What is Statistics?

Brenda Gunderson

What is Statistics?

- **Methodological** subject encompassing all aspects of **learning from data**.

  tools and methods for working with and understanding data

- **Statisticians** apply and develop data analysis methods, seek to understand their properties…

  …when do these tools provide *insight?*
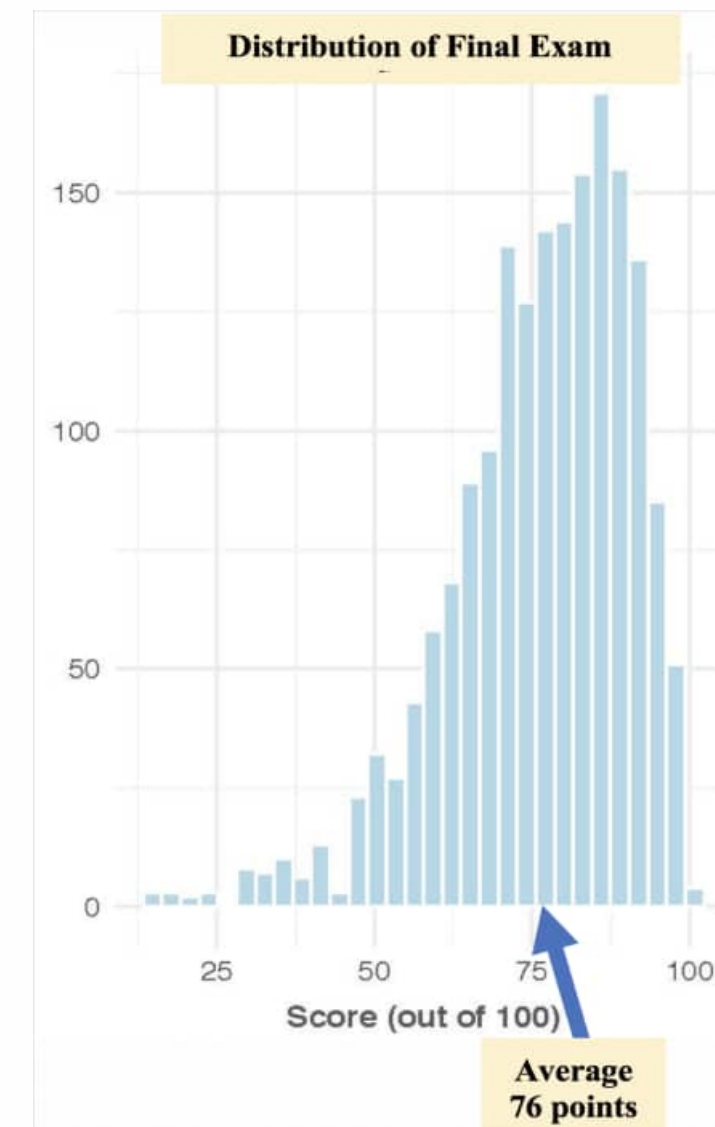 …when are they *possibly misleading?*

- **Researchers** and **workers** apply and extend statistical methodology, and contribute new ideas and methods for conducting data analysis.

A "Statistic" and the field of "Statistics"

- A **statistic** ~ numerical or graphical summary of a collection of data.
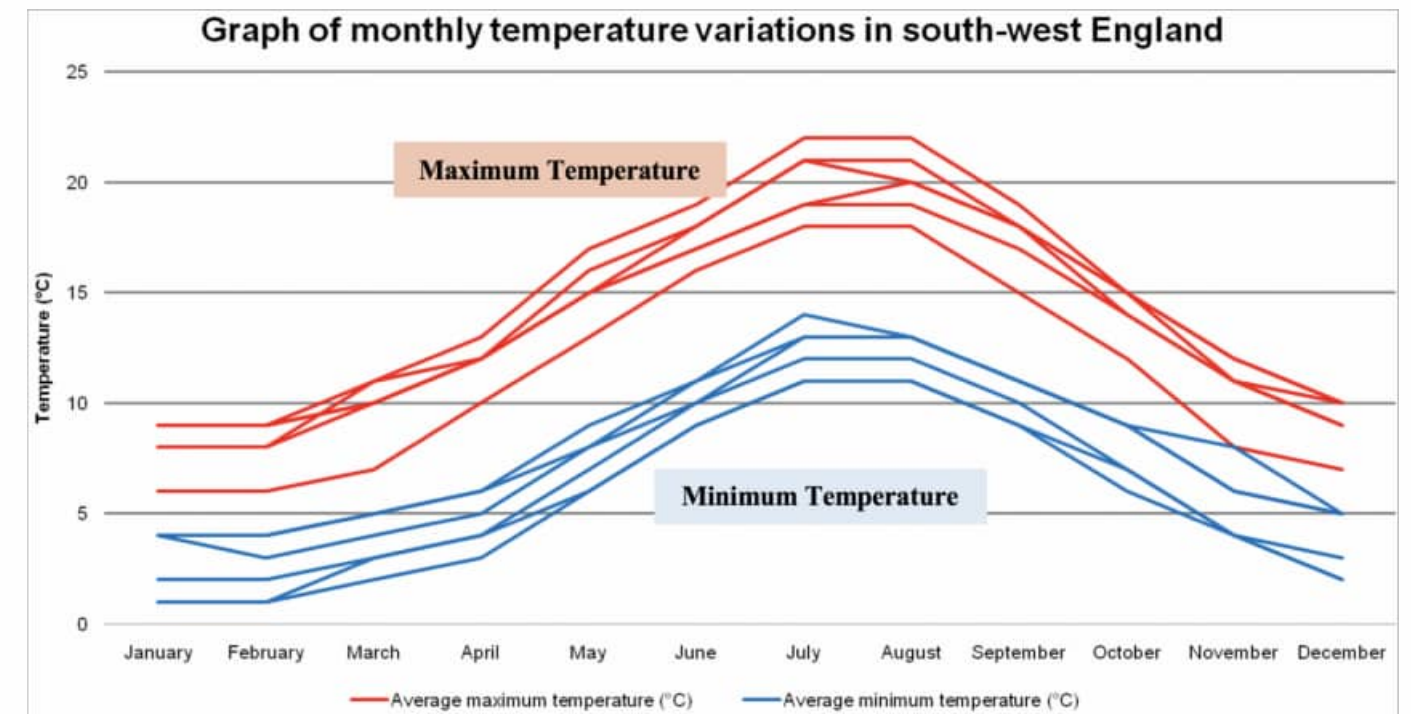
A "Statistic" and the field of "Statistics"

- A **statistic** ~ numerical or graphical summary of a collection of data.

  - Average score on final exam

Distribution of Final Exam

Score (out of 100)

Average
76 points

A "Statistic" and the field of "Statistics"

- A **statistic** ~ numerical or graphical summary of a collection of data.

  ○ Average score on final exam

  ○ Minimum temperature at a location over year

Graph of monthly temperature variations in south-west England

A "Statistic" and the field of "Statistics"

- A **statistic** ~ numerical or graphical summary of a collection of data.

  ○ Average score on final exam

  ○ Minimum temperature at a location over year

  ○ Proportion of people who are retired

A "Statistic" and the field of "Statistics"

- A **statistic** ~ numerical or graphical summary of a collection of data.

  ○ Average score on final exam

  ○ Minimum temperature at a location over year

  ○ Proportion of people who are retired

- **Statistics** ~ academic discipline focusing on research methodology. Statisticians develop new statistical tools, calculate statistics from data, and collaborate with subject-matter experts to interpret them.

The Landscape of Statistics

Evolving and dynamic field ~ Emerging **challenges** and **opportunities**

- **Properties** of statistical methods are under **continuing study**

- New application areas → **development** of new analytic methods

- New types of sensors → **new types of data**

- Advances in **computing** → sophisticated analyses on Big Data

Perspectives on Statistical Science

Statistics is a "**big tent**" discipline ~ incorporates new ideas from theory, practice, allied fields.
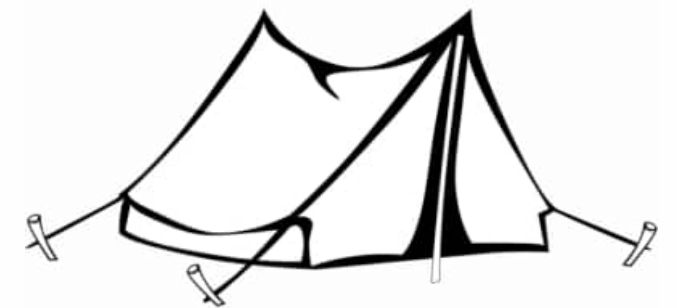
Different Perspectives:

- ○ "art of summarizing data"
- ○ "science of uncertainty"
- ○ "science of decisions"
- ○ "science of variation"
- ○ "art of forecasting"
- ○ "science of measurement"
- ○ "basis for principled data collection"

Statistics as the "art of summarizing data"

- Data can be **overwhelming**

- Making sense of data usually involves **reduction** and **summarization**

make a dataset **comprehensible** to human observer

always **depends primarily on goals of "data consumer"** to be meaningful -- many approaches

Statistics as the "science of uncertainty"

- Data can be **misleading**

- Statistics provides framework for **assessing whether claims based on data are meaningful**

- Uncertainty is inevitable, but it is highly desirable to **quantify how far our reported findings may fall from "the truth"**

Many public opinion polls report   **± margin of error**

→ potential discrepancy between
reported and actual states of public opinion

Statistics as the "science of decisions"

- Understanding data is important →
  only consequential if we act on what we have learned

- **Decision-making =** ultimate goal of any statistical analysis

- We **make decisions in face of uncertainty**!
  What are costs and benefits of different approaches?

→ at higher than average risk for cancer…
should they undergo preventative procedure?

Statistics as the "science of variation"

- Often focus on most typical or "**central**" value
- Great emphasis on understanding **variation** in data!

Average American has around $6000 of credit card debt → central value of credit card debt in US population.

10% of Americans have more than $30,000 in credit card debt → variation of credit card debt in US population.

Statistics as the "art of forecasting"

- Forecasting or prediction = central tasks in statistics

- **Cannot** know future with absolute certainty, but efficient use of available data

  → **can** sometimes make accurate predictions about future
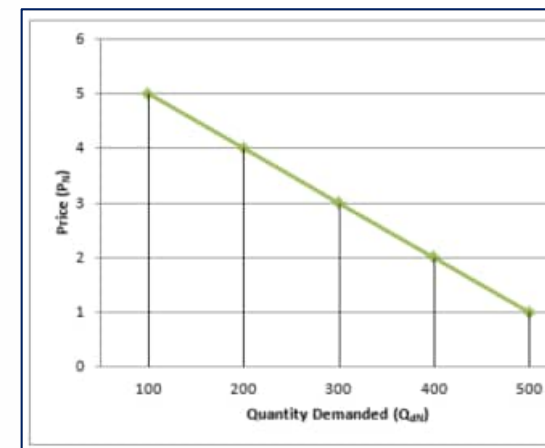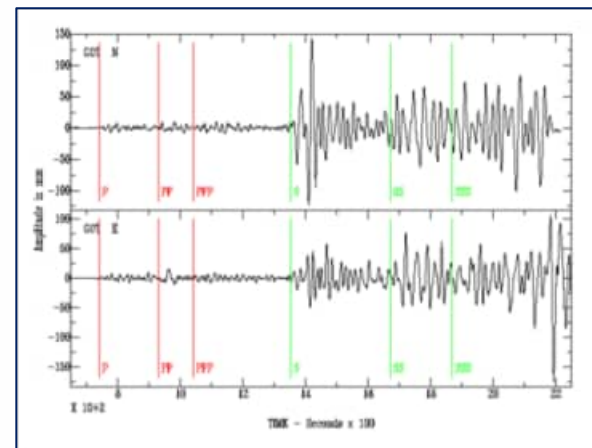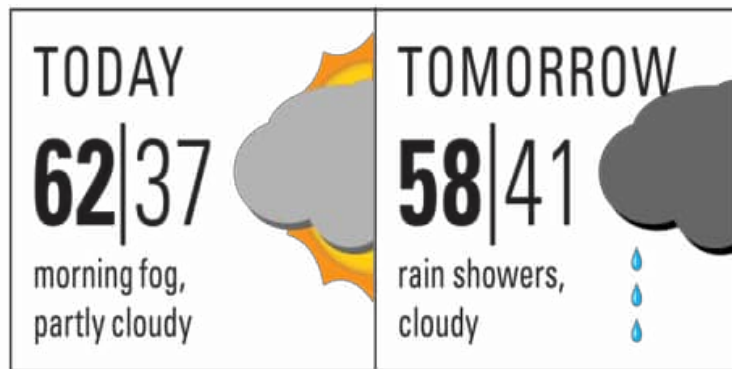
Statistics as the "science of measurement"

- **High accuracy**: person's age or height

- **More difficult**: blood pressure (*varies minute to minute*)

- **Harder**: "mood", "political ideology", "personality"

Statistics: major role in **constructing and evaluating rigorous approaches for measuring difficult-to-define concepts** and in **assessing quality**.

Statistics as the "basis for principled data collection"

- Data often expensive and difficult to collect

- Resource limitations → collect least data possible

Statistics: provides a rational way to manage this trade-off

History of Statistics Milestones

Ancient Times:

Data Collection on harvests floods population sizes

1700's:

Probability Theory
→ randomness and variation

19th Century:

Modern Statistics emerges, via genetics demography economics

20th Century:

Statistical Theory advances, new application areas, **computers**

21st Century:

"massive data", "data science" "machine learning"

Statistics and its Allied Fields

Computer science: algorithms, data structures for working with data, programming languages for manipulating data.

Mathematics: language and notation for expressing statistical concepts concisely, tools for understanding properties of statistical methods.

Probability theory: branch of mathematics ~ crucial part of foundations of statistics – to express ideas about randomness and uncertainty.

Data Science: database management, machine learning, computational infrastructure to carry out data analysis.

Frontiers of Statistics

Statistics $\Rightarrow$ **Statistics** $\Rightarrow$ **Statistics**

Emerging applications

- Computer vision
- Recommender systems
- Predictive analytics
- Fraud and anomaly detection
- Risk assessment
- Social and government services