Cogs 14B: Introduction to Statistical Analysis
Statistical Tools:
Description vs. Prediction/Inference

- Description
  - Averages
  - Variability
  - Correlation

- Prediction (Inference)
  - Regression
  - Confidence intervals/estimation
  - Tests (t-test, ANOVA)
1.- Generalizing beyond data

- Observation and generalization
- The role of inferential statistics
  - Goal: to form a conclusion about a characteristic of a population from the study of a sample taken from that population
  - Central idea: Sample outcomes vary
- Two types of inferential procedures
  - Hypothesis testing
  - Estimation
Populations

- A **real population** is a complete data set of observations about all subjects.
- A **hypothetical population** is a wider group of subjects, about which we don’t have data, but which (we assume) have similar characteristics.

Census: Count all subjects
Poll: Survey only a “representative” subset
2.- Population and Sample

- Population
  - A complete set of observations (or potential observations)
    - Real population: all potential observations are accessible at the time of sampling
    - Hypothetical population: potential observations are not accessible at the time of sampling

- Sample
  - A subset of observations from a population
Statistical inferences assume that observations represent the population

- Example: A survey worker is interested in how Americans spend their mornings. They call random phone numbers to survey respondents, asking them what they are currently doing.
  - Is this a valid sampling method?
  - How might the results be affected by the sampling method?
Population and Sample

- Sample size
  - It will depend on
    - Estimated variability among observations
    - Acceptable amount of probable error

- Techniques from Inferential Statistics use
  - Random Samples
Random Sample

- A sample produced when all potential observations in the population have *equal* chances of being selected
  - Casual? Haphazard? (not random)
  - Random numbers

- Random samples from Hypothetical populations?
  - In practice it’s impossible, but …we do “as if” …
Random Assignment of Subjects

- Important concept in design of experiments (e.g., control group vs. experimental groups)

- Purpose:
  - To ensure that (except for random differences) groups of subjects are similar with respect to any uncontrolled variables
II. Inferential Statistics (2)

- Introduction to Probability
- Basic notions
  - Trials, outcomes, events, sample space
- Basic operations
Probability

• The logic of knowledge and uncertainty

• A mathematical theory for reasoning about predictions

• “One may even say, strictly speaking, that almost all our knowledge is only probable; and in the small number of things that we are able to know with certainty, the principle means of arriving at the truth—induction and analogy—are based on probabilities”

• Laplace, P.S. Theorie Analytique des Probabilites (1812).
Defining probability

• The proportion or fraction of times that a particular event is likely to occur

• Probability summarizes our knowledge about what may happen in the future

• Probability theory tells us how to compute probabilities that give rise to rational (consistent) beliefs
  • For example, if \( P(A) > P(B) \) and \( P(B) > P(C) \), then \( P(A) > P(C) \)

• Dutch book theorem: If you place bets ignoring the rules of probability theory, you can be taken advantage of.
Mammalian brains encode uncertainty in the activity of dopamine neurons.

Figure 3 Sustained activation of dopamine neurons precedes uncertain rewards.

Christopher D. Fiorillo et al. Science 2003;299:1898-1902
new york times bestseller

noise and the noise

the signal and the noise and the noise and the noise and the noise and the noise and the noise and the noise

why so many noise predictions fail—but some don’t

nate silver

“The could turn out to be one of the more momentous books of the decade.” —The New York Times Book Review
Reasoning with probabilities:
The Monty Hall paradox
Basic notions

- **Trial**: is any operation or procedure whose outcomes cannot be predicted with certainty
- The set of all possible outcomes for a trial is the **sample space** for the trial
- **Event**: a subset of the sample space consisting of at least one outcome from the sample
  - Simple event (one outcome)
  - Compound event (more than one outcome)
Probability from frequency tables

- Blue histogram: Frequencies in the real population of previous launches
- Red normal curve: Estimated probabilities for the hypothetical population of future launches
Basic rules of probability theory

• **Probabilities are non-negative:** $0 \leq P(A) \leq 1$

• **Probabilities sum to 1:** The sum of probabilities over all possible events is 1

\[ P(A) + P(\text{not } A) = 1 \]

• **Addition rule:** If $A$ and $B$ are mutually exclusive events, then

\[ P(A \text{ or } B) = P(A) + P(B) \]

• **Multiplication rule:** If $A$ and $B$ are independent, then

\[ P(A \text{ and } B) = P(A)P(B) \]

• **Example:** What is the probability that Alice was born on a Monday, Tuesday or Friday?
  - $P(\text{Monday}) = P(\text{Tuesday}) = P(\text{Friday}) = 1/7$
  - $P(\text{M or T or F}) = 3/7$

• **Example:** What is the probability that Alice was born on Monday and Bob was born on Friday?
  - $P(\text{Alice on Monday}) = P(\text{Bob on Friday}) = 1/7$
  - $P(\text{Alice on Monday and Bob on Friday}) = 1/7 \times 1/7 = 1/49$
Independent, Dependent and Exclusive Events

- Independent: No relationship between $X$ and $Y$
- Dependent: Correlation (positive or negative) between $X$ and $Y$
- Exclusive: If $X$ is true then $Y$ is false, and if $Y$ is true then $X$ is false
Basic notions - Example

- The trial of rolling two dice has 36 equally likely outcomes (Sample space)
Basic notions

- The *probability of an event* is the sum of the probabilities of the elementary outcomes in the set
  - Event A: Dice add to 3
    - Outcomes: \{(1,2), (2,1)\}
    - \(P(A) = 2/36\)
  - Event B: Dice add to 6
    - Outcomes: \{(1,5), (2,4), (3,3), (4,2), (5,1)\}
    - \(P(B) = 5/36\)
  - Event C: Black die shows 1
    - Outcomes: \{(1,1), (2,1), (3,1), (4,1), (5,1), (6,1)\}
    - \(P(C) = 6/36\)
Basic operations

- Events (not just elementary outcomes) can be combined to make other events, using logical operations such as AND, OR, NOT
  - Event A AND Event B
  - Event A OR Event B
  - NOT Event E
  - etc

- We can calculate their probabilities!
Basic operations - Example

- Event C: White die is 1
- Event D: Black die is 1
Basic operations - Example

- **Addition rule (general case)**
  - Event C: White die is 1
  - Event D: Black die is 1
  - \[ P(C \text{ OR } D) = P(C) + P(D) - P(C \text{ AND } D) \]
  - \[ P(C \text{ OR } D) = \frac{6}{36} + \frac{6}{36} - \frac{1}{36} = \frac{11}{36} \]
Basic operations - Example

- **Addition rule (mutually exclusive events)**
  - Event E: Dice add to 3
  - Event F: Dice add to 6
  - \[ P(E \text{ OR } F) = P(E) + P(F) \]
  - \[ P(E \text{ OR } F) = \frac{2}{36} + \frac{5}{36} = \frac{7}{36} \]
Basic operations - Example

- **Conditional probability**: $P(A|B) = P(A)$ given $B$
- What is the probability that dice add to 3?
  - Event B: White die comes up 1
  - $P(A|B) = P(A \text{ AND } B) / P(B)$
  - $P(A|B) = (1/36) / (1/6) = 1/6$
Basic operations - Example

- **Conditional probability**: $P(A|B) = P(A)$ given $B$
  - $P(A|B) = P(A \text{ AND } B) / P(B)$

- **Multiplication rule**:
  - $P(A \text{ AND } B) = P(A|B) \times P(B)$

- When Events $A$ and $B$ are independent: $P(A|B)=P(A)$
  - $P(A \text{ AND } B) = P(A) \times P(B)$
Probability and Statistics

- Common and rare outcomes

![Normal Distribution Diagram]

- Rare outcomes: 
  - Critical value: -1.96
  - Area: 0.05

- Common Outcomes: 
  - Area: 0.95
  - Critical value: 1.96

- Rare outcomes: 
  - Area: 0.05
  - Critical value: -1.96
**Statistical “events”**

- A researcher administers a test to 100 subjects and computes a z-score for each one.
  - What is the probability of the “event” that Alice’s Z-score will be > 1.96?
    - 2.5%
  - What is the probability that Alice’s Z-score will be < 0?
    - 50%
  - Are these “events” independent, dependent, or mutually exclusive?
    - Mutually exclusive -- they represent non-overlapping areas under the normal curve
  - What is the probability that Alice’s Z-score will be < -2 or >0
    - 50 + 2.5 = 52.5%

![Graph showing probability distributions and calculations](image)
A researcher administers a test to 100 subjects and computes a z-score for each one.

- What is the probability of the “event” that Alice’s Z-score will be between 1 and 2?
- What is the probability that Alice’s Z-score will be > 0?
- Are these “events” independent, dependent, or mutually exclusive?
  - Dependent
- What is the probability that Alice’s Z-score will be > 0 and that it will be between 1 and 2? Can you use the sum or product rule?

\[ P(X \text{ or } Y) \neq P(X) + P(Y) \]

(can’t use sum or product rules)
Now consider two different students, Alice and Bob

What is the probability of the “event” that Alice’s Z-score will be < -2?

What is the probability that Bob’s Z-score will be > 0?

Are these “events” independent, dependent, or mutually exclusive?

Independent

What is the probability that Alice’s Z-score will be < -2 and Bob’s will be >0?

\[ P(A \text{ and } B) = P(A) \cdot P(B) = 0.025 \cdot 0.5 = 0.0125 = 1.25\% \]

Assuming Alice and Bob are two randomly chosen students, what is the probability that Alice’s Z-score will be < -2 or Bob’s score will be >0?

\[ P(A \text{ or } B) = P(A \text{ and } B) + P(A \text{ and not } B) + P(\text{not } A \text{ and } B) \]
\[ = P(A)P(B) + P(A)[1-P(B)] + [1-P(A)]P(B) \]
\[ = 0.0125 + 0.0125 + 0.975 \cdot 0.5 = 0.5125 \]
\[ = P[\text{not } (\text{not } A \text{ and not } B)] = 1-[1-P(A)][1-P(B)] \]
\[ = 1-0.975 \cdot 0.5 = 0.5125 \]
Conditional probability

- $P(A|B)$ is the probability of $A$ conditional on $B$.
- That is, if you know that $B$ happened, how likely is it that $A$ also happened?
- Example 1:
  - According to the almanac, it has rained on this date 5% of the time over the past 100 years. What is $P(\text{rain})$?
    - $P(\text{rain}) = 5\%$.
  - However, when you look out the window and see dark storm clouds. How does $P(\text{rain} \mid \text{clouds})$ compare to $P(\text{rain})$?
    - $P(\text{rain} \mid \text{clouds}) > P(\text{rain})$
- Example 2:

\[
\begin{array}{c}
\text{A} \\
\includegraphics[width=0.4\textwidth]{image1.png}
\end{array}
\quad
\begin{array}{c}
\text{B} \\
\includegraphics[width=0.4\textwidth]{image2.png}
\end{array}
\]

$P(B) = 50\%$

$P(B|A) = 100\%$
Conditional probability

- $P(X \text{ and } Y) = P(X \mid Y) \cdot P(Y)$
- $P(X \text{ and } Y) = 1 \cdot 0.5 = 0.5$

Event $X$: $1 < z < 2$
Event $Y$: $z > 0$

$P(X \text{ or } Y) \neq P(X) + P(Y)$
(can’t use sum or product rules)