Chapter 1: Introduction to Statistical Data

DSCC 462
Computational Introduction to Statistics

Anson Kahng
Fall 2022

Plan for Today

Plan for Today

- Cover the basics of statistics

Plan for Today

- Cover the basics of statistics
- Introduce types of data

What is Statistics?

What is Statistics?

- **Statistics**: Collection, organization, analysis, and interpretation of data

What is Statistics?

- **Statistics**: Collection, organization, analysis, and interpretation of data
- Descriptive statistics: methods for organizing and summarizing data

What is Statistics?

- **Statistics**: Collection, organization, analysis, and interpretation of data
- Descriptive statistics: methods for organizing and summarizing data
- Statistical inference: methods for inferring properties of a *population* based on a *sample*

Sample vs. Population

Sample vs. Population

- **Sample**: Subset of a group of interest we have data for

Sample vs. Population

- **Sample**: Subset of a group of interest we have data for
- **Population**: The entire group of interest

Sample vs. Population

- **Sample**: Subset of a group of interest we have data for
- **Population**: The entire group of interest
- Using the sample, we can:

Sample vs. Population

- **Sample**: Subset of a group of interest we have data for
- **Population**: The entire group of interest
- Using the sample, we can:
  - Create estimates and plots, perform inference, summarize results

Sample vs. Population

- **Sample**: Subset of a group of interest we have data for
- **Population**: The entire group of interest
- Using the sample, we can:
  - Create estimates and plots, perform inference, summarize results
- **Goal**: Have sample be representative of population so that we can generalize results

Sample vs. Population

- **Sample**: Subset of a group of interest we have data for
- **Population**: The entire group of interest
- Using the sample, we can:
  - Create estimates and plots, perform inference, summarize results
- **Goal**: Have sample be representative of population so that we can generalize results
- Notation:

Sample vs. Population

- **Sample**: Subset of a group of interest we have data for
- **Population**: The entire group of interest
- Using the sample, we can:
  - Create estimates and plots, perform inference, summarize results
- **Goal**: Have sample be representative of population so that we can generalize results
- Notation:
  - Population: parameter (μ, σ^2, etc.)

Sample vs. Population

- **Sample**: Subset of a group of interest we have data for
- **Population**: The entire group of interest
- Using the sample, we can:
  - Create estimates and plots, perform inference, summarize results
- **Goal**: Have sample be representative of population so that we can generalize results
- Notation:
  - Population: parameter (μ, σ^2, etc.)
  - Sample: estimate ($\bar{x}, s^2$, etc.)

Sample vs. Population: Example

Sample vs. Population: Example

- A study was conducted on 423 male children between the age of 5 and 15. From this study, it was concluded that advanced maternal age is associated with higher risk of a male child having autism

Sample vs. Population: Example

ALL

- A study was conducted on 423 male children between the age of 5 and 15. From this study, it was concluded that advanced maternal age is associated with higher risk of a male child having autism

- Population:

Sample vs. Population: Example

- A study was conducted on 423 male children between the age of 5 and 15. From this study, it was concluded that advanced maternal age is associated with higher risk of a male child having autism
- Population: male children aged 5 - 15
- Sample: 423 children in the study

Data

Data

- Data are pieces of information about **subjects** that are organized into **variables**

Data

- Data are pieces of information about **subjects** that are organized into **variables**
- **Subjects** are the particular people or objects we are interested in studying (i.e., the people in our sample)

Data

- Data are pieces of information about **subjects** that are organized into **variables**
- **Subjects** are the particular people or objects we are interested in studying (i.e., the people in our sample)
- **Variables** are the characteristics we are interested in measuring for each subject (e.g., weight, height, eye color)

Data

- Data are pieces of information about **subjects** that are organized into **variables**
- **Subjects** are the particular people or objects we are interested in studying (i.e., the people in our sample)
- **Variables** are the characteristics we are interested in measuring for each subject (e.g., weight, height, eye color)
- Different types of summaries and analyses are appropriate for different types of data

Example Dataset: Emergency Room Patients

	Age	Weight (kg)	Eye Color	Smoker?	Pain Rating
Patient 1	25	82	Brown	Yes	Medium
Patient 2	42	60	Green	No	High
Patient 3	31	105	Blue	No	Low

Example Dataset: Emergency Room Patients

	Age	Weight (kg)	Eye Color	Smoker?	Pain Rating
Patient 1	25	82	Brown	Yes	Medium
Patient 2	42	60	Green	No	High
Patient 3	31	105	Blue	No	Low

- Subjects: Patient 1, Patient 2, Patient 3

Example Dataset: Emergency Room Patients

	Age	Weight (kg)	Eye Color	Smoker?	Pain Rating
Patient 1	25	82	Brown	Yes	Medium
Patient 2	42	60	Green	No	High
Patient 3	31	105	Blue	No	Low

- Subjects: Patient 1, Patient 2, Patient 3
- Variables: Age, Weight, Eye Color, Smoker, Pain Rating

Types of Data

Types of Data

- **Categorical (qualitative) data**: Data that are measured on a scale consisting of sets of groups or categories

Types of Data

- **Categorical (qualitative) data**: Data that are measured on a scale consisting of sets of groups or categories
  - Place subjects in one of the categories

Types of Data

- **Categorical (qualitative) data**: Data that are measured on a scale consisting of sets of groups or categories
  - Place subjects in one of the categories
  - Usually, care about the count / proportion in each category

Types of Data

- **Categorical (qualitative) data**: Data that are measured on a scale consisting of sets of groups or categories
  - Place subjects in one of the categories
  - Usually, care about the count / proportion in each category
- Examples:

Types of Data

- **Categorical (qualitative) data**: Data that are measured on a scale consisting of sets of groups or categories
  - Place subjects in one of the categories
  - Usually, care about the count / proportion in each category
- Examples:
  - Nominal variables, ordinal variables, discrete interval variables with few values, continuous variables that have been grouped into a small number of categories

Types of Data

Types of Data

- **Numerical (quantitative) data**: Data that are counts or measured on a numeric scale

Types of Data

- **Numerical (quantitative) data**: Data that are counts or measured on a numeric scale
  - Discrete if measurements are integers

Types of Data

- **Numerical (quantitative) data**: Data that are counts or measured on a numeric scale
  - Discrete if measurements are integers
  - Continuous if measurements can take any value within a range

Data Type Example

Data Type Example

- **Categorical**:

Data Type Example

- **Categorical**:
  - Smoking status (yes/no)

Data Type Example

- **Categorical**:
  - Smoking status (yes/no)
  - Class year (first-year/sophomore/junior/senior/graduate student)

Data Type Example

- **Categorical**:
  - Smoking status (yes/no)
  - Class year (first-year/sophomore/junior/senior/graduate student)
  - Hair color (black, brown, blonde, red, other)

Data Type Example

- **Categorical**:
  - Smoking status (yes/no)
  - Class year (first-year/sophomore/junior/senior/graduate student)
  - Hair color (black, brown, blonde, red, other)
- **Quantitative**:

Data Type Example

- **Categorical**:
  - Smoking status (yes/no)
  - Class year (first-year/sophomore/junior/senior/graduate student)
  - Hair color (black, brown, blonde, red, other)
- **Quantitative**:
  - Age

Data Type Example

- **Categorical**:
  - Smoking status (yes/no)
  - Class year (first-year/sophomore/junior/senior/graduate student)
  - Hair color (black, brown, blonde, red, other)
- **Quantitative**:
  - Age
  - Steps per day

Data Type Example

- **Categorical**:
  - Smoking status (yes/no)
  - Class year (first-year/sophomore/junior/senior/graduate student)
  - Hair color (black, brown, blonde, red, other)
- **Quantitative**:
  - Age
  - Steps per day
  - GPA

Data Type Example

- **Categorical**:
  - Smoking status (yes/no)
  - Class year (first-year/sophomore/junior/senior/graduate student)
  - Hair color (black, brown, blonde, red, other)
- **Quantitative**:
  - Age
  - Steps per day
  - GPA
  - Number of siblings

Categorical Data

Categorical Data

- **Nominal Data**:

Categorical Data

- **Nominal Data**:
  - *Unordered* categories or classes

Categorical Data

- **Nominal Data**:
  - *Unordered* categories or classes
  - Order of the categories is irrelevant

Categorical Data

- **Nominal Data**:
  - *Unordered* categories or classes
  - Order of the categories is irrelevant
- Examples:

Categorical Data

- **Nominal Data**:
  - *Unordered* categories or classes
  - Order of the categories is irrelevant
- Examples:
  - Department: philosophy, data science, statistics, linguistics, art history

Categorical Data

- **Nominal Data**:
  - *Unordered* categories or classes
  - Order of the categories is irrelevant
- Examples:
  - Department: philosophy, data science, statistics, linguistics, art history
  - Hair color: black, brown, blonde, red, other

Categorical Data

Categorical Data

- **Ordinal Data**:

Categorical Data

- **Ordinal Data**:
  - *Ordered* categories or classes (“natural ordering”)

Categorical Data

- **Ordinal Data**:
  - *Ordered* categories or classes (“natural ordering”)
  - Distances between categories are unknown

Categorical Data

- **Ordinal Data**:
  - *Ordered* categories or classes (“natural ordering”)
  - Distances between categories are unknown
  - Care about the ordering itself, not the magnitude

Categorical Data

- **Ordinal Data**:
  - *Ordered* categories or classes (“natural ordering”)
  - Distances between categories are unknown
  - Care about the ordering itself, not the magnitude
- Examples:

Categorical Data

- **Ordinal Data**:
  - *Ordered* categories or classes (“natural ordering”)
  - Distances between categories are unknown
  - Care about the ordering itself, not the magnitude
- Examples:
  - Pain scale: low, medium, high

Categorical Data

strong disagree

strong agree

-3, -2, ..., 2, 3

- **Ordinal Data**:
  - *Ordered* categories or classes ("natural ordering")
  - Distances between categories are unknown
  - Care about the ordering itself, not the magnitude
- Examples:
  - Pain scale: low, medium, high
  - Course evaluations: unsatisfactory, neutral, satisfactory, excellent

Categorical Data

Categorical Data

- **Ranked Data**:

Categorical Data

- **Ranked Data**:
  - Arrange a group of observations from highest to lowest (or reversed) according to their magnitude

Categorical Data

- **Ranked Data**:
  - Arrange a group of observations from highest to lowest (or reversed) according to their magnitude
  - Assign ranks corresponding to each observation's place in the sequence

Categorical Data

- **Ranked Data**:
  - Arrange a group of observations from highest to lowest (or reversed) according to their magnitude
  - Assign ranks corresponding to each observation's place in the sequence
- Example:

Categorical Data

100, 3, 2, 1

1, 2, 3, 4

- **Ranked Data**:
  - Arrange a group of observations from highest to lowest (or reversed) according to their magnitude
  - Assign ranks corresponding to each observation's place in the sequence
- Example:
  - GPAs: (93.1, 86.2, 98.5, 89.8) → (2, 4, 1, 3)

2 4 1 3

Quantitative Data

Quantitative Data

- **Discrete Data**:

Quantitative Data

- **Discrete Data**:
  - Both order and magnitude are important

Quantitative Data

- **Discrete Data**:
  - Both order and magnitude are important
  - Numbers represent actual values instead of labels

Quantitative Data

- **Discrete Data**:
  - Both order and magnitude are important
  - Numbers represent actual values instead of labels
  - Often integers or counts (isolated points on a number line)

Quantitative Data

- **Discrete Data**:
  - Both order and magnitude are important
  - Numbers represent actual values instead of labels
  - Often integers or counts (isolated points on a number line)
- Examples:

Quantitative Data

- **Discrete Data**:
  - Both order and magnitude are important
  - Numbers represent actual values instead of labels
  - Often integers or counts (isolated points on a number line)
- Examples:
  - Number of steps walked in a day

Quantitative Data

- **Discrete Data**:
  - Both order and magnitude are important
  - Numbers represent actual values instead of labels
  - Often integers or counts (isolated points on a number line)
- Examples:
  - Number of steps walked in a day
  - Number of prospective students who come to Admissions this week

Quantitative Data

Quantitative Data

- **Continuous Data**:

Quantitative Data

- **Continuous Data**:
  - Data can take any value within a given interval (entire interval on a number line)

Quantitative Data

- **Continuous Data**:
  - Data can take any value within a given interval (entire interval on a number line)
  - Distance between measurements is meaningful (both order and magnitude matter)

Quantitative Data

- **Continuous Data**:
  - Data can take any value within a given interval (entire interval on a number line)
  - Distance between measurements is meaningful (both order and magnitude matter)
- Examples:

Quantitative Data

- **Continuous Data**:
  - Data can take any value within a given interval (entire interval on a number line)
  - Distance between measurements is meaningful (both order and magnitude matter)
- Examples:
  - Height in cm

Quantitative Data

- **Continuous Data**:
  - Data can take any value within a given interval (entire interval on a number line)
  - Distance between measurements is meaningful (both order and magnitude matter)
- Examples:
  - Height in cm
  - Time in minutes spent on an assignment

Quantitative Data

Quantitative Data

- **Interval** level of measurement:

Quantitative Data

- **Interval** level of measurement:
  - Data can be arranged in some order, and the difference between any two data values is meaningful

Quantitative Data

- **Interval** level of measurement:
  - Data can be arranged in some order, and the difference between any two data values is meaningful
  - There is no natural zero or starting point

Quantitative Data

- **Interval** level of measurement:
  - Data can be arranged in some order, and the difference between any two data values is meaningful
  - There is no natural zero or starting point
  - Ex: Years (1000, 1359, 2009, 2017)

Quantitative Data

- **Interval** level of measurement:
  - Data can be arranged in some order, and the difference between any two data values is meaningful
  - There is no natural zero or starting point
  - Ex: Years (1000, 1359, 2009, 2017)
- **Ratio** level of measurement:

Quantitative Data

- **Interval** level of measurement:
  - Data can be arranged in some order, and the difference between any two data values is meaningful
  - There is no natural zero or starting point
  - Ex: Years (1000, 1359, 2009, 2017)
- **Ratio** level of measurement:
  - Data can be arranged in some order, and the difference or ratio between any two data values is meaningful

Quantitative Data

- **Interval** level of measurement:
  - Data can be arranged in some order, and the difference between any two data values is meaningful
  - There is no natural zero or starting point
  - Ex: Years (1000, 1359, 2009, 2017)
- **Ratio** level of measurement:
  - Data can be arranged in some order, and the difference or ratio between any two data values is meaningful
  - There is a natural zero or starting point

Quantitative Data

- **Interval** level of measurement:
  - Data can be arranged in some order, and the difference between any two data values is meaningful
  - There is no natural zero or starting point
  - Ex: Years (1000, 1359, 2009, 2017)
- **Ratio** level of measurement:
  - Data can be arranged in some order, and the difference or ratio between any two data values is meaningful
  - There is a natural zero or starting point
  - Ex: Price of textbooks ($0 represents no cost; $100 costs twice as much as $50)

Four Levels of Measurement

disc or cont

	Nominal	Ordinal	Interval	Ratio
Categorizes and labels variables	✓	✓	✓	✓
Ranks categories in order		✓	✓	✓
Has known, equal intervals			✓	✓
Has a true or meaningful zero				✓

	Disc	Cont
Int	•	•
Rat	•	•

Categorical **Quantitative**