IEDA 2540 Statistics for Engineers
Introduction

Wei YOU

香港科技大學
THE HONG KONG UNIVERSITY OF
SCIENCE AND TECHNOLOGY

Spring, 2025

What is Statistics?

The thinking involved in statistics is not new!

- Imagine meeting two friends: one is 1.6m tall, the other 1.8m tall.
- Based on your daily experience, you'd likely guess the taller one is a man.
- Could you be wrong? Sure!
- Yet, you wouldn't hesitate to trust your gut.

Statistical Way of Thinking

- Statistics **summarize** past experience.
- When facing a new scenario, we **generalize/predict** from previous experience by making an informed guess.

Example:

- **Summarize**: "On *average*, I cycle about 100 miles a week."
- **Predict**: "We can *expect* a lot of rain at this time of year."
- **Predict**: "The earlier you start preparing, the *more likely* you are to do well in exams."

Statistical Thinking in Everyday Life

Humans learn from the environment to gain greater control.

- However, we can never know everything for certainty.
- Fortunately, we *observe* and notice *patterns* and *regularities*.

Example: Counting and measuring in a *rough-and-ready* fashion.

- Humans have the intuition to rely on habits of "quantification": "How often?", "how big?", "how difficult?", "how far?".

Summarizing one aspect of one subject.

Statistical Thinking in Everyday Life

Example: Noticing several aspects of the same subject

- the size of a potato-crop in a particular field this year
- Average temperature of the growing season.
- Sunshine.
- Amount of fertilizer used.
- Nature of soil (acidity or basicity).
- Rainfall.

Example: More interestingly, we compare several subjects with lots in common but differing in some respects.

- How the size of a potato varies between several different fields in the same year?
- Summarizing multiple aspects of multiple subjects $\rightarrow$ **data collection**.

Looking for Patterns

By instinct, when facing a collection of observations, we look for **connections**, **patterns**, **similarities** and **differences**.

Example: Studying the size of potato-crop harvested from different fields

- Are there differences in their sizes?
  - If one potato from field A is larger than one potato from field B, can we say field A is better?
  - If the average of 5 potatoes from field A is larger than that from field B, is field A better?
  - How confident are we in such claims?
- What are the differences between the fields? What could be the cause of the size difference?
  - Soil? Weather? Planting methods? Or a combination of several factors?

From Patterns to Action

More vital questions

- What can we learn from the patterns we identified?
- How do they help us act more efficiently in the future?

This is where statistics comes in.

- **Statistics** is a scientific way to:
  - **collect** and **summarize** observations,
  - **identify** patterns,
  - **generalize** conclusions,

  and avoid jumping to **hasty conclusions** by considering what could have happen or what could have gone wrong.

To which extent can we generalize from our limited experiences?

Be Cautious about Generalizing from Experience

Tendency to generalize is human nature.

Example: When we generously apply fertilizer, we harvest bigger than usual potato.

- Shall we apply more fertilizer?
- Shall we do the same for other plants?
- Shall we do the same for other fields?

Would you think it safe to generalize in this way – on the basis of experience with one field? Why, or why not?

- What is true for one subject may not be true for others!
- To generalize more confidently, we need **more experience/data**!

Dealing with Chances

Even if we have collected as much data as we can, is it possible to predict with 100% certainty?

Think about the following two arguments. Which do you think is more likely to be correct?

- "I predict *most* fields will have bigger potatoes if we do such-and-such."
- "I predict that *any specific* field will have bigger potatoes if we do such-and-such."

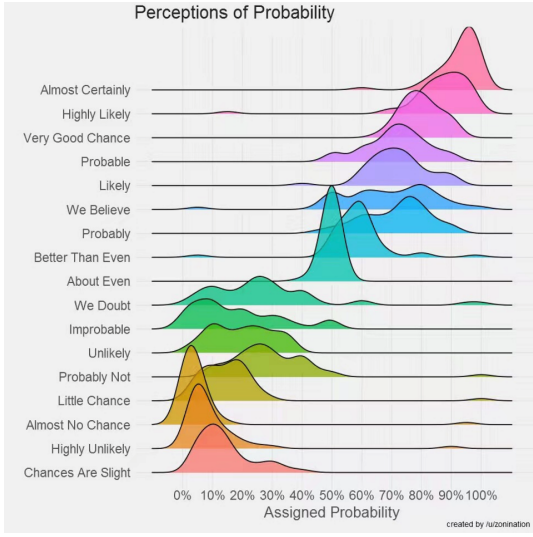
There is no 100% certainty, especially when dealing with people, things and events!

- Statistics helps us to look for **reliable regularities** and associations among things "in general," "in the long run" and "with high chance."
- At the same time, it teaches us to **manage expectation** about our predictions, especially when it comes to specific individuals.

What is Probability?

Probability is the language of uncertainty and variability.

- Probability used in daily communication: “Almost certainly”, “probably not”, ...
- What are the meaning of these phrases?



What is Probability?

Example: Probability used in daily communication

- When some one says "highly likely," what does it mean?
- With a probability approach, we need to assume a "distribution," i.e., how likely a certain meaning of "highly likely" occurs.
  Example: Imagine a colleague is always late for meetings. Someone says, "He'll be on time for once, *highly likely*," sarcastically.
  - What is the meaning of "highly likely" out of an average person's mouth? Mean.
  - How variable is the meaning of "highly likely?" Variance.

History of Probability

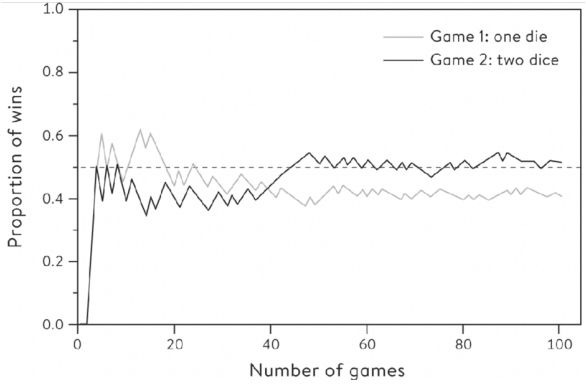
A 17th century gambler Chevalier de Méré

- **Game 1:** "Throw a fair die at most 4 times and win if you get a six."
- **Game 2:** "Throw two fair dice at most 24 times and win if you get a double-six."

Which one would you bet on?

Chevalier de Méré's Reasoning

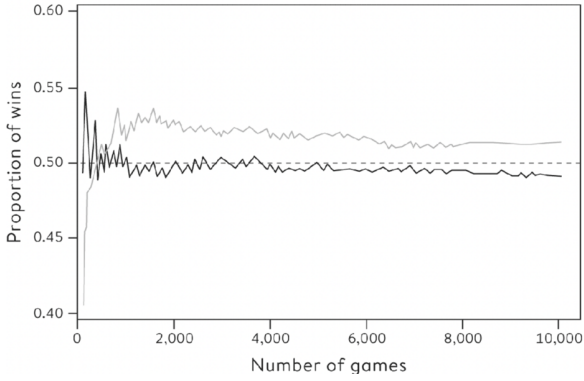
He simulated the game 100 times, and found that Game 2 has higher reward.



Chevalier de Méré's Reasoning

What went wrong?

- In the presence of randomness, we need to wisely interpret the experiments we did.



The Math Behind

Let $p = \frac{1}{6}$ be the probability of a six in one roll.

- **Game 1:** "Throw a fair die at most 4 times and win if you get a six."
  - Probability of winning:
$$1 - (1-p)^4 = 0.518.$$
- **Game 2:** "Throw two fair dice at most 24 times and win if you get a double-six."
  - Probability of winning:
$$1 - \left(1 - \frac{p}{6}\right)^{4\times 6} = 1 - \left[\left(1 - \frac{p}{6}\right)^6\right]^4 = 0.491.$$

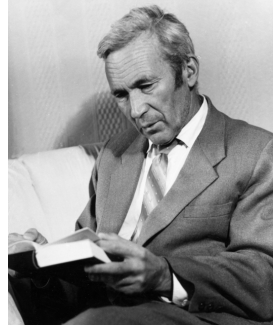
Read more here: `https://tinyurl.com/3t3y9xku`

History of Probability: Key Figures

- Chevalier de Méré complained to **Pascal** and **Fermat**, who laid the foundation of probability theory.
- Probability theory developed significantly through the 17th to 19th centuries.
- **Andrey Kolmogorov** is considered the founder of modern probability theory.

Blaise Pascal

Pierre de Fermat



History of Statistics

Statistics, as a word, is derived from the German word "Statistik" meaning "description of a state/country."

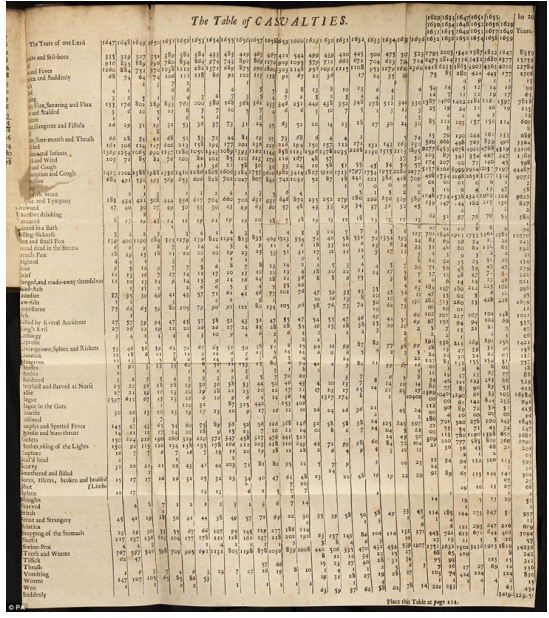
- Systematic collection of data on population and economy dates back to Renaissance (14th - 16th century).
- John Graunt (British, founder of demography, epidemiologist), natural and political observations made upon the **bills of mortality**, 1662.
- More and more regions started to systematically collect data from 1700s.
- Statistics became concerned with **inferencing conclusions** from a sample of numerical data, late 1800s.

Example: Bill of mortality

TABLE 1.2 *John Graunt's Mortality Table*

Age at Death	Number of Deaths per 100 Births
0–6	36
6–16	24
16–26	15
26–36	9
36–46	6
46–56	4
56–66	3
66–76	2
76 and greater	1

Note: The categories go up to but do not include the right-hand value. For instance, 0–6 means all ages from 0 up through 5.



Probability and Statistics Today: Virus Testing

A patient takes a virus test that returns positive. Consider:

- 0.3% of the population have Covid.
- PCR test correctly detects Covid with 95% chance.
- PCR test incorrectly detects Covid with 5% chance when absent.

What is the probability they really have the virus? Only about 5.4%!

`https://calculator.testingwisely.com/`

- How to design good tests? $\rightarrow$ Hypothesis testing.

The math:

$$\begin{aligned}\mathbb{P}(C|+) &= \frac{\mathbb{P}(+|C)\cdot\mathbb{P}(C)}{\mathbb{P}(+|C)\cdot\mathbb{P}(C)+\mathbb{P}(+|\neg C)\cdot\mathbb{P}(\neg C)} \\ &= \frac{0.95\times 0.003}{0.95\times 0.003+0.05\times 0.997}\approx 0.054\end{aligned}$$

Probability and Statistics Today: Online Retailers

Amazon uses statistics to understand and predict your preferences based on your browsing and purchasing history.

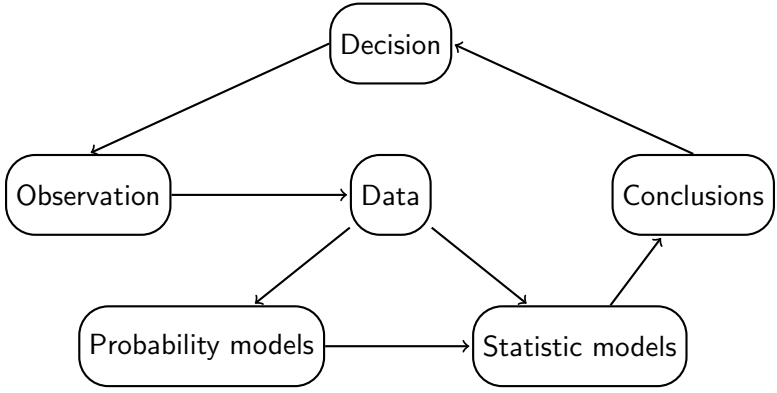
- **Data Collection:**
  - Tracks items you view, buy, and review.
  - Aggregates data from millions of users.
- **Pattern Identification:**
  - Uses statistical methods to find similarities between users and products.
  - Applies techniques like collaborative filtering and clustering.
- **Prediction:**
  - Predicts products you might be interested in.
  - Continuously updates recommendations as your behavior evolves.

Netflix Prize (2006-2009)

Netflix launched a $1 million contest to improve its movie recommendation system.

- **In-house algorithm:** Cinematch, a straightforward statistical linear model with extensive data conditioning (RMSE 0.9525).
- **Data context:** Before the era of "Big Data", over 100 million ratings of 17,770 movies from 480,189 customers were available.
- **Winner:** Team "BellKor's Pragmatic Chaos" achieved a 10% improvement with an RMSE of 0.8567.
- **Methods explored:** Regression, k-nearest-neighbor, matrix factorization, kernel ridge regression, neural networks, decision trees, and more.

What is this course about?



Statistics enable us to summarize, compare and predict more precisely than we normally would in everyday conversation.