Probability and Statstics

Dr. Syed Faisal Bukhari

Associate Professor

Department of Data Science

Faculty of Computing and Information Technology

University of the Punjab

Textbooks

- **Probability & Statistics for Engineers & Scientists**, Ninth Edition, Ronald E. Walpole, Raymond H. Myer

- **Elementary Statistics: Picturing the World,** 6th Edition, Ron Larson and Betsy Farber

- **Elementary Statistics,** 13th Edition, Mario F. Triola

Reference books

- **Probability and Statistical Inference, Ninth Edition,** Robert V. Hogg, Elliot A. Tanis, Dale L. Zimmerman
- **Probability Demystified**, Allan G. Bluman
- **Schaum's Outline of Probability,** Second Edition, Seymour Lipschutz, Marc Lipson
- **Python for Probability, Statistics, and Machine Learning,** José Unpingco
- **Practical Statistics for Data Scientists: 50 Essential Concepts,** Peter Bruce and Andrew Bruce
- **Think Stats: Probability and Statistics for Programmers,** Allen Downey

References

Readings for these lecture notes:

- **Elementary Statistics: Picturing the World,** 6th Edition, Ron Larson and Betsy Farber
- **Probability & Statistics for Engineers & Scientists**, Ninth Edition, Ronald E. Walpole, Raymond H. Myer
- **Probability Demystified**, Allan G. Bluman
- **Practical Statistics for Data Scientists: 50 Essential Concepts,** Peter Bruce and Andrew Bruce
- https://www.mymarketresearchmethods.com/types-of-data-nominal-ordinal-interval-ratio/
- http://www.thefreedictionary.com/statistics

These notes contain material from the above three resources.

Distribution of marks

- **Mid term** = 35 points
- **Final term** = 40 points
- **Sessional marks** = 25 points
  I. **Assignments** = 1 × 3 = 3 points
  II. **Quizzes** = 1.5 × 8 = 12 points
  III. **A survey based project presentation** = 10 points

Target Journals

Some of the journals that are relevant to health care and the medical field, based on computer science.

1. **Medical Decision Making**, JCR Impact Factor (2017-18) = 2.793
2. **Health Informatics Journal**, JCR Impact Factor (2017-18) = 2.297
3. **Informatics for Health and Social Care**, JCR Impact Factor (2017-18) = 1.137
4. **Health Care Analysis**, , JCR Impact Factor (2017-18) = 1.043
5. **International Journal of Health Care Quality Assurance**, JCR Impact Factor (2017-18) = 1.218

Target Journals

Some of the journals that are relevant to education, based on computer science.

1. **Computers & Education**, JCR Impact Factor (2017-18) = 5.627
2. **Computer Applications in Engineering Education**, JCR Impact Factor (2017-18) = **1.435**
3. **Journal of Computing in Higher Education**, JCR Impact Factor (2017-18) = **1.870**
4. **Acm Transactions on Computing Education**, , JCR Impact Factor (2017-18) = **1.356**
5. **Assessment & Evaluation In Higher Education**, JCR Impact Factor (2017-18) = **2.473**
6. **Educational Assessment Evaluation and Accountability**, JCR Impact Factor (2017-18) = **1.772**
7. **Computer Applications in Engineering Education = Impact Factor: 1.435**

Basic concepts [1]

Statistics is defined as

"The mathematics of the **collection**, **organization**, and

interpretation of **numerical data**, especially the analysis of population characteristics by inference from sampling"

OR

Statistics is a science which deals with collection, classification, distribution and interpretation of data.

OR

Statistics is a science of **uncertainty**.

OR

Statistics is the science of collecting, organizing, analyzing, and interpreting data in order to make decisions.

Data sets

Data consist of information coming from observations, counts, measurements, or responses.

Statistics is the science of collecting, organizing, analyzing, and interpreting data in order to make decisions.

There are **two types of data sets** you will use when studying statistics. These data sets are called **populations** and **samples.**

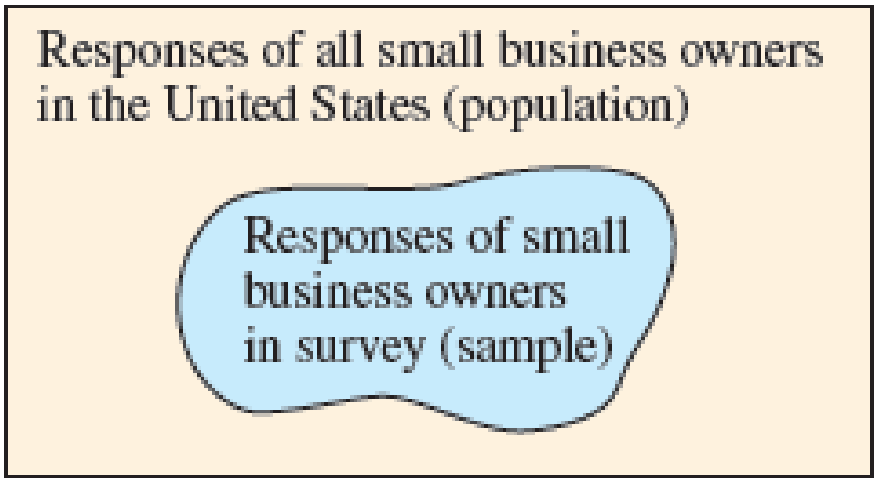
- A **population** is the collection of *all* outcomes, responses, measurements, or counts that are of interest.

- A **sample** is a subset, or part, of a population.

Identifying Data Sets

In a recent survey, **614 small business owners** in the United States were asked whether they thought their **company's Facebook presence** was valuable. **Two hundred fifty-eight (258)** of the **614 respondents** said **yes**. Identify the population and the sample. Describe the sample data set.

$\hat{p} = \frac{258}{614} = 0.4025$



`

Solution:

- The **population consists** of the **responses of all small business owners** in the United States, and the **sample consists of the responses** of the 614 small business owners in the survey.

- **Notice that the sample is a subset of the responses of all small business owners** in the United States. The sample data set consists of **258 owners** who said **yes** and **356 owners** who said **no**.

Descriptive Statistics vs. Inferential Statistics

- The study of statistics has two major branches: **descriptive statistics** and **inferential statistics.**

- **Descriptive statistics** is the branch of statistics that involves the **organization**, **summarization**, and **display of data**.

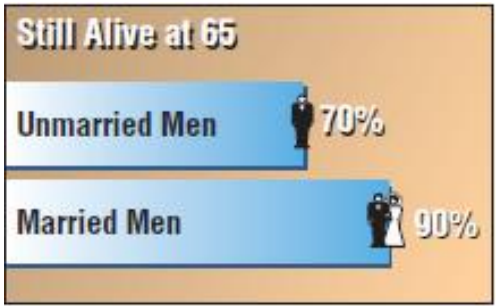
- **Inferential statistics** is the branch of statistics that involves using a **sample to draw conclusions** about a **population**. A basic tool in the study of **inferential statistics** is **probability**.

Descriptive and Inferential Statistics

Example : Determine which part of the study represents the **descriptive branch** of statistics. What conclusions might be drawn from the study using **inferential statistics**?

1. A **large sample** of men, aged 48, was studied for 18 years. For unmarried men, approximately 70% were alive at age 65. For married men, 90% were alive at age 65. *(Source: The Journal of Family Issues)*

2. In a sample of Wall Street analysts, the percentage who **incorrectly forecasted** high-tech earnings in a recent year was **44%.** *(Source: Bloomberg News)*



Solution:

1. Descriptive statistics involves statements such as **"For unmarried men, approximately 70% were alive at age 65"** and **"For married men, 90% were alive at age 65."** Also, the figure represents the descriptive branch of statistics. A possible **inference** drawn from the study is that being **married is associated with a longer life for men**.

2. The part of this study that represents the **descriptive branch** of statistics involves the **statement "the percentage [of Wall Street analysts] who incorrectly forecasted high-tech earnings in a recent year was 44%."** A possible **inference** drawn from the study is that the **stock market is difficult to forecast, even for professionals**.

Parameter vs. Statistic

A **parameter** is a numerical description of a ***population*** characteristic.

A **statistic** is a numerical description of a ***sample*** characteristic.

Distinguishing Between a Parameter and a Statistic

Example: Determine whether the numerical value describes a **population parameter** or a **sample statistic**. Explain your reasoning.

1. A recent survey of approximately **400,000 employers** reported that the average starting salary for marketing majors is **$53,400**. *(Source: National Association of Colleges and Employers)*

2. The freshman class at a university has an average SAT math score of 514.

3. In a random check of 400 retail stores, the Food and Drug Administration found that 34% of the stores were not storing fish at the proper temperature.

Solution

1. Because the average of $53,400 is based on a subset of the population, it is a **sample statistic**.

2. Because the average SAT math score of 514 is based on the entire freshman class, it is a **population parameter.**

3. Because the percent, 34%, is based on a subset of the population, it is a **sample statistic.**

Types of Data

Data sets can consist of **two types** of data: **qualitative data** and **quantitative data.**

- **Qualitative data** consist of attributes, labels, or nonnumerical entries.

- **Quantitative data** consist of numerical measurements or counts.

Classifying Data by Type

Example: The suggested retail prices of **several Honda vehicles** are shown in the table. Which data are **qualitative data** and which are **quantitative data**? Explain your reasoning. *(Source: American Honda Motor Company, Inc.)*

Model	Suggested retail price
Accord Sedan	$21,680
Civic Hybrid	$24,200
Civic Sedan	$18,165
Crosstour	$27,230
CR-V	$22,795
Fit	$15,425
Odyssey	$28,675
Pilot	$29,520
Ridgeline	$29,450

Solution

- The information shown in the table can be separated into two data sets. One data set contains the **names of vehicle models**, and the other contains the **suggested retail prices of vehicle models**.

- The **names are nonnumerical entries**, so these are **qualitative data**.

- The **suggested retail prices** are **numerical entries**, so these are **quantitative data**.

Levels of Measurement

- Another characteristic of **data** is its **level of measurement**. The **level of measurement** determines which **statistical calculations** are **meaningful**.

- The **four levels of measurement**, in order from lowest to highest, are **nominal, ordinal, interval,** and **ratio.**

Nominal vs Ordinal

- Data at the **nominal level of measurement** are **qualitative only.** Data at this level are categorized using **names**, **labels**, or **qualities**. No mathematical computations can be made at this level.

- Data at the **ordinal level of measurement** are **qualitative or quantitative**. Data at this level can be arranged in **order**, or **ranked**, but **differences between data entries are not meaningful**.

Example

Two data sets are shown. Which data set consists of data at the **nominal level**? Which data set consists of data at the **ordinal level**? Explain your reasoning. *(Source: The Numbers)*

Top five grossing movies of 2012
1. Marvel's The Avengers
2. The Dark Knight Rises
3. The Hunger Games
4. Skyfall
5. The Twilight Saga: Breaking Dawn, Part 2

Movie genres
Action
Adventure
Comedy
Drama
Horror

Solution

- The **first data set** lists **the ranks** of five movies. The data set consists of the ranks 1, 2, 3, 4, and 5. Because the **ranks can be listed in order**, these data are at the **ordinal level**. Note that the **difference** between a rank of 1 and 5 has no **mathematical meaning**.

- The **second data set** consists of the **names of movie genres**. No mathematical computations can be made with the **names** and the **names cannot be ranked**, so these data are at the **nominal level**.

Interval vs. Ratio

Data at the **interval level of measurement** can be **ordered**, and **meaningful differences** between data entries can be calculated. At the interval level, a **zero entry** simply represents a **position on a scale**; the entry is not an **inherent zero**.

Data at the **ratio level of measurement** are similar to **data at the interval level**, with the added property that **a zero entry is an inherent zero**. A **ratio of two data entries** can be **formed** so that **one data entry** can be **meaningfully** expressed as a **multiple of another**.

- An ***inherent zero*** is a zero that implies **"none."** For instance, the amount of money you have in a savings account could be **zero dollars**. In this case, the **zero** represents **no money**; it is an **inherent zero**. On the other hand, a **temperature of 0°C** does not represent a condition in which **no heat is present**. The **0°C temperature** is simply a position on the **Celsius scale**; it is **not an inherent zero**.

- To distinguish between data at the **interval level and at the ratio level**, determine whether the expression **"twice as much"** has any meaning in the context of the data.

- For instance, **$2 is twice as much as $1**, so these data are at the **ratio level**. On the other hand, **2°C is not twice as warm as 1°C**, so these data are at the **interval level**.

Classifying Data by Level

Example: Two data sets are shown at below. Which data set consists of data at the interval level? Which data set consists of data at the ratio level? Explain your reasoning.

(Source: Major League Baseball)

New York Yankees' World Series victories (years)
1923, 1927, 1928, 1932, 1936,
1937, 1938, 1939, 1941, 1943,
1947, 1949, 1950, 1951, 1952,
1953, 1956, 1958, 1961, 1962,
1977, 1978, 1996, 1998, 1999,
2000, 2009

2012 American League home run totals (by team)	
Baltimore	214
Boston	165
Chicago	211
Cleveland	136
Detroit	163
Kansas City	131
Los Angeles	187
Minnesota	131
New York	245
Oakland	195
Seattle	149
Tampa Bay	175
Texas	200
Toronto	198

Solution

- Both of these data sets contain **quantitative data**. Consider the dates of the Yankees' World Series victories. It makes sense to find **differences** between **specific dates**. For instance, the time between the **Yankees' first** and last **World Series victories** is **2009 - 1923** = **86 years**. But it does **not make sense** to say that **one year is a multiple of another**. So, these data are at the **interval level**.

- However, using the **home run totals**, you can **find differences *and* write ratios**. From the data, you can see that **Baltimore hit 39 more home runs than Tampa Bay** hit and that **New York** hit about **1.5 times** as many **home runs as Detroit hit**. So, these data are at the **ratio level**.

The tables below summarize which operations are meaningful at each of the four levels of measurement.

When identifying a data set's level of measurement, use the **highest level** that applies.

Level of measurement	Put data in categories	Arrange data in order	Subtract data values	Determine whether one data value is a multiple of another
Nominal	Yes	No	No	No
Ordinal	Yes	Yes	No	No
Interval	Yes	Yes	Yes	No
Ratio	Yes	Yes	Yes	Yes

Summary of Four Levels of Measurement

	Example of a data set	Meaningful calculations
Nominal level (Qualitative data)	*Types of Shows Televised by a Network* Comedy — Documentaries Drama — Cooking Reality Shows — Soap Operas Sports — Talk Shows	*Put in a category.* For instance, a show televised by the network could be put into one of the eight categories shown.
Ordinal level (Qualitative or quantitative data)	*Motion Picture Association of America Ratings Description* G — General Audiences PG — Parental Guidance Suggested PG-13 — Parents Strongly Cautioned R — Restricted NC-17 — No One 17 and Under Admitted	Put in a category and *put in order.* For instance, a PG rating has a stronger restriction than a G rating.
Interval level (Quantitative data)	*Average Monthly Temperatures (in degrees Fahrenheit) for Denver, CO* Jan 30.7 — Jul 74.2 Feb 32.5 — Aug 72.5 Mar 40.4 — Sep 63.4 Apr 47.4 — Oct 50.9 May 57.1 — Nov 38.3 Jun 67.4 — Dec 30.0 *(Source: National Climatic Data Center)*	Put in a category, put in order, and *find differences between values.* For instance, $72.5 - 63.4 = 9.1$°F. So, August is 9.1°F warmer than September.
Ratio level (Quantitative data)	*Average Monthly Precipitation (in inches) for Orlando, FL* Jan 2.35 — Jul 7.27 Feb 2.38 — Aug 7.13 Mar 3.77 — Sep 6.06 Apr 2.68 — Oct 3.31 May 3.45 — Nov 2.17 Jun 7.58 — Dec 2.58 *(Source: National Climatic Data Center)*	Put in a category, put in order, find differences between values, and *find ratios of values.* For instance, $\frac{7.58}{3.77} \approx 2$. So, there is about twice as much precipitation in June as in March.

Key Terms for Data Types

- ***Continuous***
  - Data that can take on any value in an interval.
  - ***Synonyms:* interval, float, numeric**

- ***Discrete***
  - Data that can only take on integer values, such as counts.
  - ***Synonyms:* integer, count**

Key Terms for Data Types

❑*Categorical*

- Data that can only take on a **specific set of values**.
- Example: Sex, type of chocolate, color
- ***Synonyms:*** enums, enumerated, factors, nominal, polychotomous

❑*Binary*

- A special case of categorical with just **two categories (0/1, True, False)**.
- ***Synonyms:*** dichotomous, logical, indicator

❑*Ordinal*

- **Categorical data** that has an **explicit ordering**.
- ***Synonyms:*** ordered factor

Data Types

- ***Binary* data** is an important special case of **categorical data** that takes on only one of two values, such as **0/1, yes/no** or **true/false**.

Synonyms: dichotomous, logical, indicator

- ***Ordinal***
  - **Categorical data** that has an **explicit ordering**.
  - ***Synonyms:*** ordered factor

  An example of this is a numerical rating **(1, 2, 3, 4, or 5)**

Data Types

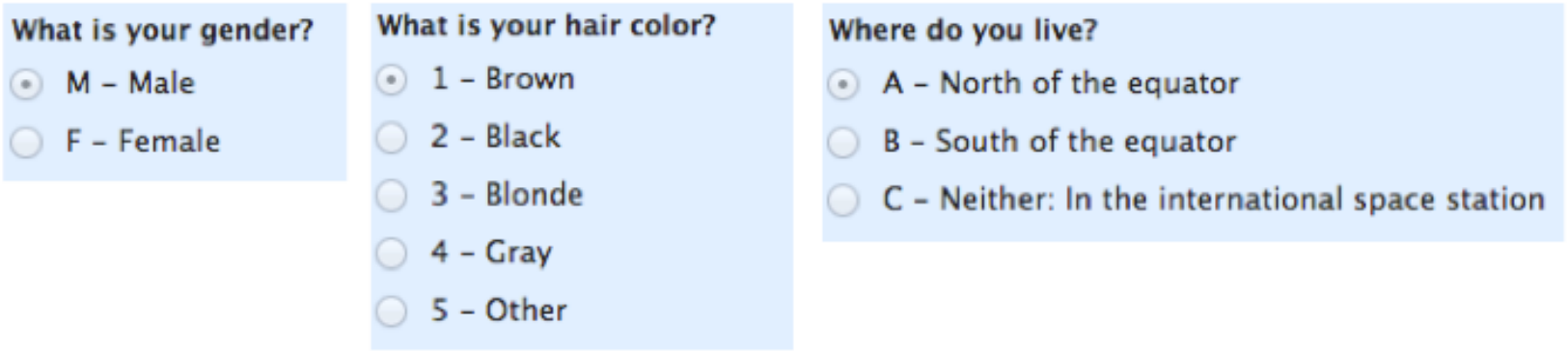
- There are **two basic types** of structured data: **numeric** and **categorical**.

- **Numeric data** comes in two forms: ***continuous***, such as **wind speed** or **time duration**, and ***discrete***, such as the count of the occurrence of an event.

- ***Categorical* data** takes only a **fixed set of values**, such as a type of **TV screen** (plasma, LCD, LED, …) or a **state name** (Alabama, Alaska, …).

Nominal scales

- Nominal scales are used for **labeling variables**, without any quantitative value.
- **"Nominal"** scales could simply be called **"labels."**
- Here are some examples, below. Notice that all of these scales are mutually exclusive (no overlap) and none of them have any **numerical significance**.
- A good way to remember all of this is that **"nominal"** sounds a lot like **"name"** and nominal scales are kind of like **"names"** or **labels**.

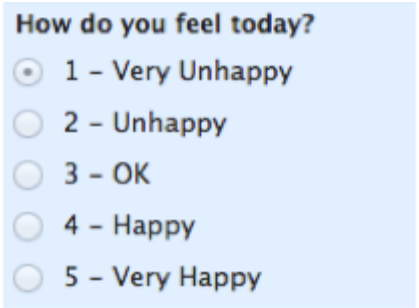


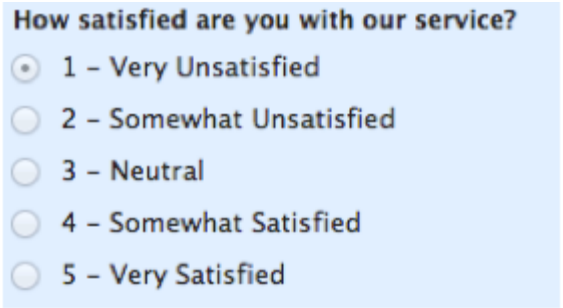
Nominal scale example

- **Type of chocolate**
  - Dark(1)
  - Milk(2)
  - White (3)
- **Sex**
  - Male(0)
  - Female(1)
- **Color**
  - Red(1)
  - Green(2)
  - Blue(3)
  - Yellow(4)

Ordinal scale

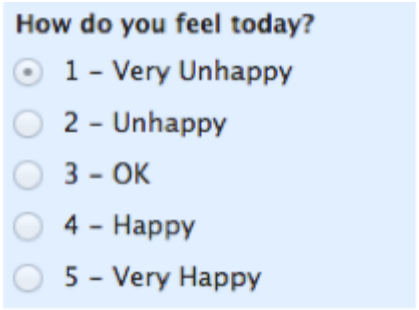
- With **ordinal scales**, it is the **order of the values** is what's **important and significant**, but the differences between each one is not really known.
- Take a look at the example on below. In each case, we know that **option 4** is better than **option 3 or option 2**, but we don't know–and **cannot quantify**–how *much* better it is.
- For example, is the difference between **"OK"** and **"Unhappy"** the same as the difference between **"Very Happy"** and **"Happy" ?** We can't say.
- Ordinal scales are typically **measures** of **non-numeric** concepts like **satisfaction**, **happiness**, **discomfort**, etc.

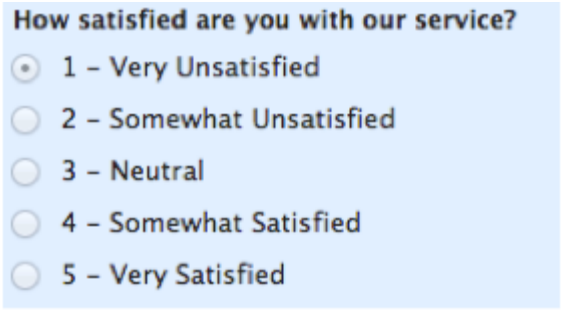




Ordinal scale example

- **"Ordinal"** is easy to remember because is sounds like **"order"** and that's the key to remember with "ordinal scales"–it is the *order* that matters, but that's all you really get from these.
- ***Advanced note*:** The best way to determine *central tendency* on a set of ordinal data is to use the **mode or median**; the mean cannot be defined from an ordinal set.





Key Ideas

- **Data** are typically classified in **software** by their type
- Data types include **continuous, discrete, categorical** (which includes binary), and **ordinal**
- **Data-typing** in software acts as a ***signal*** to the **software** on how to **process the data**