STAT 111

Recitation 1

Mo Huang

Email: mohuang@wharton.upenn.edu

Office Hours: Wednesdays 3:00 - 4:00 pm, JMHH F96

Slides (adapted from Gemma Moran): `github.com/mohuangx/STAT111-Fall2018`

September 7, 2018

Logistics

- Section 201: 11:00 am - 11:50 am
- Section 202: 12:00 pm - 12:50 pm
- Every Friday, I will:
  1. Collect the homework for that week;
  2. Give you your graded homework; and
  3. Give you the homework for the week after (also posted on Canvas).
- If you do not collect your homework, it will be available in the STAT 111 box in the Statistics Department (4th Floor, JMHH).
- If you know in advance that you will not be able to attend a recitation, give your homework to Professor Ewens on Thursdays or put it in my mailbox (at entrance to the Statistics Department)
- Questions about course materials should be asked during office hours or in an email to Professor Ewens.

Statistics and Probability

- Probability theory is *deductive* (or "top-down logic"). We start with a theory and then consider the implications resulting from that theory.
- Statistics is *inductive* (or "bottom-up logic"). We start with finite observations and then attempt to make objective statements about the world from a sample.
- Statistical methods use probability theory to draw conclusions from observed data.

Statistics and Probability: Example

- We currently cure an illness with some medicine ("current medicine") which cures 80% of patients.
- A new medicine is proposed ("new medicine") which cures 1,643 people out of 2,000.

Is the new medicine more effective?

Probability theory

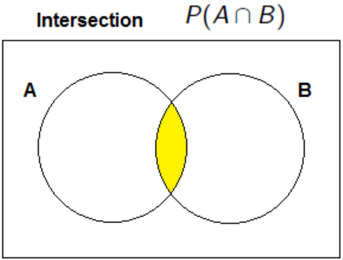
Assume new medicine is as effective as current medicine → Probability that 1,643 or more people are cured by new medicine is 0.009

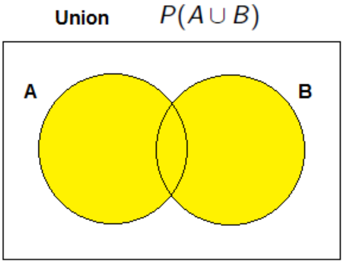
Conclude new medicine is more effective ← We observe 1,643 people cured by new medicine

Statistics

Probability: Axioms

- $0 \leq P(A) \leq 1$, where A is any event.
- $P(A^C) = 1 - P(A)$ where A^C is the complement of A.
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$





Probability: Independence

- A and B are independent if and only if $P(A \cap B) = P(A)P(B)$
- Conditional probability: $P(A|B) = \dfrac{P(A \cap B)}{P(B)}$

Suppose A and B are independent. Then

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A)P(B)}{P(B)} = P(A)$$

So we have two definitions of independence:

1. $P(A \cap B) = P(A)P(B)$
2. $P(A|B) = P(A)$

Note: This year, conditional probability is not on the syllabus but it's still an interesting and useful concept!

Probability: Questions

Q1: A smoke alarm consists of two parts, A and B. If there is smoke, part A will detect it with probability 0.96 and part B will detect it with probability 0.98. The probability that they both detect it is 0.95. What is the probability that smoke will not be detected?

A1: Let A be the event that part A detects smoke and let B be the event that part B detects smoke. What is the event that smoke is not detected?

$A \cup B =$ Smoke is detected $\rightarrow [A \cup B]^C =$ Smoke is not detected

$$\begin{aligned} P(A \cup B) &= P(A) + P(B) - P(A \cap B) \\ &= 0.96 + 0.98 - 0.95 \\ &= 0.99 \\ P([A \cup B]^C) &= 1 - P(A \cup B) \\ &= 0.01 \end{aligned}$$

Probability: Questions

Q2: A six-sided die is biased such that an odd number is twice as likely to occur as an even number. That is, the probability of an odd number is $2/9$ and an even number is $1/9$. Let A be the event that an even number occurs. Let B be the event that a number greater than or equal to 4 occurs. Find $P(A)$, $P(B)$, $P(A \cup B)$ and $P(A \cap B)$.

A2:

$$\begin{aligned} P(A) &= P(2, 4 \text{ or } 6) \\ &= 1/9 + 1/9 + 1/9 \\ &= 3/9 \\ P(A \cap B) &= P(4 \text{ or } 6) \\ &= 1/9 + 1/9 \\ &= 2/9 \end{aligned}$$

$$\begin{aligned} P(B) &= P(4, 5 \text{ or } 6) \\ &= 1/9 + 2/9 + 1/9 \\ &= 4/9 \\ P(A \cup B) &= P(A) + P(B) - P(A \cap B) \\ &= 3/9 + 4/9 - 2/9 \\ &= 5/9 \end{aligned}$$

Probability: Questions

A2: $P(A) = 3/9, \quad P(B) = 4/9, \quad P(A \cap B) = 2/9, \quad P(A \cup B) = 5/9$

Q3: Following Q2, find $P(A|B)$.

A3:

$$\begin{aligned} P(A|B) &= \frac{P(A \cap B)}{P(B)} \\ &= \frac{2/9}{4/9} \\ &= 1/2 \end{aligned}$$

Q4: Are A and B independent?

A4: No, because $P(A|B) \neq P(A)$. (Recall $P(A) = 1/3$).

Probability: Questions

Q5: An unfair coin ($P(H) = 0.4$) was flipped twice. Given that on at least one flip a head occurred, what is the probability that a head occurred on both flips?

A5: We want to find $P(HH|\text{at least one } H)$.

$$\begin{aligned}
P(\text{at least one } H) &= P(HH) + P(HT) + P(TH) \\
&= (0.4)^2 + (0.4)(0.6) + (0.6)(0.4) \\
&= 0.64. \\
P(HH|\text{at least one } H) &= \frac{P(HH \cap \text{at least one } H)}{P(\text{at least one } H)} \\
&= \frac{P(HH)}{P(\text{at least one } H)} \\
&= \frac{(0.4)^2}{0.64} \\
&= 0.25
\end{aligned}$$

Parameters (θ), Random Variables (X), and Data (x)

- A parameter represents some underlying numerical constant of a phenomenon. Represented by Greek letters.
- A random variable is a numerical outcome of interest in a future experiment. Can be modeled by a probability distribution which depends on the parameter. Represented by capital letters.
- Data is the realization or observed value of a random variable after performing the experiment. Represented by lower-case letters.

Fair coin-flipping example:

- A parameter θ is the underlying probability of obtaining a head. $\theta = 0.5$.
- A random variable X is the number of heads obtained in 10 tosses. Can be modeled by a binomial distribution dependent on θ.
- Data $x = 6$ is observing 6 heads in 10 tosses, a realization of X.

Random Variables

- Two types of random variables: discrete and continuous.
- A discrete random variable is a random variable that can only take on a countable set of numbers.
- The probability distribution of a discrete random variable is the range of values it can take *and* the probabilities of these values.
- For example:
  - Let X be the number of heads I toss if I toss a fair coin 3 times.

x	0	1	2	3
$P(X = x)$	0.125	0.375	0.375	0.125

Table: Probability distribution of X.

Random Variables: Mean

- The mean of a random variable (expected value) is the long-run average of realizations of the random variable over repeated experiments.

- This mean of random variable X is different from the sample mean, which is the average of a *finite* number of observations (x).

- Let X be a discrete random variable that can take values $\{v_1, v_2, \ldots, v_k\}$. Then the mean of X is given by:

$$\begin{aligned}\mu &= \sum_{i=1}^{k} v_i P(X = v_i) \\ &= v_1 P(X = v_1) + v_2 P(X = v_2) + \cdots + v_k P(X = v_k).\end{aligned}$$

Note: We use μ to denote the mean of a random variable.
Note: Can think of it as a weighted average, weighted by probability.

The Binomial Distribution

- The binomial distribution arises if:
  1. We plan to conduct a fixed number of experiments. We denote the number of experiments as n.
  2. In each experiment, there are two outcomes: "success" or "failure".
  3. The experiments are independent.
  4. The probability of a success is the same for each experiment.
- For example:
  1. I plan to toss a coin n times.
  2. I can toss either a head or a tail.
  3. Each coin toss is independent of the next - it doesn't matter whether I get a head or a tail on the previous toss.
  4. The probability of getting a head is the same for each toss.

The Binomial Distribution

- Let X be a binomial random variable where $\theta = P(\text{success})$ and there are n experiments. Then the probability distribution of X is given by:

$$P(X = x) = \binom{n}{x} \theta^x (1-\theta)^{n-x}, \quad \text{for } x = 0, 1, 2, \ldots, n.$$

  - θ is called a parameter: a constant whose value may be known or unknown.
  - $\binom{n}{x}$ is said as "n choose x": it is the number of ways x successes can occur in n experiments.

Note: $X \sim \mathcal{B}(n, \theta)$ means "X is a binomial random variable with n experiments and probability of success θ."

The Binomial Distribution: Mean

- For the binomial distribution, we have a simpler formula for the mean: if $X \sim \mathcal{B}(n, \theta)$, then the mean of X is

$$\mu = n\theta.$$

- For example:
  - On the previous slide, we calculated the mean of a binomial random variable with $n = 3$ and $\theta = 0.5$ to be 1.5. This is exactly $n\theta$.

Note: This formula is *only* for the binomial distribution.