Probability

Kosuke Imai

Department of Politics
Princeton University

Fall 2011

What Is Probability?

- A mathematical model of uncertainty (chance)
- READING: FPP Chapters 13 and 14; A&F 4.1

- Different interpretations:
  - Classical: physical laws
  - Frequentist: repeated experiments
  - Bayesian: subjective probability

Reverend Thomas Bayes (1702 – 1761)

Definition of Probability

- Experiment:
  1. flipping a coin
  2. rolling a die
  3. voting in an election
- Sample space Ω: all possible outcomes of the experiment
  1. $\{\text{head}, \text{tail}\}$
  2. $\{1, 2, 3, 4, 5, 6\}$
  3. $\{\text{abstain}, \text{Corzine}, \text{Christie}, \text{Daggett}\}$
- Event: any subset of outcomes in the sample space
  1. head, tail, head or tail, etc.
  2. 1, even number, odd number, does not exceed 3, etc.
  3. do not abstain, vote for major candidates, etc.

- $\Pr(A)$: probability that event A occurs
- If all outcomes are equally likely to occur, then we have

$$\Pr(A) = \frac{\text{number of elements in } A}{\text{number of elements in } \Omega}$$

Probability Axioms

From 3 axioms, the entire probability theory can be built!

1. Probability of any event is non-negative

$$\Pr(A) \geq 0$$

2. Prob. that one of the outcomes in the sample space occurs is 1

$$\Pr(\Omega) = 1$$

3. Addition Rule: If events A and B are *mutually exclusive*, then

$$\Pr(A \text{ or } B) \; = \; \Pr(A) + \Pr(B)$$

Useful Rules of Probability

1. Sometimes, it's easier to calcuate $1 - \Pr(\text{not } A) = \Pr(A)$

   Example: If $\Pr(\text{voting}) = 0.6$, then $\Pr(\text{not voting}) = 0.4$

2. Law of Total Probabilities: $\Pr(A) = \Pr(A \text{ and } B) + \Pr(A \text{ and not } B)$

   Example: If $\Pr(\text{dating}) = 0.5$ and $\Pr(\text{dating and happy}) = 0.4$, then $\Pr(\text{dating but unhappy}) = 0.1$

3. General addition rule: $\Pr(A \text{ or } B) = \Pr(A) + \Pr(B) - \Pr(A \text{ and } B)$

   Example: If $\Pr(\text{happy}) = 0.5$, then $\Pr(\text{dating or happy}) = 0.6$

Conditional Probability

- $\Pr(A \mid B)$ is the conditional probability of event A occurring given that event B occurs
  Example: $\Pr(\text{vote for Corzine} \mid \text{support Obama})$
- Multiplication Rule:

$$\underbrace{\Pr(A \text{ and } B)}_{\textit{joint probability}} = \overbrace{\Pr(A \mid B)}^{\textit{conditional probability}} \times \underbrace{\Pr(B)}_{\textit{marginal probability}}$$

- A couple is expecting twins...
  1. In a ultrasound exam, the technician was only able to determine that one of the two was a boy. What is the probability that both are boys?
  2. During the delivery, the baby that was born first was a boy. What is the probability that both are boys?

Independence

- Independence: Two events A and B are said to be independent if

$$\Pr(A \text{ and } B) = \Pr(A)\Pr(B)$$

- If A and B are independent, then

$$\Pr(A \mid B) = \Pr(A)$$

- Monty Hall Problem: You must choose one of three doors where one conceals a new car and two conceal old goats. After you randomly choose one door, the host of the game show, Monty, opens another door which does not conceal a new car. Then, Monty asks you if you would like to switch to the (unopened) third door. Should you switch?

Bayes' Rule

- From the conditional probability formula, we have the following Bayes' rule

$$\underbrace{\Pr(A \mid B)}_{\textit{conditional probability}} = \frac{\overbrace{\Pr(A \text{ and } B)}^{\textit{joint probability}}}{\underbrace{\Pr(B)}_{\textit{marginal probability}}} = \frac{\Pr(B \mid A)\Pr(A)}{\Pr(B \mid A)\Pr(A) + \Pr(B \mid \text{not} A)\Pr(\text{not} A)}$$

- Knowledge of $\Pr(A)$, $\Pr(B \mid A)$, and $\Pr(B \mid \text{not} A)$ gives you $\Pr(A \mid B)$
- Bayesian update: prior belief $\Pr(A) \stackrel{B:\text{Data}}{\Longrightarrow}$ posterior belief $\Pr(A \mid B)$

Educating OB/GYN

First Trimester Screening Test Problem

A 35 year old pregnant woman is told that 1 in 378 women of her age will have a baby with Down Syndrome (DS). A first trimester ultrasound screening procedure indicates that she is in a high-risk category. Of 100 cases of DS 86 mothers would have received a high-risk result and 14 cases of DS will be missed. There is a 1 in 20 chance for a normal pregnancy to be called high-risk. Given the result of the screening procedure, what is the probability that her baby has DS? What would be the probability if the result had been negative?

Permutations

- READING: FPP Chapter 15
- How to count # of ways to arrange objects?
- $\{A, B, C\}$ has 6 permutations: $ABC, ACB, BAC, BCA, CAB, CBA$
- Sampling without replacement: # of permutations of n elements taken k at a time

$$_nP_k = n(n-1)\cdots(n-k+1) = \frac{n!}{(n-k)!}$$

where $k \leq n$ and $0! = 1$

- Birthday problem: How many people do you need in order for the probability that at least two people have the same birthday to exceed 0.5?

Combinations

- Ways to select objects without regard to their arrangement
- # of combinations of k distinct elements from a pool of n elements

$$ {}_nC_k = \binom{n}{k} = \frac{{}_nP_k}{k!} = \frac{n!}{k!(n-k)!} $$

- One combination yields $k!$ permutations: order does not matter
- Grading problem: There are 7 students in a class. The professor decides to randomly select two students to receive an A. What is the probability that exactly one of the two best students gets an A?

  # of ways to select two students out of seven: ${}_7C_2 = 21$
  # of ways to select one best: ${}_2C_1 = 2$
  # of ways to select one of the five low ranking students: ${}_5C_1 = 5$
  # of ways to select exactly one of the two best students: $2 \times 5 = 10$
  The required probability $= 10/21 \approx 0.48$

Schwarzenegger's Veto Message (October 2009)

To the Members of the California State Assembly:

I am returning Assembly Bill 1176 without my signature.

For some time now I have lamented the fact that major issues are overlooked while many unnecessary bills come to me for consideration. Water reform, prison reform, and health care are major issues my Administration has brought to the table, but the Legislature just kicks the can down the alley.

Yet another legislative year has come and gone without the major reforms Californians overwhelmingly deserve. In light of this, and after careful consideration, I believe it is unnecessary to sign this measure at this time.

Sincerely,

Arnold Schwarzenegger

- “My goodness. What a coincidence,” said Schwarzenegger spokesman Aaron McLear. “I suppose when you do so many vetoes, something like this is bound to happen.”

- What is the probability that this “coincidence” happens by chance?
- In the precept, you will calculate this probability

People v. Collins

A purse snatching in which witnesses claimed to see a young women with blond hair in a ponytail fleeing from the scene in a yellow car driven by a black young man with a beard. A couple meeting the description was arrested a few days after the crime, but no physical evidence was found. The probability that a randomly selected couple would possess the described characteristics was estimated to be about one in 12 million. Faced with such overwhelming odds, the jury convicted the defendants. Given that there was already one couple who met the description, what is the conditional probability that there was also a second couple such as the defendants?

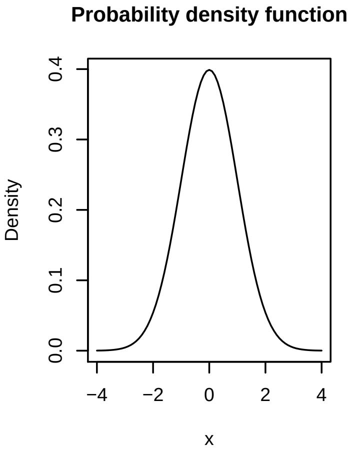
Random Variables and Probability Distributions

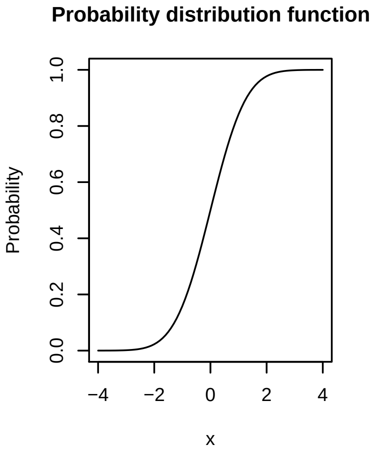
- READING: A&F 4.2–4.3
- What is a random variable?: assigns a number to an event
  1. Coin flip: `head` $= 1$ and `tail` $= 0$
  2. Gambling: `win` $= \$100$ and `lose` $= -\$10$
  3. Voting: `vote` $= 1$ and `not vote` $= 0$
  4. Survey response: `strongly agree` $= 4$, `agree` $= 3$, `disagree` $= 2$, and `strongly disagree` $= 1$
  5. Income: `earned $100K` $= 100$

- Probability model:
  1. Probability density function: $f(x)$
     - How likely does X take a particular value?
     - When X is discrete, $f(x) = \Pr(X = x)$
  2. Probability distribution function: $F(x) = \Pr(X \leq x)$
     - What is the probability that a random variable X takes a value equal to or less than x?
     - Area under the density curve
     - Non-decreasing

Standard Normal Distribution

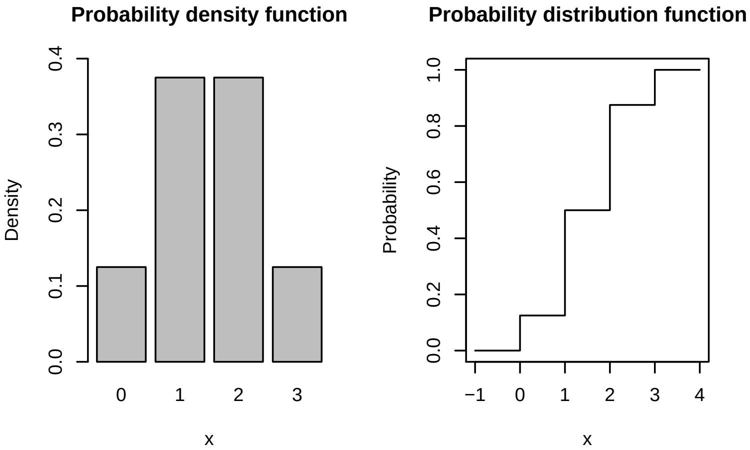
- $f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$
- Standard normal table





Binomial Distribution

- $f(x) = \binom{n}{x} p^x (1-p)^{n-x}$ where $x \in \{0, 1, \ldots, n\}$
- Example: flip a fair coin 3 times



The Probability of Your Vote Being Decisive

Minnesota's 2008 senate race was very close. The difference between the Democratic and Republican candidate was only 312 votes. Assume that each voter is equally likely to vote for either the Democrat or the Republican. And each vote is independent. What is the probability of your vote being decisive when the number of voters is equal to 400,000?

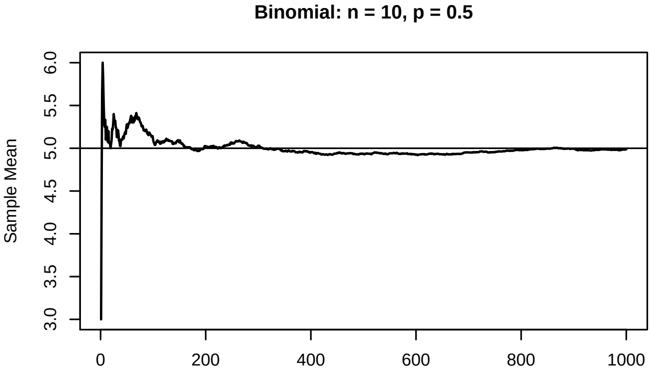
- Total number of Democratic votes: $\text{Binom}(400000, 0.5)$
- The probability of an exact tie:

$$\texttt{dbinom}(\texttt{x} = \texttt{200000}, \texttt{size} = \texttt{400000}, \texttt{prob} = \texttt{0.5}) \approx 0.126\%$$

- About 20,000 contested elections for the US Congress between 1900 and 1990, but none of these were tied
- However, six of these were within 10 votes of being tied and 49 were within 100 votes

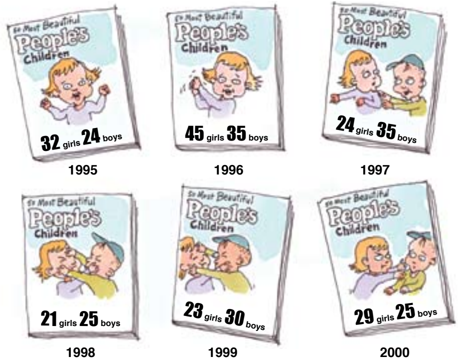
Law of Averages

- READING: FPP Chapters 16 and 17
- Law of Large Numbers: As the sample size increases, the sample average approaches to a value called "expected value"
- Example:
  1. flip a coin 10 times and count # of heads
  2. repeat it many times and compute the sample mean



Do Beautiful People Have More Girls?

- In *Journal of Theoretical Biology*,
  1. "Big and Tall Parents have More Sons" (2005)
  2. "Engineers Have More Sons, Nurses Have More Daughters" (2005)
  3. "Violent Men Have More Sons" (2006)
  4. "Beautiful Parents Have More Daughters" (2007)



- Law of Averages in action
  1. 1995: 57.1%
  2. 1996: 56.6
  3. 1997: 51.8
  4. 1998: 50.6
  5. 1999: 49.3
  6. 2000: 50.0
- No dupilicates: 47.7%
- Population frequency: 48.5%

Gelman & Weakliem, *American Scientist*

Expected Values

- Expected value of a random variable X: $\mathbb{E}(X)$
- On average, what is the value of a random variable?
- *Population* average (mean) value of $X \neq$ sample mean

$$\mathbb{E}(X) = \begin{cases} \sum_x xf(x) & \text{if } X \text{ is discrete} \\ \int xf(x)dx & \text{if } X \text{ is continuous} \end{cases}$$

- Example: throwing a die
  - Expected value: $1 \times \frac{1}{6} + 2 \times \frac{1}{6} + \cdots + 6 \times \frac{1}{6} = \frac{7}{2}$
  - Sample mean of three throws $\{1, 5, 3\}$: $\frac{1+5+3}{3} = 3$
- Inherits the properties of sample mean,

$$\begin{aligned} \mathbb{E}(X + Y) &= \mathbb{E}(X) + \mathbb{E}(Y) \\ \mathbb{E}(aX + b) &= a\mathbb{E}(X) + b \end{aligned}$$

- Standard normal: $\mathbb{E}(X) = 0$
- Binomial: $\mathbb{E}(X) = np$

Variance and Standard Deviation

- Recall (sample) standard deviation
- Population variance:

$$\mathbb{V}(X) = \mathbb{E}[\{X - \mathbb{E}(X)\}^2] = \mathbb{E}(X^2) - \{\mathbb{E}(X)\}^2$$

- Population standard deviation: $\sqrt{\mathbb{V}(X)}$
- It inherits the properties of sample variance:

$$\mathbb{V}(aX + b) \ = \ a^2\mathbb{V}(X)$$

- If X and Y are *independent*, $\mathbb{V}(X + Y) = \mathbb{V}(X) + \mathbb{V}(Y)$
- Standard normal: $\mathbb{V}(X) = 1$
- Binomial: $\mathbb{V}(X) = np(1 - p)$

Examples Using Normal Distribution

- If X and Y are normal random variables, then $aX + bY$ is also normally distributed with appropriate mean and variance
- z-score:

$$Z = \frac{X - \mathbb{E}(X)}{\sqrt{\mathbb{V}(X)}} \sim \mathcal{N}(0, 1)$$

- Sum: X_i is independently distributed as $\mathcal{N}(\mathbb{E}(X), \mathbb{V}(X))$

$$\sum_{i=1}^{n} X_i \sim \mathcal{N}(n\mathbb{E}(X), n\mathbb{V}(X))$$

- Sample mean:

$$\overline{X} \sim \mathcal{N}\left(\mathbb{E}(X), \frac{\mathbb{V}(X)}{n}\right)$$

- Regression: $Y_i = -15 + 1.2X_i + \epsilon_i$ with $X_i \sim \mathcal{N}(60, 16)$ and $\epsilon_i \sim \mathcal{N}(0, 25)$
  1. $Y_i \sim \mathcal{N}(57, 48.04)$
  2. Y_i given $X_i = 60$ is $\sim \mathcal{N}(57, 25)$

Election Polls

- Sample size n
- $X_i = 1$ if supports Obama, $X_i = 0$ if supports McCain
- Simplifying assumption: $\sum_{i=1}^{n} X_i \sim \mathrm{Binom}(n, p)$

- Obama's support rate: $\overline{X} = \sum_{i=1}^{n} X_i / n$
- $\mathbb{E}(\overline{X}) = p$ and $\mathbb{V}(\overline{X}) = p(1-p)/n$

- Margin of victory: $\delta = p - (1-p) = 2p - 1$
- Estimate: $\hat{\delta} = 2\overline{X} - 1$
- $\mathbb{E}(\hat{\delta}) = \delta$ and $\mathbb{V}(\hat{\delta}) = 4\mathbb{V}(\overline{X}) = 4p(1-p)/n$

Central Limit Theorem

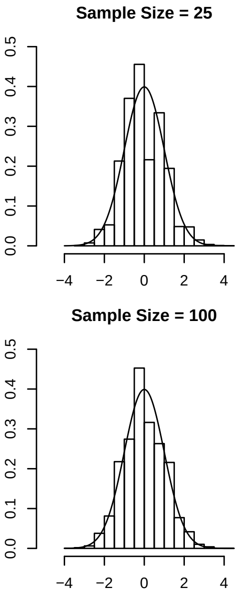
- READING: FPP Chapter 18
- What is the distribution of sample mean $\overline{X}$ when X is not normally distributed?
- Polling example: repeated (often hypothetical) polls
- The approximate (asymptotic) distribution of $\overline{X}$ is still normal!
- In particular, when n is large, we have

$$\overline{X} \sim \mathcal{N}\left(\mathbb{E}(X), \frac{\mathbb{V}(X)}{n}\right)$$

- Theorem: As the sample size increases, the distribution of the z-score for the sample mean,

$$Z = \frac{\overline{X} - \mathbb{E}(\overline{X})}{\sqrt{\mathbb{V}(\overline{X})}} = \frac{\overline{X} - \mathbb{E}(X)}{\sqrt{\mathbb{V}(X)/n}}$$

approaches to $\mathcal{N}(0, 1)$


Sample Size = 25
Sample Size = 50
Sample Size = 100
Sample Size = 500
0.0
0.1
0.2
0.3
0.4
0.5
−4
−2
0
2
4