students of a university are classified by their class standing using a nonnumeric label such as Freshman, Sophomore, Junior, or Senior.

Alternatively, a numeric code could be used for the class standing variable (e.g. 1 denotes Freshman, 2 denotes Sophomore, and so on).

# Scales of Measurement

## Interval scale

- The data have the properties of ordinal data, and the interval between observations is expressed in terms of a fixed unit of measure.
- Interval data are always numeric.

## Example

Melissa has an SAT score of 1985, while Kevin has an SAT score of 1880. Melissa scored 105 points more than Kevin.

# Scales of Measurement

## Ratio scale

- Data have all the properties of interval data and the ratio of two values is meaningful.
- Ratio data are always numerical.
- Zero value is included in the scale.

### Example:

Price of a book at a retail store is \$ 200, while the price of the same book sold online is \$100. The ratio property shows that retail stores charge twice the online price.

# Categorical and Quantitative Data

- Data can be further classified as being categorical or quantitative.
- The statistical analysis that is appropriate depends on whether the data for the variable are categorical or quantitative.
- In general, there are more alternatives for statistical analysis when the data are quantitative.

# Categorical Data

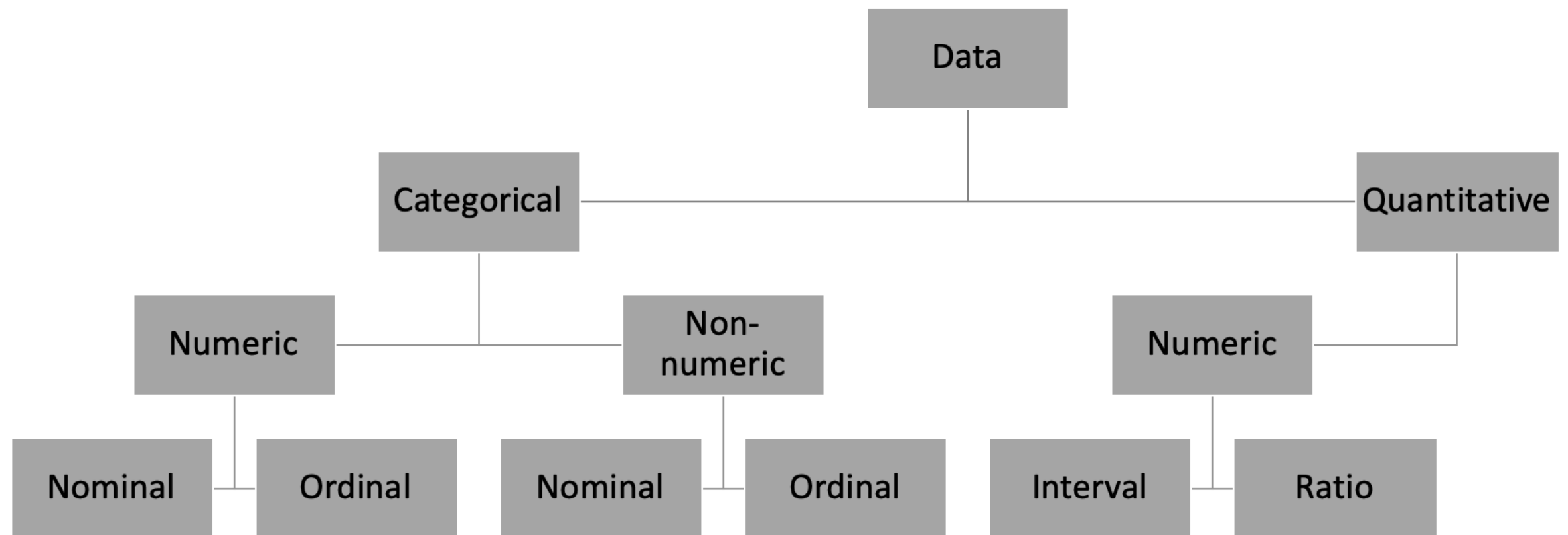
- Labels or names are used to identify an attribute of each element
- Often referred to as qualitative data
- Use either the nominal or ordinal scale of measurement
- Can be either numeric or nonnumeric
- Appropriate statistical analyses are rather limited



# Quantitative Data

- Quantitative data indicate how many or how much.
- Quantitative data are always numeric.
- Ordinary arithmetic operations are meaningful for quantitative data.

# Scales of Measurement



# Cross-Sectional Data

Cross-sectional data are collected at the same or approximately the same point in time.

## Example

Data detailing the number of building permits issued in November 2013 in each of the counties of Ohio.

# Time Series Data

Time series data are collected over several time periods.

## Example

Data detailing the number of building permits issued in Lucas County, Ohio in each of the last 36 months.

Graphs of time series data help analysts understand

- what happened in the past
- identify any trends over time, and
- project future levels for the time series

# Data Sources

## Existing Sources

- Internal company records – almost any department
- Business database services – Dow Jones & Co.
- Government agencies - U.S. Department of Labor
- Industry associations – Travel Industry Association of America
- Special-interest organizations – Graduate Management Admission Council (GMAT)
- Internet – more and more firms

# Data Sources

## Data Available From Internal Company Records

Record	Some of the Data Available
Employee records	Name, address, social security number
Production records	Part number, quantity produced, direct labor cost, material cost
Inventory records	Part number, quantity in stock, reorder level, economic order quantity
Sales records	Product number, sales volume, sales volume by region
Credit records	Customer name, credit limit, accounts receivable balance
Customer profile	Age, gender, income, household size

# Data Sources

## Data Available From Selected Government Agencies

Government Agency	Web address	Some of the Data Available
Census Bureau	<a href="http://www.census.gov">www.census.gov</a>	Population data, number of households, household income
Federal Reserve Board	<a href="http://www.federalreserve.gov">www.federalreserve.gov</a>	Data on money supply, exchange rates, discount rates
Office of Mgmt. & Budget	<a href="http://www.whitehouse.gov/omb">www.whitehouse.gov/omb</a>	Data on revenue, expenditures, debt of federal government
Department of Commerce	<a href="http://www.doc.gov">www.doc.gov</a>	Data on business activity, value of shipments, profit by industry
Bureau of Labor Statistics	<a href="http://www.bls.gov">www.bls.gov</a>	Customer spending, unemployment rate, hourly earnings, safety record

# Data Sources

## Statistical Studies – Observational

- In observational (nonexperimental) studies no attempt is made to control or influence the variables of interest.
- Example - Survey
- Studies of smokers and nonsmokers are observational studies because researchers do not determine or control who will smoke and who will not smoke.



# Data Sources

## Statistical Studies – Experimental

- In experimental studies the variable of interest is first identified. Then one or more other variables are identified and controlled so that data can be obtained about how they influence the variable of interest.
- The largest experimental study ever conducted is believed to be the 1954 Public Health Service experiment for the Salk polio vaccine. Nearly two million U.S. children (grades 1- 3) were selected.

# Data Acquisition Considerations

## Time Requirement

- Searching for information can be time consuming.
- Information may no longer be useful by the time it is available.

## Cost of Acquisition

- Organizations often charge for information even when it is not their primary business activity.

## Data Errors

- Using any data that happen to be available or were acquired with little care can lead to misleading information.

# Descriptive Statistics

- Most of the statistical information in newspapers, magazines, company reports, and other publications consists of data that are summarized and presented in a form that is easy to understand.
- Such summaries of data, which may be tabular, graphical, or numerical, are referred to as descriptive statistics.

## Example

The manager of Hudson Auto would like to have a better understanding of the cost of parts used in the engine tune-ups performed in her shop. She examines 50 customer invoices for tune-ups. The costs of parts, rounded to the nearest dollar, are listed on the next slide.

# Example: Hudson Auto Repair

Sample of Parts Cost (\$) for 50 Tune-ups

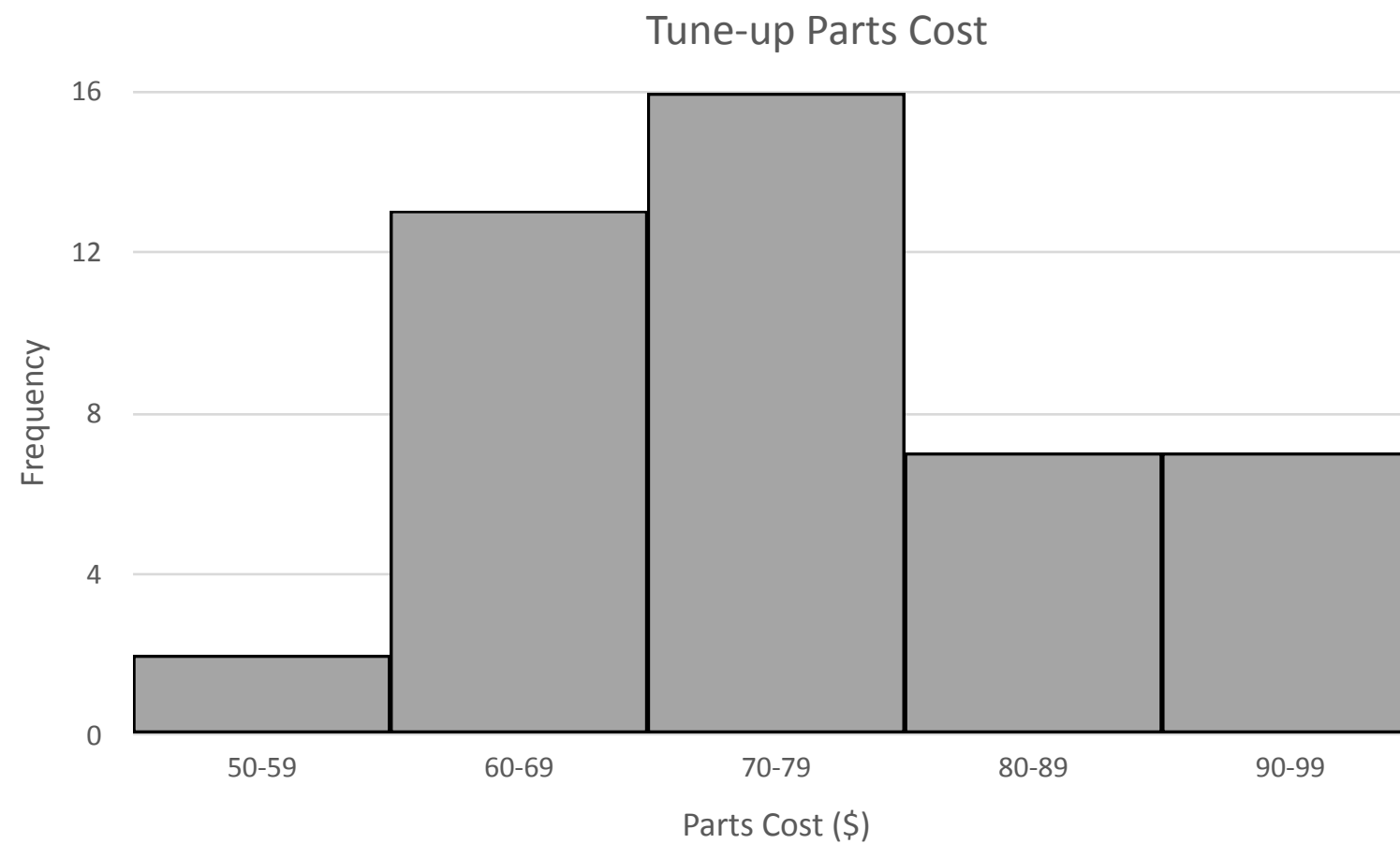
91	78	93	57	75	52	99	80	97	62
71	69	72	89	66	75	79	75	72	76
104	74	62	68	97	105	77	65	80	109
85	97	88	68	83	68	71	69	67	74
62	82	98	101	79	105	79	69	62	73

## Tabular Summary: Frequency and Percent Frequency

Parts Cost (\$)	Frequency	Percent Frequency
50-59	2	4%
60-69	13	26%
70-79	16	32%
80-89	7	14%
90-99	7	14%
100-109	5	10%
<b>TOTAL</b>	<b>50</b>	<b>100%</b>

# Graphical Summary: Histogram

Example: Hudson Auto



# Numerical Descriptive Statistics

- The most common numerical descriptive statistic is the mean (or average).
- The mean demonstrates a measure of the central tendency, or central location of the data for a variable.
- Hudson's mean cost of parts, based on the 50 tune-ups studied is \$79 (found by summing up the 50 cost values and then dividing by 50).

# Statistical Inference

**Population:** The set of all elements of interest in a particular study.

**Sample:** A subset of the population.

**Statistical inference:** The process of using data obtained from a sample to make estimates and test hypotheses about the characteristics of a population.

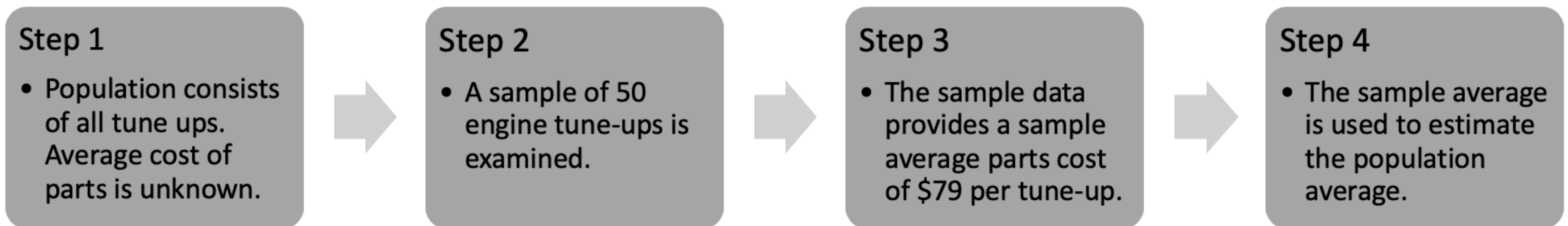
**Census:** Collecting data for the entire population.

**Sample survey:** Collecting data for a sample.



# Process of Statistical Inference

Example: Hudson Auto



# Analytics

Analytics is the scientific process of transforming data into insight for making better decisions.

## Techniques:

- Descriptive analytics: This describes what has happened in the past.
- Predictive analytics: Use models constructed from past data to predict the future or to assess the impact of one variable on another.
- Prescriptive analytics: The set of analytical techniques that yield a best course of action.

# Big data and Data Mining:

Big data: Large and complex data set.

Three V's of Big data:

⇒ Volume : Amount of available data

⇒ Velocity: Speed at which data is collected and processed

⇒ Variety: Different data types

# Data warehousing

Data warehousing is the process of capturing, storing, and maintaining the data.

- Organizations obtain large amounts of data on a daily basis by means of magnetic card readers, bar code scanners, point of sale terminals, and touch screen monitors.
- Wal-Mart captures data on 20-30 million transactions per day.
- Visa processes 6,800 payment transactions per second.

# Data Mining

- Methods for developing useful decision-making information from large databases.
- Using a combination of procedures from statistics, mathematics, and computer science, analysts “mine the data” to convert it into useful information.
- The most effective data mining systems use automated procedures to discover relationships in the data and predict future outcomes prompted by general and even vague queries by the user.

# Data Mining Applications

- The major applications of data mining have been made by companies with a strong consumer focus such as retail, financial, and communication firms.
- Data mining is used to identify related products that customers who have already purchased a specific product are also likely to purchase (and then pop-ups are used to draw attention to those related products).
- Data mining is also used to identify customers who should receive special discount offers based on their past purchasing volumes.

# Data Mining Requirements

- Statistical methodology such as multiple regression, logistic regression, and correlation are heavily used.
- Also needed are computer science technologies involving artificial intelligence and machine learning.
- A significant investment in time and money is required as well.

# Data Mining Model Reliability

- Finding a statistical model that works well for a particular sample of data does not necessarily mean that it can be reliably applied to other data.
- With the enormous amount of data available, the data set can be partitioned into a training set (for model development) and a test set (for validating the model).
- There is, however, a danger of overfitting the model to the point that misleading associations and conclusions appear to exist.
- Careful interpretation of results and extensive testing is important.



# Ethical Guidelines for Statistical Practice

- In a statistical study, unethical behavior can take a variety of forms including:
  - Improper sampling
  - Inappropriate analysis of the data
  - Development of misleading graphs
  - Use of inappropriate summary statistics
  - Biased interpretation of the statistical results
- One should strive to be fair, thorough, objective, and neutral as you collect, analyze, and present data.
- As a consumer of statistics, one should also be aware of the possibility of unethical behavior by others.

# Ethical Guidelines for Statistical Practice

- The American Statistical Association developed the report “Ethical Guidelines for Statistical Practice”.
- It contains 67 guidelines organized into 8 topic areas:
  - Professionalism
  - Responsibilities to Funders, Clients, Employers
  - Responsibilities in Publications and Testimony
  - Responsibilities to Research Subjects
  - Responsibilities to Research Team Colleagues
  - Responsibilities to Other Statisticians/Practitioners
  - Responsibilities Regarding Allegations of Misconduct
  - Responsibilities of Employers Including Organizations, Individuals, Attorneys, or Other Clients

# Wrap-up

---

- 课程说明
- 基本概念
  - 数据、数据集
  - 计量的尺度、分类
  - 样本、总体、抽样、描述与推断