Statistical Data Analysis

Assist. Prof. Dr. Zeyneb KURT

(Slides have been prepared by
Prof. Dr. Nizamettin AYDIN,
updated by Zeyneb KURT)

zeyneb@ce.yildiz.edu.tr

https://www.ce.yildiz.edu.tr/personal/zeyneb

Data types

- Our first requirement is to find a way to represent information (data) in a form that is mutually comprehensible by human and machine.
  - We need to develop schemes for representing all conceivable types of information - language, images, actions, etc.
  - Specifically, the devices that make up a computer are switches that can be on or off, i.e. at high or low voltage.
  - Thus they naturally provide us with two symbols to work with:
    - we can call them on and off, or 0 and 1.

What kinds of data do we need to represent?

Numbers

signed, unsigned, integers, floating point, complex, rational, irrational,…

Text

characters, strings, …

Images

pixels, colors, shapes, …

Sound

Logical

true, false

Instructions

…

Data type:

– representation and operations within the computer

Number Systems – Representation

- Positive radix, positional number systems
- A number with *radix* ***r*** is represented by a string of digits:

$$A_{n-1}A_{n-2} \cdots A_1A_0 \, . \, A_{-1}\, A_{-2} \cdots A_{-m+1}\, A_{-m}$$

in which $\mathbf{0} \le A_i < r$ and . is the *radix point*.

- The string of digits represents the power series:

$$(\text{Number})_r = \left(\sum_{i=0}^{i=n-1} A_i \cdot r^i\right) + \left(\sum_{j=-m}^{j=-1} A_j \cdot r^j\right)$$

(Integer Portion)+ (Fraction Portion)

Decimal Numbers

- "decimal" means that we have ten digits to use in our representation (the symbols 0 through 9)
- What is 3546?
  - it is *three* thousands plus *five* hundreds plus *four* tens plus *six* ones.
  - i.e. $3546 = 3.10^3 + 5.10^2 + 4.10^1 + 6.10^0$
- How about negative numbers?
  - we use two more symbols to distinguish positive and negative:

    + and -

Unsigned Binary Integers

$$Y = \text{“abc”} = a.2^2 + b.2^1 + c.2^0$$

(where the digits a, b, c can each take on the values of 0 or 1 only)

N = number of bits

Range is:
$0 \le i < 2^N - 1$

	3-bits	5-bits	8-bits
0	000	00000	00000000
1	001	00001	00000001
2	010	00010	00000010
3	011	00011	00000011
4	100	00100	00000100

Problem:

- How do we represent *negative* numbers?

Signed Binary Integers

-2s Complement representation-

- Transformation
  - To transform a into -a, invert all bits in a and add 1 to the result

Range is:
$-2^{N-1} < i < 2^{N-1} - 1$

Advantages:

- Operations need not check the sign
- Only one representation for zero
- Efficient use of all the bits

-16	10000
…	…
-3	11101
-2	11110
-1	11111
0	00000
+1	00001
+2	00010
+3	00011
…	…
+15	01111

Limitations of integer representations

- Most numbers are not integer!
  - Even with integers, there are two other considerations:
- Range:
  - The magnitude of the numbers we can represent is determined by how many bits we use:
    - e.g. with 32 bits the largest number we can represent is about +/- 2 billion, far too small for many purposes.
- Precision:
  - The exactness with which we can specify a number:
    - e.g. a 32 bit number gives us 31 bits of precision, or roughly 9 figure precision in decimal repesentation.
- We need another data type!

Real numbers

- Our decimal system handles non-integer *real* numbers by adding yet another symbol - the decimal point (.) to make a *fixed point* notation:
  - e.g. $3456.78 = 3.10^3 + 4.10^2 + 5.10^1 + 6.10^0 + 7.10^{-1} + 8.10^{-2}$
- The scientific notation or *floating point* allows us to represent very large and very small numbers (integer or real), with as much or as little precision as needed:
  - Unit of electric charge e = $1.602\ 176\ 462 \times 10^{-19}$ Coulomb
  - Volume of universe = 1×10^{85} cm^3
    - the two components of these numbers are called the mantissa and the exponent

Real numbers in binary

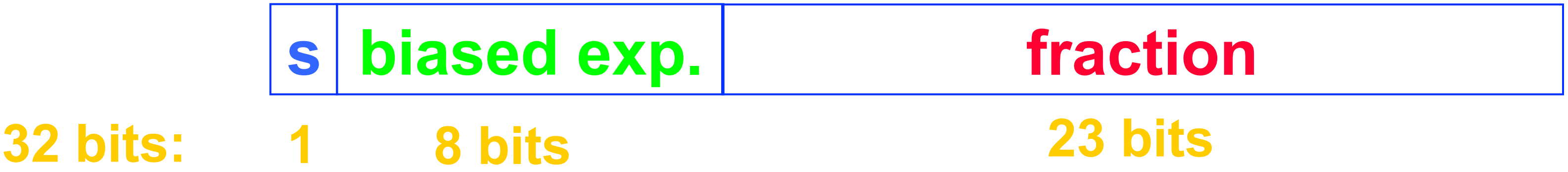
- We mimic the decimal floating point notation to create a "hybrid" binary floating point number:
  - We first use a "binary point" to separate whole numbers from fractional numbers to make a fixed point notation:
    - e.g. 00011001.110 = $1.2^{4} + 1.10^{3} + 1.10^{1} + 1.2^{-1} + 1.2^{-2}$ => 25.75

      ($2^{-1} = 0.5$ and $2^{-2} = 0.25$, etc.)
  - We then "float" the binary point:
    - 00011001.110 => 1.1001110 x 2^{4}

      mantissa = 1.1001110, exponent = 4
  - Now we have to express this without the extra symbols (x, 2, .)
    - by convention, we divide the available bits into three fields:

sign, mantissa, exponent

IEEE-754 fp numbers - 1



$$N = (-1)^{s} \times 1.\text{fraction} \times 2^{(\text{biased exp.} - 127)}$$

- Sign: 1 bit (0 for +ve numbers, 1 for –ve numbers)
- Mantissa: 23 bits
  - We “normalize” the mantissa by dropping the leading 1 and recording only its fractional part (i.e. “1.” will be omitted, “fraction” will be stored)
- Exponent:
  - In order to handle both +ve and -ve exponents, we add 127 to the actual exponent to create a “biased exponent”:
    - 2^{-127} => biased exponent = 0000 0000 (= 0)
    - 2^{0} => biased exponent = 0111 1111 (= 127)
    - 2^{+127} => biased exponent = 1111 1110 (= 254)

IEEE-754 fp numbers - 2

- Example: Find the corresponding fp representation of 25.75
    - 25.75 => 00011001.110 => 1.1001110 x 2^4
    - sign bit = 0 (+ve)
    - normalized mantissa (fraction) = 100 1110 0000 0000 0000 0000
    - biased exponent = 4 + 127 = 131 => 1000 0011
    - so 25.75 => 0 1000 0011 100 1110 0000 0000 0000 0000 => x41CE0000
- Values represented by convention:
    - Infinity (+ and -): exponent = 255 (1111 1111) and fraction = 0
    - NaN (not a number): exponent = 255 and fraction ≠ 0
    - Zero (0): exponent = 0 and fraction = 0
        - note: exponent = 0 => fraction is *de-normalized,* i.e no hidden 1

IEEE-754 fp numbers - 3

- Double precision (64 bit) floating point

	s	biased exp.	fraction
64 bits:	1	11 bits	52 bits

$$N = (-1)^{s} \times 1.\text{fraction} \times 2^{(\text{biased exp.} - 1023)}$$

- Range & Precision:
  - 32 bit:
    - mantissa of 23 bits + 1 => approx. 7 digits decimal
    - $2^{+/-127}$ => approx. $10^{+/-38}$
  - 64 bit:
    - mantissa of 52 bits + 1 => approx. 15 digits decimal
    - $2^{+/-1023}$ => approx. $10^{+/-306}$

Binary Numbers and Binary Coding

- Flexibility of representation
  - Within constraints below, can assign any binary combination (called a code word) to any data as long as data is uniquely encoded.
- Information Types
  - Numeric
    - Must represent range of data needed
    - Very desirable to represent data such that simple, straightforward computation for common arithmetic operations permitted
    - Tight relation to binary numbers
  - Non-numeric
    - Greater flexibility since arithmetic operations not applied.
    - Not tied to binary numbers

Non-numeric Binary Codes

- Given *n* binary digits (called bits), a binary code is a mapping from a set of represented elements to a subset of the 2^n binary numbers.
- Example: A binary code for the seven colors of the rainbow
- Code 100 is not used

Color	Binary Number
Red	**000**
Orange	**001**
Yellow	**010**
Green	**011**
Blue	**101**
Indigo	**110**
Violet	**111**

Number of Bits Required

- Given M elements to be represented by a binary code, the minimum number of bits, *n*, needed, satisfies the following relationships:

  $2^n > M > 2^{(n-1)}$

  $n = \lceil \log_2 M \rceil$ where $\lceil x \rceil$, called the *ceiling function,* is the integer greater than or equal to x.

- Example: How many bits are required to represent decimal digits with a binary code?
  - 4 bits are required ($n = \lceil \log_2 9 \rceil = 4$)

Number of Elements Represented

- Given n digits in radix r, there are r^n distinct elements that can be represented.
- But, you can represent m elements, $m < r^n$
- Examples:
  - You can represent 4 elements in radix $r = 2$ with $n = 2$ digits: (00, 01, 10, 11).
  - You can represent 4 elements in radix $r = 2$ with $n = 4$ digits: (0001, 0010, 0100, 1000).

Binary Coded Decimal (BCD)

- In the 8421 Binary Coded Decimal (BCD) representation each decimal digit is converted to its 4-bit pure binary equivalent
- This code is the simplest, most intuitive binary code for decimal digits and uses the same powers of 2 as a binary number,
  - but only encodes the first ten values from 0 to 9.
    - For example: $(57)_{dec}$ ➔ $(?)_{bcd}$

(5 7) dec

= (0101 0111)bcd

Error-Detection Codes

- Redundancy (e.g. extra information), in the form of extra bits, can be incorporated into binary code words to detect and correct errors.
- A simple form of redundancy is parity, an extra bit appended onto the code word to make the number of 1's odd or even.
  - Parity can detect all single-bit errors and some multiple-bit errors.
- A code word has even parity if the number of 1's in the code word is even.
- A code word has odd parity if the number of 1's in the code word is odd.

4-Bit Parity Code Example

- Fill in the even and odd parity bits:

Even Parity Message _ Parity	Odd Parity Message _ Parity
000 _	000 _
001 _	001 _
010 _	010 _
011 _	011 _
100 _	100 _
101 _	101 _
110 _	110 _
111 _	111 _

- The codeword "1111" has even parity and the codeword "1110" has odd parity. Both can be used to represent 3-bit data.

ASCII Character Codes

- American Standard Code for Information Interchange
- This code is a popular code used to represent information sent as character-based data.
- It uses 7- bits to represent
  - 94 Graphic printing characters
  - 34 Non-printing characters
- Some non-printing characters are used for text format
  - e.g. BS = Backspace, CR = carriage return
- Other non-printing characters are used for record marking and flow control
  - e.g. STX = start text areas, ETX = end text areas.

ASCII Properties

- **ASCII has some interesting properties:**
- Digits 0 to 9 span Hexadecimal values 30_{16} to 39_{16}
- Upper case A-Z span 41_{16} to $5A_{16}$
- Lower case a-z span 61_{16} to $7A_{16}$
  - Lower to upper case translation (and vice versa) occurs by flipping bit 6
- Delete (DEL) is all bits set,
  - a carryover from when punched paper tape was used to store messages

UNICODE

- UNICODE extends ASCII to 65,536 universal characters codes
  - For encoding characters in world languages
  - Available in many modern applications
  - 2 byte (16-bit) code words

Warning: Conversion or Coding?

- Do NOT mix up "conversion of a decimal number to a binary number" with "coding a decimal number with a binary code".
- $13_{10} = 1101_2$
  - This is conversion
- $13 \Leftrightarrow 0001\ 0011_{BCD}$
  - This is coding

Another use for bits: Logic

- Beyond numbers
  - *logical variables* can be *true* or *false*, *on* or *off*, etc., and so are readily represented by the binary system.
  - A logical variable A can take the values *false = 0* or *true = 1* only.
  - The manipulation of logical variables is known as Boolean Algebra, and has its own set of operations
    - which are not to be confused with the arithmetical operations.
  - Some basic operations: NOT, AND, OR, XOR

Basic Logic Operations

- Truth Tables of Basic Operations

NOT

A	A'
0	1
1	0

AND

A	B	A.B
0	0	0
0	1	0
1	0	0
1	1	1

OR

A	B	A+B
0	0	0
0	1	1
1	0	1
1	1	1

- Equivalent Notations
  - not A = A' = $\overline{A}$
  - A and B = A.B = A∧B = A intersection B
  - A or B = A+B = A∨B = A union B

More Logic Operations

XOR		
A	B	$A \oplus B$
0	0	0
0	1	1
1	0	1
1	1	0

XNOR		
A	B	$(A \oplus B)'$
0	0	1
0	1	0
1	0	0
1	1	1

– Exclusive OR (XOR): either A or B is 1, not both

– $A \oplus B = A.B' + A'.B$

Statistical Data Analysis

The role of statistical analysis in science

The role of statistical analysis in science

- This course discusses some statistical methods,
  - which involve applying statistical methods to various problems such as biological, economics, social, health, etc.
- We use empirical evidence to study populations and make informed decisions
- To study a population, we measure a set of characteristics,
  - which are referred to as variables
- The objective of many scientific studies is to learn about the variation of a specific characteristic in the population of interest

The role of statistical analysis in science

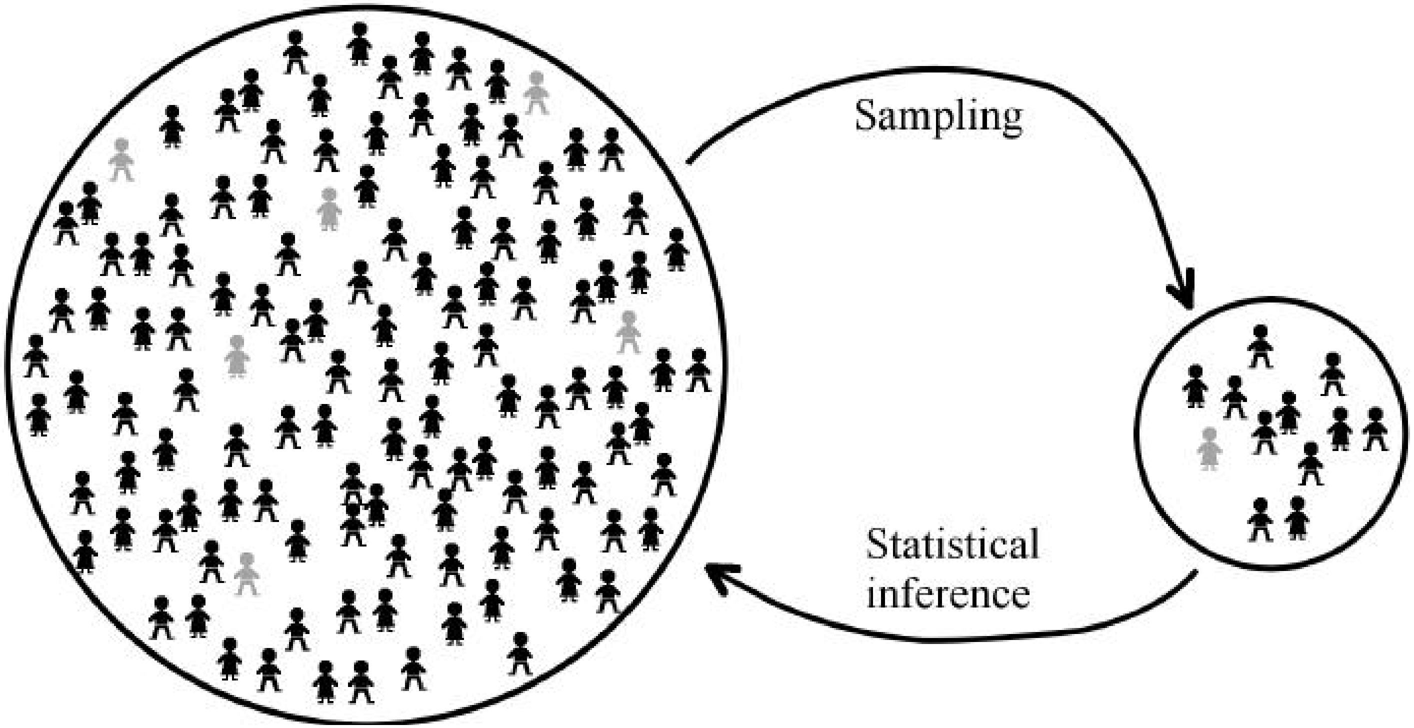
- In many studies, we are interested in possible relationships among different variables.
- We refer to the variables that are the main focus of our study as
  - the response (or target) variables.
- In contrast, we call variables that explain or predict the variation in the response variable as
  - explanatory variables
  - predictors

  depending on the role of these variables.
- Statistical analysis begins with a scientific problem usually presented in the form of
  - a hypothesis testing
  - a prediction problem.

Description of samples and populations

- Statistics is about making statements about a population from data observed from a representative sample of the population.
- A population
  - a collection of subjects whose properties are to be analyzed.
  - contains all subjects of interest.
- A sample
  - a part of the population of interest
  - a subset selected by some means from the population.

Description of samples and populations



- we sample subjects from a large population and use the information obtained from the sample to infer characteristics about the general population.

Description of samples and populations

- A parameter
  - a numerical value that describes a characteristic of a population
- A statistic
  - a numerical measurement that describes a characteristic of a sample
- We use a statistic to infer something about a parameter.

Description of samples and populations

- {For example, we are interested in the average height of a population of individuals.
  - The average height of the population, *m*, is a parameter,
    - but it would be too expensive and/or time-consuming to measure the height of all individuals in the population.
  - Instead we draw a random sample of, say, 12 individuals and measure the height of each of them.
    - The average of those 12 individuals in the sample is our statistic,
      - if the sample is representative of the population and the sample is sufficiently large, we have confidence in using the statistic as an estimate or guess of the true population parameter *m*. }

Description of samples and populations

- The distinction between population and sample depends on the context and the type of inference that you wish to perform.
  - If we were to deduce the average height of the total population, then the 12 individuals are indeed a sample.
  - If for some reason we were only interested in the height of these 12 individuals, and had no intention to make further inferences beyond the 12,
    - then the 12 individuals themselves would constitute the population.

Sampling

- The samples are selected randomly
  - i.e., with some probability from the population.
- Unless stated otherwise, these randomly selected members of populations are assumed to be independent.
- The selected members are called sampling units.
- The individual entities from which we collect information are called observation units, or simply observations.
- Our sample must be representative of the population, and their environments should be comparable to that of the whole population.

Sampling

- Some of the most widely used sampling designs
  - Simple Random Sampling
    - the chance of being selected is the same for any group of *n* members in the population
  - Stratified Sampling
    - The population is first partitioned into subpopulation and sampling is performed separately within each subpopulation
    - a.k.a. strata
  - Cluster Sampling
    - Group observations units into clusters and then sample from these clusters

Designing Studies

- Once a research question is defined, the next step is designing a study in order to answer that question.
- This amounts to figuring out what process you will use to get the data you need.
- After obtaining the sample, the next step is gathering the relevant information from the selected members.
- There are two major types of studies
  - observational studies
  - experiments

Observational studies and experiments

- In observational studies, researchers are passive examiners,
  - trying to have the least impact on the data collection process.
- Observational studies are quite helpful in detecting relationships among characteristics.
- When studying the relationships between characteristics, it is important to distinguish between association and causality.
  - The realationship is casual if one characteristic influences the other one.
- It is usually easier to establish causality by using experiments.
  - In experiments, researchers attempt to control the process as much as possible.
  - An experiment imposes one or more treatments on the participants in such a way that clear comparisons can be made.

Data exploration

- After collecting data, the next step towards statistical inference and decision making is to perform data exploration,
  - which involves visualizing and summarizing the data.
    - The objective of data visualization is to obtain a high level understanding of the sample and their observed (measured) characteristics.
- To make the data more manageable, we need to further reduce the amount of information in some meaningful ways so that we can focus on the key aspects of the data.
  - Summary statistics are used for this purpose.

Data exploration

- Using data exploration techniques, we can learn about the distribution of a variable.
  - The distribution of a variable tells us
    - the possible values it can take,
    - the chance of observing those values,
    - how often we expect to see them in a random sample from the population.
- Through data exploration, we might detect previously unknown patterns and relationships that are worth further investigation.
  - We can also identify possible data issues, such as unexpected or unusual measurements, known as outliers.

Statistical inference

- We collect data on a sample from the population in order to learn about the whole population.
  - {For example, Mackowiak, et al. (1992) measure the normal body temperature for 148 people to learn about the normal body temperature for the entire population.
    - In this case, we say we are estimating the unknown population average.
      - However, the characteristics and relationships in the whole population remain unknown.
    - Therefore, there is always some uncertainty associated with our estimations.}

Statistical inference

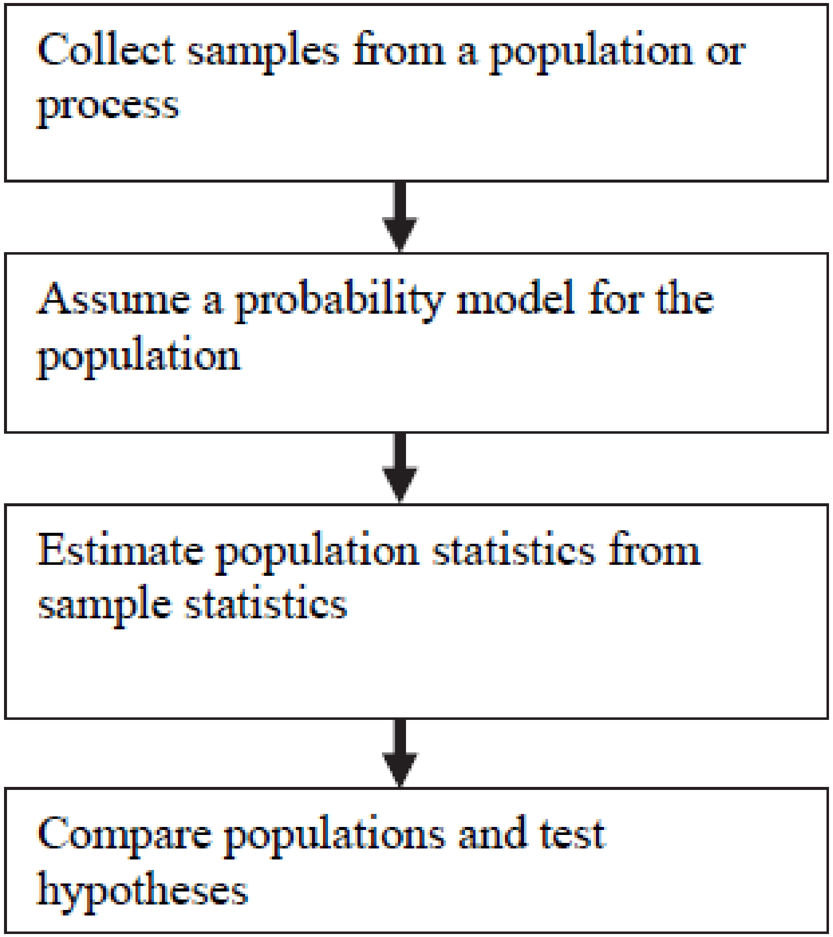
- The mathematical tool to address uncertainty in Statistics
  - probability.
- The process of using the data to draw conclusions about the whole population, while acknowledging the extent of our uncertainty about our findings, is called statistical inference.
- The knowledge we acquire from data through statistical inference allows us to make decisions with respect to the scientific problem that motivated our study and our data analysis.

Computation

- We usually use computer programs to perform most of our statistical analysis and inference.
- The computer programs commonly used for this purpose
  - R,
  - Pyton,
  - SAS,
  - STATA,
  - SPSS,
  - MINITAB,
  - MATLAB,
  - …
- R is free and the most common software among statisticians
- You are encouraged to learn R for additional flexibility in your data analysis.

Summary

- The steps for performing statistical analysis of data.



Collect samples from a population or process

Assume a probability model for the population

Estimate population statistics from sample statistics

Compare populations and test hypotheses

Why statistics?

- Reasons for using statistical data summary and analysis:
  - The real world is full of random events that cannot be described by exact mathematical expressions
  - Variability is a natural and normal characteristic of the natural world
  - We like to make decisions with some confidence.
    - This means that we need to find trends within the variability

Questions to address

- There are several basic questions we hope to address when using numerical and graphical summary of data:
  - Can we differentiate between groups or populations?
    - probably the most frequent aim of biomedical research
  - Are there correlations between variables or populations?
  - Are processes under control?
    - Such a question may arise if there are tight controls on the manufacturing specifications for a medical device

Statistical Data Analysis

Data Types

Data types

- The type(s) of data collected in a study determine
  - the type of statistical analysis that can be used
  - which hypotheses can be tested
  - which model we can use for prediction.
- Broadly speaking, data can be classified into two major types:
  - categorical
  - quantitative

Categorical data

- Categorical data can be grouped into categories based on some qualitative trait.
- The resulting data are merely labels or categories,
  - {examples include
    - gender (male and female)
    - ethnicity (e.g., Caucasian, Asian, African)}
- We can further sub-classify categorical data into two types:
  - nominal
  - ordinal

Categorical data

- **Nominal data**
  - When there is no natural ordering of the categories we call the data nominal.
    - {Hair color is an example of nominal data}
  - Observations are distinguished by name only, and there is no agreed upon ordering.
    - It does not make sense to say “brown” comes before “blonde” or “gray”.
  - Other examples include
    - gender, race, smoking status (smoker or non-smoker), or disease status.

Categorical data

- **Ordinal data**
  - When the categories may be ordered, the data are called ordinal variables.
    - {Categorical variables that judge pain (e.g., none, little, heavy) or income (low-level income, middle-level income, or high-level income) are examples of ordinal variables.}
      - [We know that households with low-level income earn less than households in the middle-level bracket, which in turn earn less than the high-level households.
      - Hence there is an ordering to these categories.]

Categorical data

- It is worth emphasizing that the difference between two categories cannot be measured even though there exists an ordering for ordinal data.
  - {We know that high-income households earn more than low- and medium-income households,}
    - [but not how much more.]
  - {Also we cannot say that the difference between low- and medium-income households is the same as the difference between medium- and high-income households.}

Quantitative data

- Quantitative data are numerical measurements where
  - the numbers are associated with a scale measure rather than just being simple labels.
- Quantitative data fall in two categories:
  - discrete
  - continuous

Quantitative data

- **Discrete quantitative data**
  - numeric data variables that have a finite or countable number of possible values.
    - When data represent counts, they are discrete.
      - {Examples include household size or the number of kittens in a litter.}
    - For discrete quantitative data there is a proper quantitative interpretation of the values:
      - {the difference between a household of size 9 and a household of size 7 is the same as the difference between a household of size 5 and a household of size 3.}

Quantitative data

- **Continuous quantitative data**
  - The real numbers are continuous with no gaps;
    - physically measurable quantities like length, volume, time, mass, etc., are generally considered continuous.
  - However, while the data in theory are continuous, we often have some limitations in the level of detail that is feasible to measure.
    - In some experiments, for example, we measure time in days or weight in kilograms even though a finer resolution could have been used: hours or seconds and grams.
      - In practice, variables are never measured with infinite precision, but regarding a variable as continuous is still a valid assumption.

categorical vs quantitative data

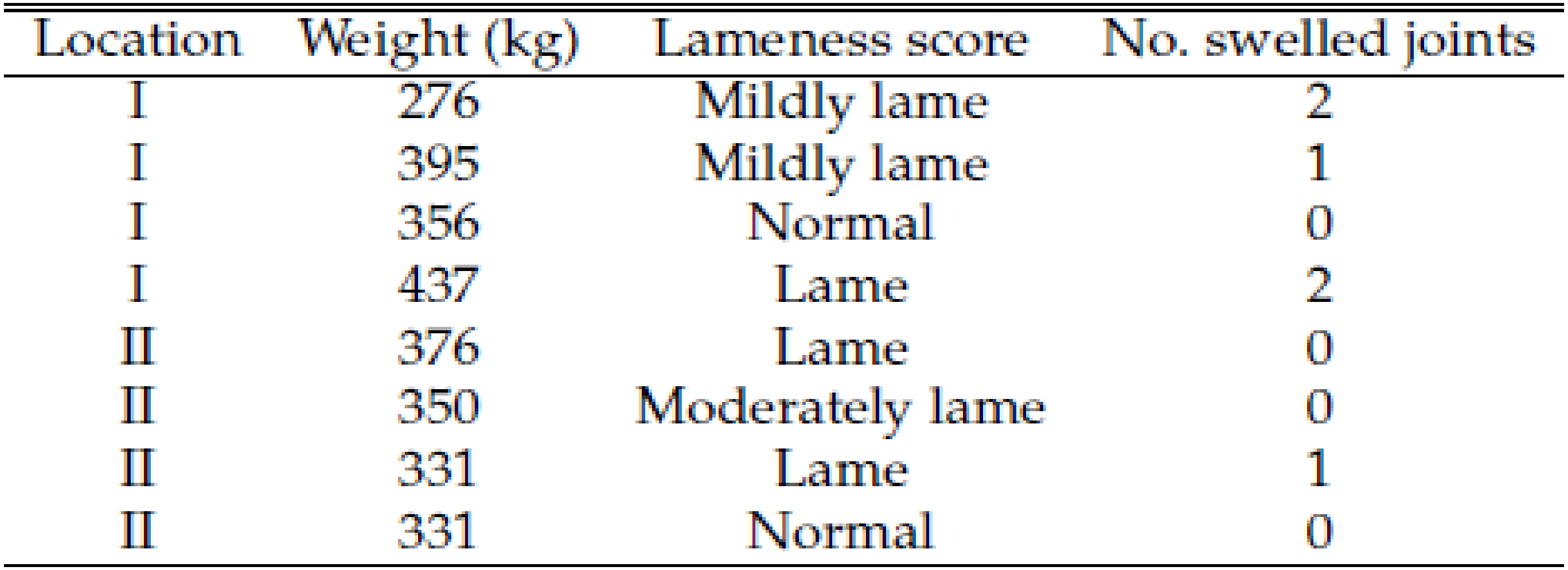
- Categorical data are typically summarized using frequencies or proportions of observations in each category
- Quantitative data typically are summarized using averages or means.

Example (Laminitis in cattle)

- {Danscher et al. (2009) examined 8 heifers in a study to evaluate acute laminitis in cattle after oligofructose overload.
  - Due to logistic reasons, the 8 animals were examined at two different locations.
  - For each of the 8 animals the location, weight, lameness score, and number of swelled joints were recorded 72 hours after oligofructose was administered.
  - The data is shown in the next Table
    - These data contain all four different types of data.}

Example (Laminitis in cattle)

- Data on acute laminitis for eight heifers



Location	Weight (kg)	Lameness score	No. swelled joints
I	276	Mildly lame	2
I	395	Mildly lame	1
I	356	Normal	0
I	437	Lame	2
II	376	Lame	0
II	350	Moderately lame	0
II	331	Lame	1
II	331	Normal	0

- [Laminitis: a disease that affects the feet of ungulates, and is found mostly in horses and cattle]
- [Heifer: a young cow; especially one that has not had a calf]
- [Oligofructose: a form of dietary fiber found in vegetables and other plants, but it's available as a supplement also]

Example (Laminitis in cattle)

- Location is a nominal (categorical) variable as it has a finite set of categories with no specific ordering.
  - Although the location is labeled with Roman numerals, they have no numeric meaning or ordering and might as well be renamed A and B.
- Weight is a quantitative continuous variable even though it is only reported in whole kilograms.
  - The weight measurements are actual measurements on the continuous scale and taking differences between the values is meaningful.
- Lameness score is an ordinal (categorical) variable where the order is defined by the clinicians who investigate the animals:
  - normal, mildly lame, moderately lame, lame, and severely lame.
- The number of swelled joints is a quantitative discrete variable
  - we can count the actual number of swelled joints on each animal.

Describing Data

- Once data are collected, the next step is to summarize it all to get a handle on the big picture.
- Statisticians describe data in two major ways:
  - with pictures
    - that is, charts and graphs
  - with numbers,
    - called descriptive statistics.

Charts and graphs

- Data are summarized in a visual way using charts and/or graphs
  - Some of the basic graphs used include pie charts and bar charts
  - Some data are numerical
  - Data representing counts or measurements need a different type of graph that either keeps track of the numbers themselves or groups them into numerical groupings.
    - One major type of graph that is used to graph numerical data is a histogram.

Descriptive statistics

- Numbers that describe a data set in terms of its important features
  - Categorical data are typically summarized using
    - the number of individuals in each group (the frequency)
    - the percentage of individuals in each group (the relative frequency)
    - Numerical data represent measurements or counts, where the actual numbers have meaning (such as height and weight)

Descriptive statistics

- With numerical data, more features can be summarized besides the number or percentage in each group.
  - Some of these features include
    - measures of center
    - measures of spread
    - measures of the relationship between two variables
- Some descriptive statistics are better than others, and some are more appropriate than others in certain situations

Data have types

- In a conventional programming language, data items are stored in memory locations associated with variables.
- The variables are declared to have values of certain types, and the storage allocated for each variable's value is associated with that type, perhaps something like four bytes for integers, one byte for each character in a string, etc.
- When your program runs, it must have the right instructions to process these bytes because the bytes themselves have no information about what type of data they represent.

Data have types

- For example, a floating point arithmetic instruction attempts to operate on a 4-byte sequence that is not a valid floating point number but was a valid integer.
- If the data are simply mistaken for the wrong type, just processed without notice, very inappropriate results may be obtained, without any idea why the strange results happened.
- Manifestations could be things like garbled text appearing in a display of patient data, a graph having strange anomalies, or colors coming out wrong.

Data have types

- Alternately, the internal binary representation of data in a running program could carry with each piece of data a (binary) tag,
  - identifying its type.
- The running program could then use these tags to determine what to do with the data.
- It makes possible the idea of "run-time dispatch,"
  - ie.,an operator can have many different implementations, one for each type of input it might receive.
- When it executes, it checks the type of the data and chooses the right method for that type of data.

Data have types

- For functions with multiple inputs, there could be methods for all the combinations of different types for each of the inputs.
- Such functions are called generic functions because their source code is organized to perform a generic operation but with different details depending on what kind of data are input.

Data have types

- A system where the data themselves carry type is the basis for key ideas of object-oriented programming.
- The association of types with data is consistent with the idea of a type hierarchy, with general types, subtypes, sub-subtypes, etc.
- The specialized types inherit the properties associated with their parent types.
- A generic function method applicable to a type will also be applicable to all its subtypes.
- The ideas of object-oriented programming depend on having type hierarchies and user-definable types that extend the type hierarchy.