CS 237: Probability in Computing

Wayne Snyder
Computer Science Department
Boston University
Overview of CS 237

Probability Theory

✧ What is a random process and how can we model it and analyze it?
✧ Finite probability and combinatorics (permutations and combinations)
✧ Conditional Probability
✧ Random Variables
✧ Joint probability of two random processes, Independence, Correlation
✧ Distributions (Discrete and Continuous)
✧ Central Limit Theorem
Overview of CS 237

**Descriptive Statistics:**
- What does the data look like?
- How can we summarize it?
- How can we display it graphically as an aid to understanding?
  - Diagrams, bar charts, scatter plots, etc.

**Inferential Statistics:**
- Sampling theory and statistical experiments
- Reasoning about sample statistics
- Standard Statistical Procedures:
  - Estimating means with Confidence Intervals
  - Hypothesis Testing
  - 2D Data: Correlation and Regression
Overview of CS 237

Probabilistic Algorithms: Using random processes to compute

- **Monte Carlo Algorithms:** Result may be incorrect with a small random error (e.g., calculation of π with a dart board)

- **Las Vegas Algorithms:** Result will be correct but running time is random (e.g., testing a program for correctness by randomly generating inputs)

- **Probabilistic Data Structures:** Randomness in storing data (e.g., hash tables, randomized BSTs, Bloom Filters)

- **Discrete Event Simulation:** Analyze a computer system or network by generating random workflows and measuring the performance using statistical tools (e.g., simulating an MM1 queueing system)
Randomness: Life is Uncertain!

Information Theory Definition of Randomness:

“The maximum entropy” OR “the minimum of information”

Example: Flip a coin – before you look at it: Is it heads or tails?

Our experience is filled with random events! Your chance in 2017 of

- being audited by IRS: 1 in 175
- finding a pearl in an oyster: 1 in 12,000
- winning $1,000,000 in Powerball: 1 in 11.6 million
- being killed by
  - a champagne cork:
  - a shark
  - a vending machine
  - a cow
  - hot tap water

Which are most and least likely?
Randomness: Life is Uncertain!

Information Theory Definition of Randomness:

“The maximum entropy OR the minimum of information”

Example: Flip a coin – before you look at it: Is it heads or tails?

Our experience is filled with random events! Your chance in 2017 of

- being audited by IRS: 1 in 175
- finding a pearl in an oyster: 1 in 12,000
- winning $1,000,000 in Powerball: 1 in 11.6 million
- being killed by
  - hot tap water: 1 in 3 million
  - a champagne cork: 1 in 13 million
  - a cow: 1 in 16 million
  - a vending machine: 1 in 146 million
  - a shark: 1 in 319 million
Randomness and Non-Randomness

But of course life is not completely uncertain, and in the last 350 years or so we have developed mathematical tools for understanding the difference:

“Probability Theory is the mathematical study of random phenomena.”
(Encyclopedia Britannica)

“Statistics is the science of learning from data, and of measuring, controlling, and communicating uncertainty....”
(American Statistical Association)

Much of this work has to do with discovering structure within a group or sequence of random events; many random phenomena behave in ways that are unpredictable in the short term, but have non-random characteristics when viewed as a whole – we seek to understand the difference!
Randomness and Non-Randomness

Examples of patterns within random events:

✧ **Example 1:** Flip a coin over and over; what is the average number of heads?

![Graph showing average number of heads after 10 flips]

Average after 10 flips = 0.3
Randomness and Non-Randomness

Examples of patterns within random events:

✧ **Example 1:** Flip a coin over and over; what is the average number of heads?
Randomness and Non-Randomness

Examples of patterns within random events:

◇ **Example 1:** Flip a coin over and over; what is the average number of heads?

![Graph showing the average number of heads after 50 flips]

Average after 50 flips = 0.48
Randomness and Non-Randomness

Examples of patterns within random events:

✧ **Example 1:** Flip a coin over and over; what is the average number of heads?
Randomness and Non-Randomness

Examples of patterns within random events:

**Example 1:** Flip a coin over and over; what is the average number of heads?
Randomness and Non-Randomness

Examples of patterns within random events:

✧ **Example 1**: Flip a coin over and over; what is the average number of heads?

![Graph showing average number of heads as coin flips increase.]
Randomness and Non-Randomness

Examples of patterns within random events:

- **Example 1:** Flip a coin over and over; what is the average number of heads?

The average number of heads ALWAYS approaches 0.5 as N gets larger!

One coin flip is random, but many coin flips are non-random....

Patterns emerge when we repeat random experiments over and over....
Randomness and Non-Randomness

Examples of patterns emerging when we “zoom out” and look at large numbers of seemingly random events:

✧ **Example 1:** Flip a coin over and over; what is the average number of heads?

✧ **Example 2:** What is the height of a human being?
Randomness and Non-Randomness

Examples of patterns emerging when we “zoom out” and look at large numbers of seemingly random events:

✧ **Example 1:** Flip a coin over and over; what is the average number of heads?

✧ **Example 2:** What is the height of a human being?

✧ **Example 3:** What is the IQ of a human being?

![IQ Normal Curve](image-url)
Randomness and Non-Randomness

Examples of patterns emerging when we “zoom out” and look at large numbers of seemingly random events:

✧ **Example 1:** Flip a coin over and over; what is the average number of heads?

✧ **Example 2:** What is the height of a human being?

✧ **Example 3:** What is the IQ of a human being?

✧ **Example 4:** What is the distribution of measurement errors?
Randomness and Non-Randomness

Examples of patterns emerging when we “zoom out” and look at large numbers of seemingly random events:

✧ **Example 1:** Flip a coin over and over; what is the average number of heads?
✧ **Example 2:** What is the height of a human being?
✧ **Example 3:** What is the IQ of a human being?
✧ **Example 4:** What is the distribution of measurement errors?
✧ **Example 5:** What is the expected return on a stock?
Descriptive Statistics

..... seeks to find these patterns in order to precisely describe a collection of data (the “population”) using:

- **Distributions**: the geometrical shape and extent of a data set;
- **Summary Statistics**: quantitative measures of its center, spread, symmetry, etc.
- **Data Graphics**: visual representations for specific applications

The **Distribution** of a data set is a graph of values vs. frequency (a “histogram”) in the population:

- **Example 1**: Toss a single die 20 times and record the number of dots each time:

The frequencies are as follows:

<table>
<thead>
<tr>
<th>Value</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>7</td>
<td>0</td>
<td>3</td>
</tr>
</tbody>
</table>
Descriptive Statistics

The **Distribution** of a data set is a graph of values vs. frequency (a “histogram”) in the population:

✧ **Example 1:** Toss a single die 20 times and record the number of dots each time:

```
Population: [4, 4, 4, 3, 2, 4, 6, 2, 1, 4, 1, 4, 4, 3, 6, 3, 6, 1, 2, 2]
```

![Histogram of Tosses of a Single Die -- 20 Trials](image-url)
Descriptive Statistics

The **Distribution** of a data set is a graph of values vs. frequency in the population:

✧ **Example 1:** Toss a single die 20 times and record the number of dots each time:

✧ **Example 2:** Flip a coin 10 times and count the number of heads; repeat 20 times and record the number of heads each time:
Descriptive Statistics

As in the case of the coin flips approaching an average of 0.5 heads per flip, each method of creating a population converges to a **Theoretical Distribution** as the number of trials increases:
Descriptive Statistics

There is a large number of Theoretical Distributions to describe different kinds of populations; these abstractions aid in the analysis of real data sets:

Uniform Distribution

Geometric Distribution

Binomial Distribution

Normal Distribution
Descriptive Statistics

Summary Statistics

✧ **Measures of Extent:**
  - Minimum, Maximum; Range; Quantiles

✧ **Measures of Center Point:**
  - Mean ("average"): balance point
  - Median: midpoint; 50\textsuperscript{th} percentile

✧ **Measures of Spread:**
  - Variance: mean of squared deviations from \( \mu \)
  - \( \sigma \) \textbf{Standard Deviation} = \sqrt{\text{Variance}}

The Standard deviation is particularly useful for the \textbf{Normal Distribution}, which follows the \textbf{68-95-99 rule:}

The Standard Deviation can indicate:

✧ Percentages centered around mean
✧ How unusual a value is: is it an outlier?
Descriptive Statistics

Example: Heights of 25000 US citizens were measured and found to be normally distributed with a mean of 68 inches and a standard deviation of 1.8 inches:

That means that, in our population,

- 68% are between 66.1 and 69.9 inches;
- 95% are between 64.2 and 71.8 inches;
- 99.7% are between 62.3 and 73.7 inches.
Descriptive Statistics

Data Graphics

- Box plots, violin plots
- Histograms 2D and 3D
- Scatterplots 2D and 3D
- Special purpose graphics
Inferential Statistics

.... is the process of deducing properties of a population by examining samples; properties of interest include:

- the underlying theoretical distribution;
- summary statistics such as the mean;
- tests of hypotheses;
- correlation and regression (for multi-dimensional data).

We conduct statistical experiments involving a (relatively small) random sample from a larger population, which can not be examined as a whole; properties of the sample are used to infer properties of the population. Often the estimate comes with a “confidence interval” telling us how certain we are of our result.

Think Polling!
Inferential Statistics

We will look at one example of a statistical procedure: inferring the population mean from the sample mean along with a “confidence interval.”
Inferential Statistics

IMPORTANT FACT: For samples of at least 30 values, the sample means for ANY population have an approximately normal distribution whose standard deviation decreases the larger the sample is:

![Image showing normal, skewed, uniform, and irregular population distributions with corresponding sampling distributions of sample means for different sample sizes (n=3, n=5, n=10, n=20).]
Review of CS 237

**Descriptive Statistics:**

✧ What does the data look like?
✧ How can we summarize it?
✧ How can we display it graphically as an aid to understanding?
  ✧ Diagrams, bar charts, scatter plots, etc.

**Inferential Statistics:**

✧ Sampling theory and statistical experiments
✧ Reasoning about sample statistics
✧ Standard Statistical Procedures:
  ✧ Estimating means with Confidence Intervals
  ✧ Hypothesis Testing
  ✧ 2D Data: Correlation and Regression
Review of CS 237

**Descriptive Statistics:**
- What does the data look like?
- How can we summarize it?
- How can we display it graphically?
  - Diagrams, bar charts, scatter plots, etc.

**Inferential Statistics:**
- Sampling theory and statistical experiments
- Reasoning about sample statistics
- Standard Statistical Procedures:
  - Estimating means with Confidence Intervals
  - Hypothesis Testing
  - 2D Data: Correlation and Regression
Review of CS 237

Descriptive Statistics:
✧ What does the data look like?
✧ How can we summarize it?
✧ How can we display it graphically?
   ✧ Diagrams, bar charts, scatter plots,

Inferential Statistics:
✧ Sampling theory and statistical experiments
✧ Reasoning about sample statistics
✧ Standard Statistical Procedures:
   ✧ Estimating means with Confidence Intervals
   ✧ Hypothesis Testing
   ✧ 2D Data: Correlation and Regression
Review of CS 237

**Descriptive Statistics:**
- What does the data look like?
- How can we summarize it?
- How can we display it graphically?
  - Diagrams, bar charts, scatter plots, etc.

**Inferential Statistics:**
- Sampling theory and statistical experiments
- Reasoning about sample statistics
- Standard Statistical Procedures:
  - Estimating means with Confidence Intervals
  - Hypothesis Testing
  - 2D Data: Correlation and Regression
  - Pattern Recognition/Machine Learning