SKKU Biostats and Big data

CHOONG-WAN WOO | COCOAN lab | http://cocoanlab.github.io

Lecture 10
Probability

Review: Key Points

Chapter 11 and 12: Randomness, Sampling

- Population parameters vs. sample statistics
- Sampling methods:
  - Simple sampling, stratified sampling, cluster sampling, multistage sampling, systematic sampling
- Common mistakes in sampling…

Chapter 13: Experiments and Observational Studies

- Observational studies: Valuable for discovering trends and relationships, but correlation ≠ causation
- **Four principles of experimental design:** Control. Randomize, Replicate, Block
- Good experiments: Randomized, Comparative (control), Double-blind, Placebo-controlled
- Confounding

Random phenomena

- Refers to a situations where we know what kinds of outcomes can possibly occur, but don't know which particular outcome will happen.

- **Trial:** Each occasion when we observe a random phenomenon
- **Outcome/Event**: the value of the random phenomenon
- **Sample space (S)**: all possible outcomes
  - E.g., $\mathbf{S} = \{\text{red}, \text{green}, \text{yellow}\}$, $\mathbf{S} = \{\text{H}, \text{T}\}$, $\mathbf{S} = \{\text{HH}, \text{HT}, \text{TH}, \text{TT}\}$

Law of Large Numbers (LLN)

- When we repeat a random process over and over, the proportion of times that an event occurs settle down to one number, which is the **probability** of the event.

- Two key assumption of LLN:
  - **No changes** of the random phenomena over the repeats
  - **independence** (all the events should be independent): an event doesn't influence the outcomes of others.

- This tells us nothing about "*short-run*" behavior. And the *long run* is really long.
  - Don't expect the probability tells you a *short-term trend* of something.

Law of Large Numbers (LLN)

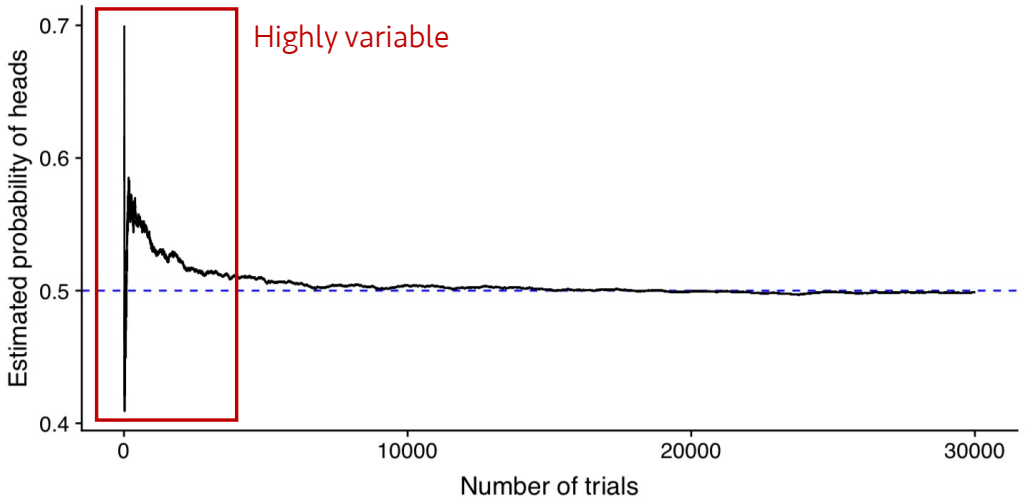
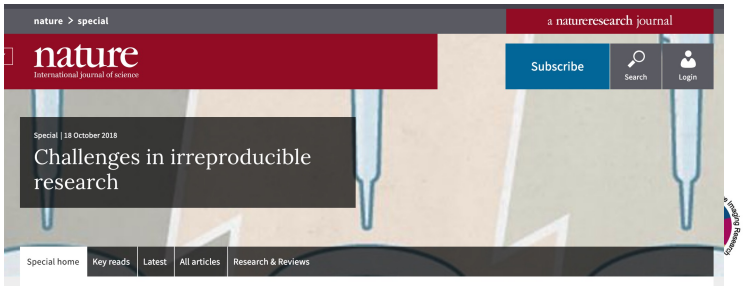


Figure 3.1: A demonstration of the law of large numbers. A coin was flipped 30,000 times, and after each flip the probability of heads was computed based on the number of heads and tail collected up to that point. It takes about 15,000 flips for the probability to settle at the true probability of 0.5.

- Many people forget this and overinterpret results from small samples: *Law of small numbers* (Tversky and Kahneman)
- One of the major sources for the recent replication crisis!
- https://www.nature.com/collections/prbfkwmwvz



Quiz 10-1

https://forms.gle/KYB4Sk7BJdat4Rtf6

Rules of formal probability

- **Rule 1.** For any event A, $0 \leq \boldsymbol{P}(\mathbf{A}) \leq 1$
- **Rule 2.** $\boldsymbol{P}(\boldsymbol{S}) = 1$: the set of all possible outcomes of a trial must have probability 1
- **Rule 3.** $\boldsymbol{P}(\mathbf{A}^C) = 1 - P(\mathbf{A})$: the probability of an event not occurring is 1 minus the probability that it occurs.
- **Rule 4 (Addition rule).** $\boldsymbol{P}(\mathbf{A}\ or\ \mathbf{B}) = \boldsymbol{P}(\mathbf{A}) + \boldsymbol{P}(\mathbf{B})$, provided that A and B are disjoint (or mutually exclusive).
- **Rule 5 (Multiplication rule).** $\boldsymbol{P}(\mathbf{A}\ and\ \mathbf{B}) = \boldsymbol{P}(\mathbf{A}) \times \boldsymbol{P}(\mathbf{B})$, provided that A and B are independent.

The General Addition Rules

- Example: Survey of college students:
  - 56% live on campus, 62% have a campus meal plan, and 42% do both.

- What's the probability of "living on campus or having a campus meal plan"?
  - $P(\mathbf{L}\ or\ \mathbf{M}) = P(\mathbf{L}) + P(\mathbf{M}) - P(\mathbf{L}\ and\ \mathbf{M})$
  - = 0.56+0.62-0.42 = 0.76

- What's the probability of "living *off* campus and not having a campus meal plan"?
  - $P(\mathbf{L}^c\ and\ \mathbf{M}^c) = 1 - P(\mathbf{L}\ or\ \mathbf{M}) = 1\ -\ \big(P(\mathbf{L}) + P(\mathbf{M}) - P(\mathbf{L}\ and\ \mathbf{M})\big)$ = 0.24

- General Addition Rules:
  - $P(\mathbf{A}\ or\ \mathbf{B}) = P(\mathbf{A}) + P(\mathbf{B}) - P(\mathbf{A}\ and\ \mathbf{B})$

Conditional Probability

- Example: surveyed 478 elementary school students and asked whether their primary goal was to get good grades, to be popular, or to be good at sports.
- Contingency table:

		Goals			
		Grades	Popular	Sports	Total
Sex	Boy	117	50	60	227
	Girl	130	91	30	251
	Total	247	141	90	478

- *P*(Sports) = 90/478 = 0.188
- *P*(Girl) = 251/478 = 0.525
- *P*(Sports and Girl) = 30/478 = 0.063
- *P*(Sports | Girl) = Probability of Sports *given* Girl = 30/251 = 0.120

  = P(Sports and Girl) / P(Girl) = (30/478) / (251/478) = 0.063 / 0.525 = 0.120

General Multiplication Rule

- Reminder:
  - Rule 5 (Multiplication rule). $P(\mathbf{A}\ and\ \mathbf{B}) = P(\mathbf{A}) \times P(\mathbf{B})$, provided that A and B are independent.

- Now, this is its general form:
  - $P(\mathbf{A}\ and\ \mathbf{B}) = P(\mathbf{A}) \times P(\mathbf{B}|\mathbf{A}) = P(\mathbf{B}) \times P(\mathbf{A}|\mathbf{B})$

- It makes sense: *P*(Girl *and* Sports) = *P*(Girl) x *P*(Sports | Girl) = 251/487 x 30/251 = 30/487

		Goals			
		Grades	Popular	Sports	Total
Sex	Boy	117	50	60	227
	Girl	130	91	30	251
	Total	247	141	90	478

Independence

- Events A and B are independent if and only if:
  - $P(\mathbf{B}|\mathbf{A}) = P(\mathbf{B})$

- Thus, if A and B are independent,
  - $P(\mathbf{A}\ and\ \mathbf{B}) = P(\mathbf{A}) \times P(\mathbf{B}|\mathbf{A}) = P(\mathbf{A}) \times P(\mathbf{B})$

- Independence ≠ Disjoint
  - **Disjoint:** *P*(A *and* B) = 0: the events cannot happen simultaneously
  - **Independence:** the occurrence of A does not change the probability of B, $P(\mathbf{B}|\mathbf{A}) = P(\mathbf{B})$

Quiz 10-2

https://forms.gle/z4uUg3sW6pwvxXrFA

Reversing the Conditioning and Bayes' Rule

- Let's say we have $P(\mathbf{A}|\mathbf{B})$, then can we calculate $P(\mathbf{B}|\mathbf{A})$ from it?
  - To calculate it, we need to know $P(\mathbf{A})$ and $P(\mathbf{B})$

- Because of this: $P(\mathbf{A}\ and\ \mathbf{B}) = P(\mathbf{A}) \times P(\mathbf{B}|\mathbf{A})$
  - $P(\mathbf{B}|\mathbf{A}) = \frac{P(\mathbf{A}\ and\ \mathbf{B})}{P(\mathbf{A})}$, then remember this:

  - Now, its general form:
  - $\boldsymbol{P(\mathbf{A}\ and\ \mathbf{B}) = P(\mathbf{A}) \times P(\mathbf{B}|\mathbf{A}) = P(\mathbf{B}) \times P(\mathbf{A}|\mathbf{B})}$

- Then, we can express the equation like the following: $P(\mathbf{B}|\mathbf{A}) = \frac{P(\mathbf{B}) \times P(\mathbf{A}|\mathbf{B})}{P(\mathbf{A})}$
- where *P*(A) can be expressed as $P(\mathbf{A}) = P(\mathbf{A}|\mathbf{B})P(\mathbf{B}) + P(\mathbf{A}|\mathbf{B^c})P(\mathbf{B^c})$
- Therefore,

$$P(\mathbf{B}|\mathbf{A}) = \frac{P(\mathbf{A}|\mathbf{B})P(\mathbf{B})}{P(\mathbf{A}|\mathbf{B})P(\mathbf{B}) + P(\mathbf{A}|\mathbf{B^C})P(\mathbf{B^C})}$$

Practice:

- *P*(test+ | disease) = 0.8
- *P*(disease | test+) = 0.1
- How is this possible? In which cases (e.g., what numbers or ratio of P(test+) and P(disease)), can this happen?

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

$$P(\mathbf{T}+|\mathbf{D}) = \frac{P(\mathbf{D}|\mathbf{T}+) \times P(\mathbf{T}+)}{P(\mathbf{D})}$$

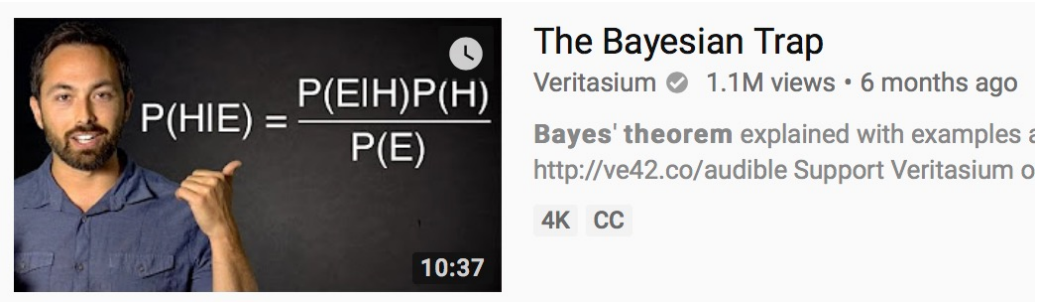
$$\frac{P(\mathbf{T}+|\mathbf{D})}{P(\mathbf{D}|\mathbf{T}+)} = \frac{P(\mathbf{T}+)}{P(\mathbf{D})}$$

$$\frac{0.8}{0.1} = \frac{P(\mathbf{T}+)}{P(\mathbf{D})}$$

- Too many test positive when the disease prevalence is low.. Is this a good test?

A good video on Bayes' theorem

https://www.youtube.com/watch?v=R13BD8qKeTg



First test result came out to be positive:

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

Best $P(\mathrm{B})$: prevalence

If the second test results came out to be positive again,

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

Best $P(\mathrm{B})$ this time: $P(\mathrm{B}|\mathrm{A})$ from the first test result

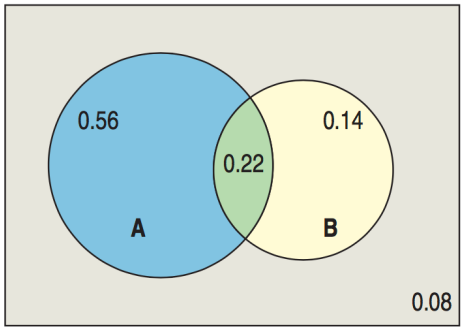
Then, $P(\mathrm{B}|\mathrm{A})$ becomes much higher than the first time.

Picturing Probability

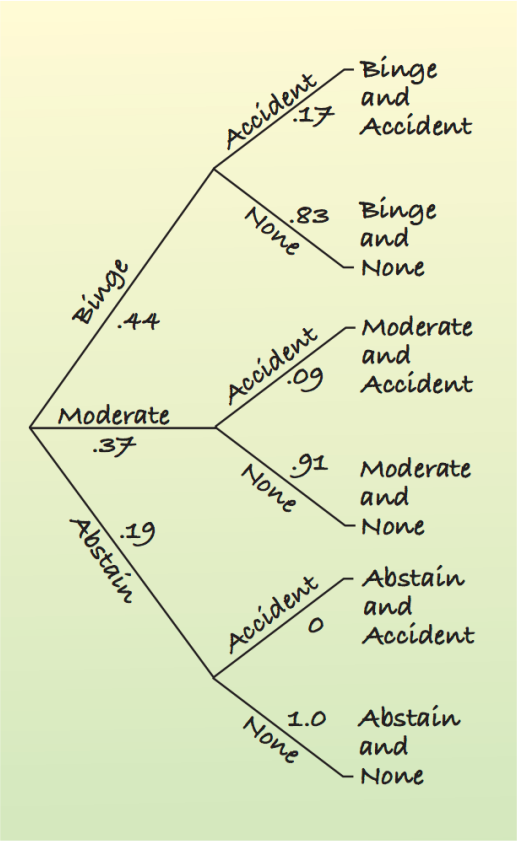
- Tables

Blood Test	Breath Test: Yes	Breath Test: No	Total
Yes	0.22	0.14	0.36
No	0.56	0.08	0.64
Total	0.78	0.22	1.00

- Venn Diagrams



- Trees



Visit https://seeing-theory.brown.edu

and run Ch.1 Chance Event and Ch.2 Conditional Probability

Quiz 10-3

https://forms.gle/1D9uzu5VTrjNqDQK8

Key Points

Chapter 14: Randomness and Probability

- Terms: Trial, outcome/event, sample space ($\mathbf{S}$)
- Law of large numbers (LLN)
- Five basic rules of probability: $0 \leq P(\mathbf{A}) \leq 1$, $P(\boldsymbol{S}) = 1$, $P(\mathbf{A}^C) = 1 - P(\mathbf{A})$,

  $P(\mathbf{A}\ or\ \mathbf{B}) = P(\mathbf{A}) + P(\mathbf{B})$, when A and B are disjoint (or mutually exclusive),

  $P(\mathbf{A}\ and\ \mathbf{B}) = P(\mathbf{A}) \times P(\mathbf{B})$, when A and B are independent.

Chapter 15: Probability rules

- General addition rule: $P(\mathbf{A}\ or\ \mathbf{B}) = P(\mathbf{A}) + P(\mathbf{B}) - P(\mathbf{A}\ and\ \mathbf{B})$
- General multiplication rule: $P(\mathbf{A}\ and\ \mathbf{B}) = P(\mathbf{A}) \times P(\mathbf{B}|\mathbf{A}) = P(\mathbf{B}) \times P(\mathbf{A}|\mathbf{B})$
- Independence: $P(\mathbf{B}|\mathbf{A}) = P(\mathbf{B})$
- Bayes' Rule: $P(\mathbf{B}|\mathbf{A}) = \dfrac{P(\mathbf{A}|\mathbf{B})P(\mathbf{B})}{P(\mathbf{A}|\mathbf{B})P(\mathbf{B}) + P(\mathbf{A}|\mathbf{B^C})P(\mathbf{B^C})}$