MSML610: Advanced Machine Learning

Probability

Instructor: Dr. GP Saggese - `gsaggese@umd.edu`

References:

Probability

- **Probability**
  - Probability definition
  - Probability measure
  - Independent events
  - Conditional probability
  - Law of total probability
  - Bayes theorem
- Random variables
- Mathematical expectation of RVs
- Probability inequalities
- Statistical Inference

Probability definition

- Probability
  - **Probability definition**
  - Probability measure
  - Independent events
  - Conditional probability
  - Law of total probability
  - Bayes theorem
- Random variables
- Mathematical expectation of RVs
- Probability inequalities
- Statistical Inference

What is probability?

- Probability is the mathematical language for quantifying uncertainty
  - It provides a framework for understanding and modeling the likelihood of different outcomes
  - E.g., the probability of flipping a coin and it landing on heads is 0.5
- Probability lets you imagine possible universes and quantify their probability
  - Each universe represents a different possible outcome / scenario
  - Assign probabilities to these universes based on available information or assumptions
  - E.g., in a simple dice game
    - Assuming a fair die
    - There are 6 universes where you roll a 1, 2, . . . , 6 on a die
    - Each universe has a probability of 1/6

Sample outcome and sample space

- **Sample space** Ω is the set of all possible outcomes of a random experiment, e.g.,
  - Toss a coin once: $\Omega = \{H, T\}$
  - Toss a die: $\Omega = \{1, 2, 3, 4, 5, 6\}$
  - Toss a coin twice: $\Omega = \{HH, HT, TH, TT\}$
  - Measure bulb lifetime in hours: real number in $0 \leq 4000$
- **Sample outcome** is the realization of an experiment
  - E.g., toss a coin twice
    - Sample outcomes are $\omega \in \Omega = \{HH, HT, TH, TT\}$
- The same experiment can be represented in different sample spaces Ω
  - E.g., toss a coin twice, possible sample spaces are:
    - $\Omega = \{HH, HT, TH, TT\}$
    - $\Omega = \{\text{Results from first and second toss are the same or not}\}$
    - $\Omega = \{\text{The first toss is H}\}$
    - $\Omega = \{\text{There is only one H, or not}\}$
    - $\Omega = \{\text{Number of heads}\} = \{0, 1, 2\}$

Event

- **Event** is a subset of the sample space Ω
  - Combine outcomes $\omega \in \Omega$ of an experiment that interest us
- **The event A happened** when the outcome ω belongs to $A \subseteq \Omega$, e.g.,
  - Toss a coin twice
  - Sample space: $\Omega = \{TT, TH, HT, HH\}$
  - Event *"the first toss is heads"* is $A = \{HH, HT\}$
- Interesting events
  - Impossible event: $\emptyset$
  - Certain event: Ω
  - Complement of event A: $-A = \Omega - A$
  - Outcomes in A but not in B: $A - B = A \cap (\Omega - B)$

Event space $\mathcal{F}$

- **Event space** $\mathcal{F}$ is the set of all possible events in Ω, i.e., set of subsets of sample space Ω

$$\mathcal{F} \triangleq \mathcal{P}(\Omega) = \{A : A \subseteq \Omega\}$$

- **Sample space vs event space**
  - Sample space Ω contains all possible outcomes of a random experiment
  - Events $\mathcal{F}$ are subsets of the possible outcomes of the experiment
  - "Sample space vs event space" provides flexibility to formulate the problem for better resolution/understanding
    1. Define sample space to include outcomes that already encode interesting events
    2. Define outcomes at maximum granularity (E.g., sample space is $\{HHHH, HHHT, ..., TTTT\}$) and then combine outcomes in events

Summary of definitions

Def	**Symbol**	**Meaning**
Sample outcome	ω	Outcome of a random experiment
Sample space	Ω	All possible outcomes of an experiments
Event	$A \subseteq \Omega$	Combines together sample outcomes
Event space	$\mathcal{F} = \mathcal{P}(\Omega)$	Set of all the possible events

Properties of event space

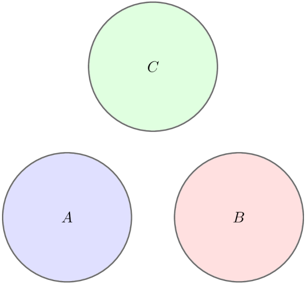
- An event space $\mathcal{F}$ must have these properties to define a probability function:
    1. Impossible event is included: $\emptyset \in \mathcal{F}$
    2. Closed under complement:
       $A \in \mathcal{F} \implies \Omega - A \in \mathcal{F}$
    3. Closed under *finite* union:
       $A_1, ..., A_n \in \mathcal{F} \implies \bigcup_i A_i \in \mathcal{F}$
- **Note**: Using $\mathcal{F} = \mathcal{P}(\Omega)$ ensures these properties

Two mutually exclusive events

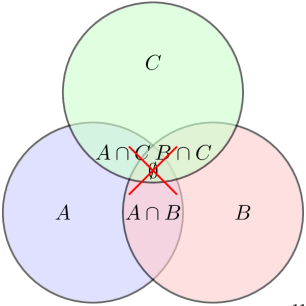
- Aka disjoint, distinct
- Two events A, B are mutually exclusive $\iff A \cap B = \emptyset$
  - The events cannot happen at the same time
  - No outcome can belong to both
- E.g., roll a die
  - Sample space $\Omega = \{1, 2, 3, 4, 5, 6\}$
  - Events $A = \{1, 2\}$ and $B = \{3, 4\}$ are mutually exclusive
  - Events "odd number" and "even number" are mutually exclusive

Mutually vs pairwise exclusive

- Two exclusive events can't happen at the same time
- It's not obvious what it means that *several* events are *exclusive*
- **Pairwise exclusive events:**
  - Events $A_1, A_2, \dots, A_n$ are pairwise exclusive $\iff A_i \cap A_j = \emptyset \; \forall i \neq j$
  - Any pair of events has no intersection
- **Mutually exclusive events:**
  - Events $A_1, A_2, \dots, A_n$ are mutually exclusive $\iff \bigcap_i A_i = \emptyset$
  - All events have no intersection

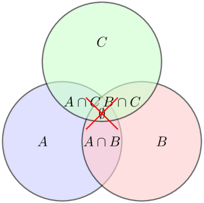


Pairwise exclusive events.

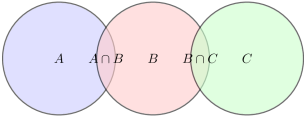


Mutually exclusive as Venn diagrams

- Pairwise exclusive is a stronger property than mutually exclusive:
  - Any pair of events has no intersection (pairwise exclusive) implies all events have no intersection (mutually exclusive)
    - The reverse is not true
  - E.g., 3 events A, B, C can have no common element, but A and B can have a non-empty intersection
- Pairwise exclusive sets mean any event does not overlap with any other event: they are separated
- Mutual, but not pairwise, exclusivity means there is a chain of sets without a single intersection



Mutually but not Pairwise Exclusive Sets



Mutually but not Pairwise Exclusive)

Partition of Ω

- Partition is a sequence of sets $A_1, A_2, \ldots$ such that:
  - Union is Ω: $\cup A_i = \Omega$
  - Pairwise exclusive: $A_i \cap A_j = \emptyset$
- Monotone increasing $\iff A_1 \subseteq A_2 \subseteq \ldots$
  - We define $A_n \to A$ as:
  $$\lim_{n\to\infty} A_n = \cup A_i = A$$
- Monotone decreasing $\iff A_1 \supseteq A_2 \supseteq \ldots$
  - We define $A_n \to A$ as:
  $$\lim_{n\to\infty} A_n = \cap A_i = A$$

Probability measure

- Probability
  - Probability definition
  - **Probability measure**
  - Independent events
  - Conditional probability
  - Law of total probability
  - Bayes theorem
- Random variables
- Mathematical expectation of RVs
- Probability inequalities
- Statistical Inference

Probability measure

- A probability measure is a function:

$$\text{Pr}: \text{ event space } \mathcal{F} \to \mathbb{R}$$

  satisfying the 3 axioms:
  - Probability is non-negative: $\text{Pr}(A) \geq 0$
  - Probability of *the certain event* is 1: $\text{Pr}(\Omega) = 1$
  - Probability of a *finite* union of disjoint events is the sum of their probabilities: if $A_1, ..., A_n$ are pairwise disjoint events, then $\text{Pr}(\cup A_i) = \sum_i \text{Pr}(A_i)$
- Probability is defined:
  - On the event space $\mathcal{F} = \mathcal{P}(\Omega)$
  - Not on the sample space Ω (i.e., set of all realizations of an experiment)
- This ensures the 3 properties of the event space $\mathcal{F}$ hold
- A probability measure associates a probability with each "possible world"

Set operations on events and probability measure

- For any events $A, B \in \mathcal{F}$
  - $\Pr(\emptyset) = 0$
  - $0 \leq \Pr(A) \leq 1$
  - $\Pr(-A) = \Pr(\Omega - A) = 1 - \Pr(A)$
  - $A \subseteq B \implies \Pr(A) \leq \Pr(B)$
  - $A \subseteq B \implies \Pr(B - A) = \Pr(B) - \Pr(A)$
  - $\Pr(A \cap B) \leq \min(\Pr(A), \Pr(B))$
  - $\Pr(A) = \Pr(A \cap B) + \Pr(A \cap -B)$
  - $\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B)$

Union upper / lower bound

- **Union upper bound**

$$\Pr(A \cup B) \leq \Pr(A) + \Pr(B)$$

  - Consequence of:
    - $\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B)$
    - $\Pr(\cdot) \geq 0$
  - Useful to upper bound probability when $\Pr(A \cap B)$ is unknown or hard to compute
- **Union lower bound**

$$\Pr(A \cup B) \geq \max(\Pr(A), \Pr(B))$$

  - From the relationship between probability of union and intersection of A and B
- **Intersection bound**

$$\Pr(A \cap B) \leq \min(\Pr(A), \Pr(B))$$

  - From the relationship between probability of union and intersection of A and B, $\Pr(A \cap B) = \Pr(A) + \Pr(B) - \Pr(A \cup B) \leq \Pr(A) + \Pr(B)$
  - This is the dual of $\Pr(A \cup B) \geq \max(\Pr(A), \Pr(B))$

Independent events

- Probability
  - Probability definition
  - Probability measure
  - **Independent events**
  - Conditional probability
  - Law of total probability
  - Bayes theorem
- Random variables
- Mathematical expectation of RVs
- Probability inequalities
- Statistical Inference

Two independent events

- Two events A and B are independent $\iff$

$$\Pr(A \cap B) = \Pr(A) \cdot \Pr(B)$$

    - In words, the probability of the intersection of independent events is the product of the probabilities of the events
    - Aka "multiplication rule"
- Complement of independent events are independent:

$$\Pr(\neg(A \cup B)) = \Pr(\neg A)\Pr(\neg B)$$

Exclusive vs independent events

- Mutually exclusive (aka disjoint, distinct):

$$\Pr(A \cup B) = \Pr(A) + \Pr(B) \text{ (addition rule)}$$

- Independent:

$$\Pr(A \cap B) = \Pr(A) \cdot \Pr(B) \text{ (multiplication rule)}$$

- Two mutually exclusive A and B with both non-null probability cannot be independent
  - In fact if $\Pr(A) > 0$ and $\Pr(B) > 0$, then $\Pr(A \cap B) = \Pr(\emptyset) = 0 \neq \Pr(A) \cdot \Pr(B)$
  - In words, typical exclusive events are not independent

Set of mutually / pairwise independent events

- A finite set of events $\{A_i : i \in I\}$ is **mutually independent** *iff*

$$\Pr(\cap_k A_k) = \prod_k \Pr(A_k)$$

  - The probability of every subset of events can be factored into the product of the probabilities
  - Mutual independence $\iff$ each event is independent from any intersection of a subset of the remaining events

- A finite set of events $\{A_i : i \in I\}$ is **pairwise independent** *iff*

$$\Pr(A_i \cap A_j) = \Pr(A_i) \Pr(A_j) \; \forall i, j \in I, i \neq j$$

  E.g., a pair of events are independent

- For more than 2 events, mutually independent events $\implies$ pairwise independent events, but the converse is not true
  - Note: this is the opposite relationship with respect to mutually vs pairwise exclusivity since pairwise exclusive is stronger than mutually exclusive

Probabilistic Principle of Inclusion-Exclusion

- Aka PPIE

$$\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B)$$

- **E.g.,**
  - Probability of throwing two dice and getting at least one 6
  - Interpret "union" as "at least one" 6 and use the complement:

$$\Pr(\text{one } 6) = \Pr(\text{at least one } 6) = 1 - \Pr(\text{no } 6)$$

  - Or use PPIE and independence:

$$\begin{aligned}
&\Pr(A = 6 \cup B = 6) \\
&\qquad = \Pr(A = 6) + \Pr(B = 6) - \Pr(A = 6 \cap B = 6) \\
&\qquad = \Pr(6) + \Pr(6) - \Pr(A = 6) \cdot \Pr(B = 6)
\end{aligned}$$

PPIE with N events

- The probability of the union of n events $\Pr(\cup A_i)$ is the sum / subtraction of the probability of intersection of all subsets of events

$$\Pr(\cup A_i) = \sum_{i=1}^{k} (-1)^{k+1} (\sum_{1 < i_1 \leq \ldots < i_k \leq n} \Pr(A_{i_1} \cap \ldots \cap A_{i_k}))$$

- PPIE for 3 events:

$$\begin{aligned} &\Pr(A \cup B \cup C) = \\ &\qquad \Pr(A) + \Pr(B) + \Pr(C) - \\ &\qquad (\Pr(A \cap B) + \Pr(A \cap C) + \Pr(B \cap C)) + \\ &\qquad \Pr(A \cap B \cap C) \end{aligned}$$

Conditional probability

- Probability
  - Probability definition
  - Probability measure
  - Independent events
  - **Conditional probability**
  - Law of total probability
  - Bayes theorem
- Random variables
- Mathematical expectation of RVs
- Probability inequalities
- Statistical Inference

Conditional probability

- Given an event B with $\Pr(B) > 0$, the conditional probability of A given B is:
$$\Pr(A|B) \triangleq \frac{\Pr(A \cap B)}{\Pr(B)}$$
- Conditional probability $\Pr(\cdot|B)$ is a probability:
  - $\Pr(B) > 0$ since conditioning to an event that cannot happen is undefined, like $\frac{0}{0}$
  - It can be proved that it verifies the 3 axioms of a probability measure
    - E.g., $\Pr(A \cap B) \leq \Pr(B)$, so $\Pr(A|B) \leq 1$
  - The rules of probability apply to events left of the bar, but not right of the bar, $\Pr(X|\cdot)$
    - E.g., $\Pr(X|A \cup B) \neq \Pr(X|A) + \Pr(X|B)$

Conditional probability: intuition

- $\Pr(A|B)$ is the probability of A when B has happened
  - I.e., the fraction of times that A happens when B has already happened
  - It changes the sample space Ω to reflect a world where B has happened, so we normalize by $\Pr(B)$
- E.g., if the probability of it raining today A is 10%, given that it's cloudy B, and clouds appear in 50% of the days, then $\Pr(A|B) = \frac{\Pr(A \cap B)}{\Pr(B)}$
- Sometimes it is easier to compute a probability in a world where B has happened
  - E.g., if you know a card drawn is a king, the probability it is a heart is $1/4$
- Conditional probability refines predictions with new evidence

Conditional probability: example

What is the probability of getting 1 from a die, given that the die yielded an odd number?

Solution

1. Computing the probability directly in the new world where the event "die is odd", i.e., in a different sample space Ω
   - We get that there are 3 possible outcomes, equally probable, so $\frac{1}{3}$
2. Using the definition of conditional probability, without changing the sample space

$$\Pr(X = 1 | X \text{ is odd}) \triangleq \frac{\Pr(X = 1 \wedge X \text{ is odd})}{\Pr(X \text{ is odd})} = \frac{1}{6} \Big/ \frac{1}{2} = \frac{1}{3}$$

Conditional probability: example

- The probability that “it is Friday and that a student is absent” is 0.03
- Today is Friday: what is the probability that the student is absent?

Solution

- The answer is not $\Pr(A \cap B)$ since that is the probability of both events happening at the same time, while we know that one event has already happened
- $A =$ “Friday” and $B =$ “student is absent”
- We want to know $\Pr(B|A) = \frac{\Pr(B \cap A)}{\Pr(A)}$ and we know that $\Pr(A) = 1/7$ and $\Pr(A \cap B) = 0.03$

Probability of the intersection of two non-independent events

- It holds:

$$\Pr(A \cap B) = \Pr(A|B) \cdot \Pr(B) = \Pr(B|A) \cdot \Pr(A)$$

- This is useful to compute the probability of the intersection event when A and B are not independent
- If they are independent the probability is factored in the product

$$\Pr(A \cap B) = \Pr(A) \cdot \Pr(B)$$

Conditional probability: example marbles

- In a bag there are 2 blue marbles and 3 red marbles
- What is the chance of drawing 2 marbles and both are blue?

Solution

1. Using conditional probability:

   $$\text{Pr(1 B and 1 B)} = \text{Pr(1 B}|\text{bag)}\cdot\text{Pr(1 B}|\text{bag - 1 B)} = 2/5\times 1/4 = 2/20 = 1/1$$

2. By counting:
   - There are $5!/(3! \times 2!) = 10$ possible ways of picking 2 marbles from a set of 2 blue and 3 red marbles (Mississippi formula)
   - We are interested in only one
3. Distinguish the marbles as all different (i.e., B1, B2, R1, R2, R3)
   - There are 20 possibles ways of picking the marbles (5×4) and we are interested in two permutations

Prosecutor's fallacy

- The prosecutor's fallacy represents that:

$$\Pr(A|B) \neq \Pr(B|A)$$

Problem

- There is a medical test for a disease D which has outcomes $+$ and $-$
- The test is fairly accurate and has $\Pr(+|D)$ (true positive rate) $=$ $\Pr(-|\overline{D})$ (true negative rate) $= 90\%$
- You take the test and get a positive: what's the probability that you have the disease?

Solution

- You want to know $\Pr(D|+)$ and not $\Pr(+|D)$ which is 90%
- In fact $\Pr(D|+)$ depends (according to Bayes' theorem) on $\Pr(+|D) = 90\%$, but also on $\Pr(D)$ (how likely is the disease) and $\Pr(+)$ (how often the machine report a positive)
- E.g., if the disease is vanishingly rare $\Pr(D) \to 0$

Prosecutor's fallacy: example

- The probability that a person is Argentinian being the Pope, is not the probability that a person is the Pope being Argentinian
- Numerically

$$\Pr(\text{x is Pope}|\text{x is from Argentina}) = \frac{\Pr(\text{x is Pope} \wedge \text{x is from Argentina})}{\Pr(\text{x is from Argentina})} = \frac{1}{47,000,000}$$

$$\Pr(\text{x is from Argentina}|\text{x is Pope}) = \frac{\Pr(\text{x is from Argentina} \wedge \text{x is Pope})}{\Pr(\text{x is Pope})} = \frac{1}{1} = 1$$

Independent events and conditional probability

- A and B are independent $\iff \Pr(A|B) = \Pr(A)$
- In words, knowing that B happened does not change the probability of A: that's why the events are said to be "independent"
- The (less intuitive) definition $\Pr(A \cap B) = \Pr(A) \cdot \Pr(B)$ is equivalent to this property
- A and B are independent *iff*

$$\Pr(A|B) = \Pr(A|\neg B) \wedge \Pr(B|A) = \Pr(B|\neg A)$$

- A and B are dependent *iff*

$$\Pr(A|B) \neq \Pr(A|\neg B) \vee \Pr(B|A) \neq \Pr(B|\neg A)$$

- Dependent events can be contemporaneous or not, e.g.,
  - 2 horses winning the same race
  - 2 marbles picked from the same bag one after another
- When we say "at least one of coin flip is ..." introduces a dependency between events

Odds: definition

- Given an event A with a success probability p (i.e., modelled as a Bernoulli), the odds of A are defined as

$$\frac{p}{1-p}$$

- Odds are the ratio between payoff when winning and losing

$$\text{odds} = \frac{\text{lose payoff}}{\text{win payoff}}$$

  that makes the game fair

- E.g.,
  - A game is fair when the expected earnings from playing it are equal to 0
  - If odds are $1/3$ one should be paid 3 times more when winning than when losing
  - When odds are < 1 then "the odds are against you"

Interpretation of odds

- Consider a game where:
  - You flip a coin with probability p of head
  - If it comes up heads you win $X > 0$, otherwise you lose $Y > 0$
- What is the value of X and Y for the game to be fair?

Solution

- The game is fair when expected earnings are 0:

$$\mathbb{E}[\text{earnings}] = pX - (1-p)Y = 0 \implies \frac{Y}{X} = \frac{p}{1-p} = \text{odds}$$

- Thus

$$Y = \frac{p}{1-p}X = \text{odds} \cdot X$$

- The odds indicate how many times one should be paid in case of losing in a game with probability p
  - If $p = 0.5$, then odds $= 1$, so one should be paid the same as winning since the game is fair
  - If $p > 0.5$, since odds $= \frac{1}{1/p-1}$, then $1/p < 2$, $(1/p - 1) < 1$, and odds > 1; $Y = \text{odds} \cdot X > X$, i.e., you need to be paid more if you lose than if you win

Law of total probability

- Probability
  - Probability definition
  - Probability measure
  - Independent events
  - Conditional probability
  - **Law of total probability**
  - Bayes theorem
- Random variables
- Mathematical expectation of RVs
- Probability inequalities
- Statistical Inference

Law of total probability

- Let $A_1, \dots, A_k$ be a partition of the sample space Ω with $\Pr(A_i) > 0$
- For any event B:

$$\Pr(B) = \sum_{i=1}^{k} \Pr(B|A_i)\Pr(A_i)$$

- It expresses the probability of an event using conditional probabilities of events partitioning the event space Ω
  - Computing conditional probabilities can be easier as our perspective changes when conditioning on an event

Proof

- Define $C_i = B \cap A_i$
- All C_i are disjoint and their union is B (i.e., C_i are a partition of B)
- By the axiom about the probability of the union of disjoint events and by the definition of conditional probability:

$$\Pr(B) = \Pr(\cup_i (B \cap A_i)) = \sum_i \Pr(B \cap A_i) = \sum_i \Pr(B|A_i) \cdot \Pr(A_i)$$

Law of total probability for two events

- For any event B with $\Pr(B) > 0$

$$\begin{aligned}\Pr(A) &= \Pr(A|B)\Pr(B) + \Pr(A|\overline{B})\Pr(\overline{B}) \\ &= \Pr(A|B)\Pr(B) + \Pr(A|\overline{B})(1 - \Pr(B))\end{aligned}$$

- So one needs 3 quantities to compute $\Pr(A)$
  - $\Pr(A|B)$
  - $\Pr(A|\overline{B})$
  - $\Pr(B)$

Bayes theorem

- Probability
  - Probability definition
  - Probability measure
  - Independent events
  - Conditional probability
  - Law of total probability
  - **Bayes theorem**
- Random variables
- Mathematical expectation of RVs
- Probability inequalities
- Statistical Inference

Bayes' theorem for 2 events

- If $\Pr(A) > 0, \Pr(B) > 0$, then

$$\Pr(A|B) = \frac{\Pr(B|A) \cdot \Pr(A)}{\Pr(B)}$$

- MEM: one multiplies by $\Pr(A)/\Pr(B)$ as in the original formula
- Bayes' theorem for 2 events allows to invert the conditioning of events
- Using the Law of total probability, we can express everything in terms of the same conditional probabilities:

$$\Pr(A|B) = \frac{\Pr(B|A) \cdot \Pr(A)}{\Pr(B|A) \cdot \Pr(A) + \Pr(B|\overline{A}) \cdot \Pr(\overline{A})}$$

General form of Bayes' theorem

- Assume (same hypothesis of law of total probability)
  - $A_1, ..., A_k$ be a partition of sample space Ω
  - $\Pr(A_i) > 0$
  - $\Pr(B) > 0$
- Then:

$$\Pr(A_i|B) = \frac{\Pr(B|A_i) \cdot \Pr(A_i)}{\Pr(B)}$$

- Bayes' theorem:
  - Computes the probability of different events A_i partitioning Ω
  - After a given event B has happened
  - In terms of the inverted conditioned probabilities $B|A_i$

Interpretation of Bayes' theorem as update of beliefs

- Bayes' theorem states:

$$\Pr(A_i|B) = \frac{\Pr(B|A_i) \cdot \Pr(A_i)}{\Pr(B)}$$

  where:
    - $\Pr(A_i|B)$ = posterior probability of A_i
    - $\Pr(B|A_i)$ = conditional (inverted) probability
    - $\Pr(A_i)$ = prior probability of A_i
    - $\Pr(B)$ = probability of B
- Bayes' theorem expresses the posterior probability of each event A_i using:
    - The conditional probabilities of $B|A_i$ (known or estimated)
    - The prior probability of A_i, i.e., the probability before event B
    - The probability of the "updating event" B
- In other words, the events of interest are:
    - A_i, for which we have prior probabilities
    - Then an event B occurs
    - Using Bayes' theorem, we update our belief about A_i after B occurs

Bayes' theorem + Law of total probability

- Under the same hypothesis of Bayes' theorem:

$$\Pr(A_i|B) = \frac{\Pr(B|A_i) \cdot \Pr(A_i)}{\sum_{j=1}^{k} \Pr(B|A_j) \cdot \Pr(A_j)}$$

  where:
  - $\Pr(A_i|B)$ = posterior probability of A_i
  - $\Pr(B|A_i)$ = conditional (inverted) probability
  - $\Pr(A_i)$ = prior probability of A_i
- All conditional probabilities on the RHS are of the same type and inverted

Bayes' theorem vs Law of total probability

- Both Bayes' theorem and Law of total probability use:
  - An event
  - A partitioning of the sample space
- Law of total probability
  - The probability of the given event A is expressed in terms of the probabilities of the partitioning events B_i
- Bayes' theorem
  - The probabilities of the partitioning events A_i are updated given an event B

Making decisions using Bayes' theorem

- Bayes' theorem is used to make decisions (e.g., choose among outcomes) using an event and prior information
- Important points:
  - Bayes' theorem calculates the probability of an event based on prior knowledge of conditions related to the event
  - It applies to various fields, such as finance, healthcare, and machine learning, where decision-making under uncertainty is required
  - The formula for Bayes' theorem is:

$$\Pr(A|B) = \frac{\Pr(B|A) \cdot \Pr(A)}{\Pr(B)}$$

- Examples:
  - In a medical diagnosis, Bayes' theorem can be used to determine the probability of a disease given a positive test result, considering the overall prevalence of the disease and the accuracy of the test
  - In spam filtering, Bayes' theorem helps decide the probability that an incoming email is spam based on features like word frequency
  - In weather forecasting, it can update the probability of rain based on new data
  - In financial markets, it can assist in estimating the probability of stock

Bayes' theorem: detect spam email

- We want to partition emails in 3 categories:
  - A_1=`spam`, A_2=`low priority`, A_3=`high priority`
- Receive an email with the word `free`: what is the probability that it is spam?

Solution

- Prior knowledge (e.g., from a large corpus of emails):
  - $\Pr(A_1) = .7$, $\Pr(A_2) = .2$, $\Pr(A_3) = .1$
  - Sum is 1 since the events are a partition of the sample space
- B is the event "email contains the word free"
- From previous experience (e.g., a large corpus of emails) we know:
  - $\Pr(B|A_1) = .9$, $\Pr(B|A_2) = .01$, $\Pr(B|A_3) = .01$
  - Sum is not 1 since $\Pr(B|.)$ is not a probability function
- Using Bayes' theorem, compute the probability of the event

$$A_1|B = \text{"email is spam, given that it contains free"}$$

$$\Pr(A_1|B) = \frac{\Pr(B|A_1) \cdot \Pr(A_1)}{\Pr(B|A_1) \cdot \Pr(A_1) + \Pr(B|A_2) \cdot \Pr(A_2) + \Pr(B|A_3) \cdot \Pr(A_3)}$$

Bayes' theorem: rain example

- Consider the events:
  - W = "weatherman predicts rain"
  - R = "it rains"
- A weatherman:
  - Predicts rain correctly 90% of the time when it rains
  - Predicts rain incorrectly 10% of the time when it does not rain
- Given it rains 5 days a year, what's the probability it rains tomorrow given the weatherman predicts rain?

Solution

- Assume
  - $\Pr(W|R) = 90\%$
  - $\Pr(W|\overline{R}) = 10\%$
  - $\Pr(R) = 5/365$
  - $\Pr(W|R) = .9$
  - $\Pr(W|\overline{R}) = .1$
- Using Bayes:

$$\Pr(R|W) = \frac{\Pr(W|R)\Pr(R)}{\Pr(W|R)\Pr(R) + \Pr(W|\overline{R})\Pr(\overline{R})} = 0.9{*}5/(0.9{*}5+0.1{*}360) = $$

Frequentist interpretation

- Probabilities can be interpreted as the outcome of long-run experiments
- This is a way to empirically estimate probabilities, e.g.,
  - Q: What is the probability of a car tire exploding when filled 50% beyond the manufacturer's recommendation?
  - A: Fill 100 tires and see how many explode
- There is a problem for one-time events, e.g.,
  - Q: What is the probability of life on Mars?
    - The true probability is 0 or 1, depending on life being on Mars or not
    - You can use scientific knowledge to estimate it
    - If we go to Mars and find life, then the probability is 1
    - Proving the probability is 0 is more difficult: you need to check everywhere on Mars

Bayesian interpretation

- Probabilities measure individual uncertainty about events
  - Knowledge of the world as a one-time event
- Probabilities quantify uncertainty and extend logic to uncertain statements
  - Uncertainty is common in the real world
  - E.g., there is noise, we make mistakes, we don't understand
- Bayesian statistics is:
  - a procedure to make statements using probabilities
  - an extension of true-false logic when dealing with uncertainty

Random variables

Random variables

- Probability
- Random variables
    - **Random variables**
    - CDF, PMF, PDF of Random Variables
    - Joint distributions
    - Marginal distributions
    - Independent RVs
    - Conditional PDF RVs
- Mathematical expectation of RVs
- Probability inequalities
- Statistical Inference

Random variable (RV)

- A RV X is a function defined from the sample space to real numbers:

$$X : \Omega \to \mathbb{R}$$

- A RV is concerned with outcomes of an experiment (i.e., sample space), and not with events
- Events are subsets of the sample space $A \subseteq \Omega$ that are transformed by X into subsets of $A \subseteq \mathbb{R}$, and vice versa
- A random variable can be:
  - Discrete: takes finite *or* countably infinite values
  - Continuous: takes uncountably infinite values

Random variables link sample space and data

- Sample space Ω is a set, but to process data we need numbers
  - Random variable is the link between Ω and numbers
- A relation for a RV $X \in A_0 \subseteq \mathbb{R}$ (e.g., $X = 1$ or $X \geq 1$) corresponds to an event $X^{-1}(A_0) \subseteq \Omega$
  - For simplicity, we refer to $X \in A_0$ as "event A_0" since the relation induces an event
  - Once the RV is introduced, it is like the sample space disappears and the outcome of an experiment is just a number
- A RV can have a simple conceptual description of an experiment, which can help understanding and reason about the problem, e.g.,
  - "$X =$ number of heads when tossing 5 fair coins"
  - $X \neq 0$ is the event "there are no heads tossing 5 fair coins"

RVs are not defined in a unique way

- Different RVs X can be associated with the same sample space Ω
- RVs introduce degrees of freedom in describing and solving a problem, just like a sample space
- E.g., for 2 coin tosses:
    - X as "binary representation" $=$ $\{$HH: 0, HT: 1, TH: 2, TT: 3$\}$
    - X as "number of tails" $=$ $\{$HH: 0, HT: 1, TH: 1, TT: 2$\}$
    - X as "number of heads" $=$ $\{$HH: 2, HT: 1, TH: 1, TT: 0$\}$
    - X as "two coin tosses are equal" $=$ $\{$HH: 1, HT: 0, TH: 0, TT: 1$\}$
- Different outcomes of the experiment can be associated with the same number
    - E.g., we might want to associate distinct numbers to interesting events
    - E.g., for 2 coin tosses: $\{$HH: 0, HT: 1, TH: 1, TT: 2$\}$

CDF, PMF, PDF of Random Variables

- Probability
- Random variables
  - Random variables
  - **CDF, PMF, PDF of Random Variables**
  - Joint distributions
  - Marginal distributions
  - Independent RVs
  - Conditional PDF RVs
- Mathematical expectation of RVs
- Probability inequalities
- Statistical Inference

Cumulative Distribution Function of a RV

- The **Cumulative Distribution Function (CDF)** of a *continuous* or *discrete* RV X is defined as:

$$F_X(x) \triangleq \Pr(X \leq x) \text{ for } x \in \mathbb{R}$$

- Why it is useful:
    - CDF combines RV X (to infer events) together with $\Pr(\cdot)$ into a function $\mathbb{R} \to [0, 1]$
    - CDF is a way to infer "standard" events using the total ordering in $\mathbb{R}$

Properties of CDF

1. Limits:

$$\lim_{x \to -\infty} F(x) = 0$$

$$\lim_{x \to +\infty} F(x) = 1$$

2. Not decreasing:

$$x_1 < x_2 \implies F(x_1) \leq F(x_2)$$

3. Continuous from the right:

$$\lim_{\varepsilon \to 0^+} F(x + \varepsilon) = F(x) \; \forall x$$

CDF of discrete RV in terms of probability

- The CDF of a discrete RV evaluated in one point x is the sum of probability of all outcomes $u \leq x$:

$$F_X(x) \triangleq \Pr(X \leq x) = \sum_{u \leq x} \Pr(X = u)$$

- E.g.,
  - Experiment: Toss a fair coin twice
  - RV $X =$ "count the number of tails"
  - X is $\{HH : 0, HT : 1, TH : 1, TT : 2\}$
  - $F_X(x)$ is $\{0 : 1/4, 1 : 3/4, 2 : 1\}$
- Plot of $F_X(x)$ for a discrete RV:
  - Is a staircase function
  - The jump at x_i is equal to $\Pr(x_i)$
  - The step has the same value on the right
  - Is monotonically increasing function

Probability Mass Function for a discrete RV

- The **Probability Mass Function (PMF)** of a discrete RV X is a function $f_X(x)$ such that:

$$f_X(x) = \begin{cases} \Pr(X = x_i) & x_i \in \{x_1, \ldots, x_n\} \\ 0 & \text{otherwise} \end{cases}$$

- A finite PMF can always be represented with a table
  - E.g., for 2 coin tosses {HH: 0, HT: 1, TH: 1, TT: 2}, and PMF is {0: 1/4, 1: 1/2, 2: 1/4}

Properties of PMF

1. Integral is 1: $\sum_{x_i} f_X(x) = 1$
2. Always non-negative: $f_X(x) \geq 0$

- These properties descend from the properties of CDF
- CDF of discrete RV in terms of PMF

$$F_X(x) = \Pr(X \leq x) = \sum_{u \leq x} f_X(u)$$

where $f_X(u)$ is the PMF of X

PMF: example of coin flip

- X represents the outcome of a coin flip (Bernoulli)
- $X = 0$ represents tails and $X = 1$ represents heads, with a given probability p:

$$f_X(x) = \begin{cases} p & \mathrm{X} = 1 \\ 1 - p & \mathrm{X} = 0 \end{cases}$$

- The PMF of X can be written as one-line:

$$f_X(x) = p^x (1 - p)^{(1-x)}$$

with $x = 0, 1$

Discrete RV in terms of continuous RV

- A discrete RV $f_X(x)$ can be expressed a continuous RV $f_X^*(x)$ with Dirac delta impulses in its PDF
  - Otherwise the probability of a single event would be 0 as in a continuous RV

$$f_X^*(x) = \sum_{i=1}^{n} f_X(x_i)\delta(x - x_i)$$

- With this definition all formulas for continuous RV apply to a discrete RV
  - E.g., CDF is just the integral of the PDF and it has jumps in the deltas

Integrals in terms of PDF and CDF

- Any integral involving a PDF in the form (e.g., same form of theorem of mean):

$$\int g(x) f_X(x) dx$$

can be rewritten in terms of CDF:

$$\int g(x) dF_X$$

- This is because of relationship between CDF and PDF in terms of derivative

$$dF_X = \frac{df_X}{dx}$$

Empirical CDF of a (discrete or continuous) RV

- Consider
  - X (discrete or continuous) with a certain CDF $F(x)$
  - Take IID samples $X_1, ..., X_n$ from X
- The empirical CDF is defined as:

$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} I(X_i \leq x)$$

- Note that X_i can be
  - A RV, then $\hat{F}_n(x)$ is the empirical CDF RV
  - A realization of a RV x_i, then $\hat{F}_n(x)$ is a sample realization of the empirical CDF RV

Empirical PMF

- The empirical PMF is defined as:

$$\hat{f}_n(x) \triangleq \frac{d\hat{F}_n(x)}{dx}$$

- Using the definition of empirical CDF:

$$\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \leq x)$$

- Then

$$\hat{f}_n(x) = \frac{d}{dx}\frac{1}{n}\sum_i I(X_i \leq x) = \frac{1}{n}\sum_i \frac{d}{dx} I(X_i \leq x) = \frac{1}{n}\sum_i \delta(x - x_i)$$

Empirical PMF of a (discrete or continuous) RV

- We can represent the PMF using n samples, summing on the m support points x_i:

$$f_X^*(x) = \frac{1}{n}\sum_{i=1}^{m} f_X(x_i)\delta(x - x_i)$$

  or in terms of the n samples x_j where multiple samples are accounted one at a time:

$$f_X^*(x) = \frac{1}{n}\sum_{j=1}^{n} \delta(x - x_j)$$

Integral of empirical CDF / PMF

- Given a relationship like

$$y(x) = \int g(x) d\hat{F}_n = \int g(x) \hat{f}_n(x) dx$$

using the expression for the empirical PMF:

$$\hat{f}_n(x) = \frac{1}{n} \sum_{i=1}^{n} \delta(x - x_i)$$

we get:

$$y(x) = \int g(x) \frac{1}{n} \sum_i \delta(x - x_i) dx = \frac{1}{n} \sum_i \int g(x) \delta(x - x_i) dx = \frac{1}{n} \sum_i g(x_i)$$

Probability Density Function for continuous RV var

- The **Probability Density Function (PDF)** is defined as

$$f_X(x) = \left.\frac{dF_X(u)}{du}\right|_{u=x}$$

- The PDF is defined in all points where $F_X(x)$ is continuous and thus derivable

2 properties of PDF

- From the axioms of probability it follows:
  1. Integral is 1: $\int_{-\infty}^{+\infty} f_X(x)dx = 1$
  2. Always non-negative: $f_X(u) \geq 0$

CDF for continuous RV in terms of PDF

- The CDF is defined as:

$$F_X(x) = \Pr(X \leq x) = \int_{-\infty}^{x} f_X(u)du$$

where $f_X(u)$ is the PDF of X

Probability of event in terms of CDF / PDF of continuous RV

- For an event $A = [a, b] \subseteq \mathbb{R}$:

$$\Pr(a \leq X \leq b) = \Pr(a < X < b) = F_X(b) - F_X(a) = \int_a^b f_X(u)du$$

- For a generic event $A \subseteq \mathbb{R}$ which corresponds to a subset of $X^{-1}(A) = A_X \subseteq \Omega_X$:

$$\Pr(A_X) = \Pr(X \in A) = \Pr(A) = \int_A f_X(u)du$$

Probability of a single value

- The probability that a continuous RV takes any particular value $\Pr(X = a)$ is 0
- This is different from a discrete RV

Joint distributions

Joint CDF for 2 RV: definition

- The joint CDF of two RVs X and Y is defined as:

$$F_{X,Y}(x, y) = \Pr(X \leq x, Y \leq y)$$

Joint CDF: intuition

- It is the same as the probability of two events to happen jointly, but using the events induced by X and Y onto $\mathbb{R}$

Joint CDF for discrete RV in terms of probability

- The joint CDF of two discrete RVs X and Y can be expressed as:

$$F_{X,Y}(x,y) = \sum_{u \leq x, v \leq y} \Pr(X = u, Y = v)$$

Joint PMF for discrete RV in terms of probability

- The joint PMF of 2 discrete RVs:

$$f_{X,Y}(x,y) = \begin{cases} \Pr(X = x, Y = y) & x \in \{x_1, ..., x_n\}, y \in \{y_1, ..., y_m\} \\ 0 & \text{otherwise} \end{cases}$$

- Note that a joint PMF can be represented with a bidimensional table

Joint PMF properties

- The joint PMF has the properties:
  1. Always non-negative: $f_{X,Y}(x,y) \geq 0$
  2. Integral is 1: $\sum_x \sum_y f_{X,Y}(x,y) = 1$

Joint PDF for continuous RV in terms of joint CDF

- The joint PDF of 2 RVs X and Y

$$f_{X,Y}(x,y) = \frac{\partial^2 F_{X,Y}(u,v)}{\partial x \partial y}$$

- The joint PDF has all the properties of a PDF (always positive, integral is 1)

Joint CDF for discrete RV in terms of joint PDF

- One can get a joint CDF by integrating the joint PDF:

$$F_{X,Y}(x, y) = \sum_{u \leq x, v \leq y} f_{X,Y}(u, v)$$

Joint CDF for continuous RV in terms of joint PDF

$$F_{X,Y}(x,y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f_{X,Y}(u,v) du dv$$

- This comes through the definition of joint PDF as partial derivative

Probability of event in terms of joint PDF for continuous RV

- Given a joint PDF $f_{X,Y}(x,y)$, the probability of event in sample space $A \subseteq \Omega_X \times \Omega_Y$, i.e., $S = X(\Omega_X) \times Y(\Omega_Y) \subseteq \mathbb{R}^2$

$$\Pr(A) = \int_S f_{X,Y}(u,v)dudv$$

- This is a key relationship for marginal and conditional PDFs

Marginal distributions

- Probability
- Random variables
  - Random variables
  - CDF, PMF, PDF of Random Variables
  - Joint distributions
  - **Marginal distributions**
  - Independent RVs
  - Conditional PDF RVs
- Mathematical expectation of RVs
- Probability inequalities
- Statistical Inference

Marginal CDF for discrete / continuous RV in terms of joint CDF

- One can get a marginal CDF from the joint CDF by setting variables to $+\infty$

$$F_X(x) \triangleq \Pr(X \leq x) = \Pr(X \leq x, Y \leq \infty) \triangleq F_{X,Y}(x, \infty)$$

Marginal PMF for discrete RV in terms of joint PMF

- Given two discrete RVs X (defined on Ω_X) and Y (defined on Ω_Y) one can get a marginal PMF from the joint PMF through summing on one variable
- The marginal PMF of X is defined as:

$$
\begin{aligned}
f_X(x) &\triangleq \Pr(X = x) \\
&= \Pr(X = x, Y \in \mathbb{R}) \\
&= \sum_{y_i \in Y(\Omega_Y)} \Pr(X = x, Y = y_i) \text{ (since all events } X = x \wedge Y = y_i \text{ are dist} \\
&= \sum_{y_i \in Y(\Omega_Y)} f_{X,Y}(x, y_i)
\end{aligned}
$$

- Note that $f_X(x)$ is a PMF with all the associated properties

Marginal PDF for continuous RV in terms of joint PDF

- Given two continuous RVs X and Y, one can get a marginal PDF from the joint PDF through integrating on one variable

$$
\begin{aligned}
f_X(x) &\triangleq \frac{dF_X(u)}{du} = \frac{d}{du}\Pr(X \le x) = \frac{d}{du}\Pr(X \le x, Y \le +\infty) \\
&= \frac{d}{du}\int_{v=-\infty}^{u}\int_{y=-\infty}^{\infty} f_{X,Y}(v,y)dvdy \\
&= \int_{y=-\infty}^{\infty} f_{X,Y}(x,y)dy
\end{aligned}
$$

Independent RVs

- Probability
- Random variables
  - Random variables
  - CDF, PMF, PDF of Random Variables
  - Joint distributions
  - Marginal distributions
  - **Independent RVs**
  - Conditional PDF RVs
- Mathematical expectation of RVs
- Probability inequalities
- Statistical Inference

Def of independent RV

- The RVs X and Y are independent $\iff$ the events $X \leq x$ and $Y \leq y$ are independent for all x and y

$$\Pr(X \leq x, Y \leq y) = \Pr(X \leq x) \cdot \Pr(Y \leq y)$$

- This is equivalent to events $X \in A$ and $Y \in B$ being independent

CDF of independent RV

- If RV X and Y are independent their joint CDF can be factored into the product of marginal CDF:

  $$F_{X,Y}(x,y) \triangleq \Pr(X \leq x, Y \leq y) = \Pr(X \leq x) \cdot \Pr(Y \leq y) \triangleq F_X(x) F_Y(y) \ \forall x,$$

- Also the converse it true

PDF / PMF of independent RVs

- If RV X and Y are independent their joint PDF (or PMF) can be factored into the product of marginal PDF (or PMF):

$$f_{X,Y}(x,y) \triangleq \frac{\partial F_{X,Y}(x,y)}{\partial x \partial y} = \frac{\partial F_X(x) F_Y(y)}{\partial x \partial y} = \frac{\partial F_X(x)}{\partial x} \cdot \frac{\partial F_Y(y)}{\partial y} = f_X(x) \cdot f_Y(y)$$

- Also the converse it true

Characterization of PDF and CDF of independent RV

- RV X and Y are independent $\iff$ their joint PDF / PMF / CDF factors in terms of marginal PDF / PMF / CDF

Marginal PDF / PMF / CDF

- It refers to a distribution of a single RV in a set-up where multiple RVs exist
- E.g., the marginal PDF X is the joint PDF of X and Y integrated over Y so when we talk about marginal we refer to a single RV

Conditional PDF RVs

Def of conditional PDF for RVs

- X and Y are RVs, the conditional PDF of X given Y is defined as:

$$f_{X|Y}(x, y) \triangleq \frac{f_{X,Y}(x, y)}{f_Y(y)}$$

- A similar definition holds for PMFs
- MEM: A PDF / PMF is like a probability of events, so definition of conditional prob extends to PDF / PMF in the same way

Conditional probability in terms of conditional PDF

$$\Pr(X \in A | Y = y) = \int_{X \in A} f_{X|Y}(x, y) dx$$

- Note that - The conditional probability is a function of y - $\Pr(X \in A|Y = y)$ is intended as the limit $dy \to 0$ of $\Pr(X \in A|y \leq Y \leq y + dy)$ since the event $Y = y$ has probability 0

  - It can be proved by writing conditional prob in terms of its definition

Marginal PDF for continuous RV in terms of conditional PDF

- Write PDF of X in terms of joint PDF of X and Y

$$f_X(x) = \int_{y=-\infty}^{+\infty} f_{X,Y}(x,y)dy$$

- Then express the joint PDF in terms of conditional PDF:

$$f_X(x) = \int_{y=-\infty}^{+\infty} f_{X|Y}(x,y)f_Y(y)dy$$

- This is similar to the law of total probability, since we express a probability as summation of the conditional probability multiplied by the probability that we are conditioning on

Summary of relationships

- $\Pr(x) \rightarrow$ CDF
- PDF $= \frac{d}{dx}$ CDF
- PDF / PMF is close to prob (it is like a prob density)
- $\Pr() = \int$ PDF
- CDF $= \int_{-\infty}^{x}$ PDF
- Marginal prob $= \int$ joint PDF
- Marginal prob $= \int$ cond PDF $\times$ marginal PDF (like "law of total probability")

Mathematical expectation of RVs

- Probability
- Random variables
- **Mathematical expectation of RVs**
  - Mean
  - Variance and covariance
  - Statistics of RVs
- Probability inequalities
- Statistical Inference

Mean

- Probability
- Random variables
- Mathematical expectation of RVs
    - **Mean**
    - Variance and covariance
    - Statistics of RVs
- Probability inequalities
- Statistical Inference

Mean of discrete RV: definition

- The mean of a discrete RV is defined as:

$$\mathbb{E}[X] \triangleq \sum_i x_i f_X(x_i)$$

- We use the PMF in the definition since in this way it is more similar to the definition of mean for continuous RVs
- MEM:
  1. Sum of each value multiplied by its prob
  2. Weighted average of the values of a PDF
  3. Dot product of a vector of values and prob of values
- If X can take countably infinite values, then the infinite series should converge in absolute value

Mean of discrete RV in terms of probability

$$\mathbb{E}[X] = \sum_i x_i \Pr(X = x_i)$$

Alternative names and symbols for mean of a RV

- The mean is also called:
    - Mathematical expectation
    - Expectation
    - Expected value
    - First moment of a RV
- It is indicated as μ_X or $\mathbb{E}[X]$
- Note that the average of values (e.g., sample mean) is indicated as $\overline{X}$

What is the mean of a biased coin?

- $\Pr(X = 1) = p$ and $\Pr(X = 0) = 1 - p$, then $\mathbb{E}[X] = 0 \times (1 - p) + 1 \times p = p$
- Note that the mean of the RV can be a value that the RV cannot assume

Mean of continuous RV: definition

- The mean is defined as:

$$\mathbb{E}[X] \triangleq \int_{-\infty}^{+\infty} x f_X(x) dx$$

- The mean is well-defined if the integral converges in absolute value

Mean as measure of central tendency

- Draw many IID samples from a RV X
- Compute the (sample) average $\overline{X}_n = \frac{X_1 + \ldots + X_n}{n}$
- Then $\overline{X}_n$ will approximate the mean of X:

$$\lim_{n \to \infty} \overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} = \frac{X_1 + X_2 + \ldots + X_n}{n} \approx \mathbb{E}[X]$$

- Thus the mean can be interpreted as the average value of the RV using infinite IID samples
- This is the Law of Large Numbers (LNN)

Intuition of law of large numbers for discrete vars

- In the case of discrete RV:
  - We take an infinite number of samples for X and sum them
  - Group the samples by values
  - The average is the sum of the values for X multiplied by the probability (since the frequency converges to the probability), which is the def of mean
- Note that the mean is not the most frequent value (that's the mode)

Mean as center of mass

- The mean can be interpreted as the center of mass of the PDF / PMF, i.e., where one needs to put a wedge to “balance” the PDF / PMF

Mean as minimum value for squared errors

- The mean is the value y that minimizes the quantity $\sum_i (x_i - y)^2$ on an infinite number of trials
- This can be proved either by
  - Calculus or
  - Adding and subtracting $\mathbb{E}[X]$ and showing that the value is minimum when $y = \mathbb{E}[X]$

Theorem of the mean

- Aka theorem of the lazy statistician
- Given a RV X (discrete or continuous) and a scalar function $g(x)$, then $Y = g(X)$ is a RV
- ***Thesis***

$$\mathbb{E}[Y] = \mathbb{E}[g(X)] = \int_{-\infty}^{+\infty} g(x) f_X(x) dx$$

- ***Notes***
- One does not have to compute the PDF of Y to compute its mean, but use only the PDF of X and the function $g(x)$
- This formula holds trivially also for the identity function, since $X = Y = I(X)$
- This formula holds also for function of multiple RV using the joint PDF

Theorem of the mean: proof

- WLOG consider a *discrete* RV
- By definition

$$\mathbb{E}[Y] \triangleq \sum_{y_i \in \Omega_Y} y_i f_Y(y_i) = \sum_{y_i \in \Omega_Y} y_i \Pr(Y = y_i)$$

- Consider the generic y_i and express $\Pr(Y = y_i)$ in terms of $\Pr(X = \ldots)$

$$\begin{aligned}
\Pr(Y = y_i) &= \Pr(g(X) = y_i) \\
&= \Pr(X \in g^{-1}(y_i)) \\
&\text{(since a set of points } x_{ij} \text{ correspond to each } y_i) \\
&= \Pr(X = x_{i1} \cup X = x_{i2} \cup \ldots \cup X = x_{iN}) \\
&\text{(since all events are distinct)} \\
&= \Pr(X = x_{i1}) + \Pr(X = x_{i2}) + \ldots \Pr(X = x_{iN})
\end{aligned}$$

- Thus we can write $\mathbb{E}[Y]$:

$$\mathbb{E}[Y] = \sum_{y_i \in \Omega_Y} y_i (\Pr(X = x_{i1}) + \ldots + \Pr(X = x_{iN}))$$

- Now we should note that the previous summation is over all and only the possible x_i

Indicator variable of an event

- Consider a RV X and an event $A \subseteq \mathbb{R}$ (which corresponds to an event in the sample space $X^{-1}(A) \subseteq \Omega_X$)
- The indicator variable of an event A for RV X is a RV defined as:

$$I_A(X) \triangleq \begin{cases} 1 & \text{if } X \in A \\ 0 & \text{otherwise} \end{cases}$$

Intuition of indicator RV

- It is a way to synthesize specific events as RV from an already exiting RV
- It is a transformed version $g(X)$ of a RV

Example of indicator variable

- Consider X the result of a die toss:

$$\Omega_X = \{1, 2, 3, 4, 5, 6\}$$

- Consider the event $A \subset \Omega_X = \{\text{die outcome is even}\}$
- The indicator variable $I_A(X)$ is a RV that is 1 when the outcome die is even

Mean of an indicator variable

- Consider a RV X, an event A, and the indicator variable $I_A(X)$, then

$$\mathbb{E}[I_A(X)] = \Pr(X \in A) = \int x f_{I_A}(x) dx$$

Mean of an indicator variable: proof

- The mean $\mathbb{E}[I_A(X)]$ is

$$= \int_{-\infty}^{\infty} I_A(x) f_X(x) dx$$

(because of theorem on the mean of a function of a RV)

$$= \int_A I_A(x) f_X(x) dx$$

(since I_A is 0 outside A)

$$= \int_A f_X(x) dx$$

(since I_A is 1 inside A)

$$= \Pr(X \in A)$$

(by property of PDF)

Linearity of mean

- If $X_1, \ldots, X_n$ are RVs and $a_1, \ldots, a_n$ constant, then

$$\mathbb{E}[\sum_i a_i X_i] = \sum_i a_i \mathbb{E}[X_i]$$

- It can be proved by theorem of mean of RV
- Note that there is no assumption made on the RVs, i.e., the mean is linear even for RVs that are not independent or mutually exclusive

Mean of product of independent RVs

- If $X_1, \ldots, X_n$ are independent RVs, then:

$$\mathbb{E}[\prod_i X_i] = \prod_i \mathbb{E}[X_i]$$

- It can be proved by theorem of mean of RV and factorization of PDFs

Conditional mean

- The conditional mean of X given Y is defined as:

$$\mathbb{E}[X|Y=y] \triangleq \int_{-\infty}^{\infty} x f_{X|Y}(x,y)dx$$

  where the conditional PDF of X given Y is $f_{X|Y}(x,y) \triangleq \frac{f_{X,Y}(x,y)}{f_Y(y)}$
- Note that the conditional mean $\mathbb{E}[X|Y=y]$ is a function of y, while $\mathbb{E}[X]$ is a number

Conditional mean of independent variables

- If X and Y are independent, then $\mathbb{E}[X|Y = y] = \mathbb{E}[X]$

Law of total expectation

- Aka Law of iterated expectation, Adam's law
- The unconditional mean can be expressed in terms of conditional mean:

$$\mathbb{E}[X] = \mathbb{E}_Y[\mathbb{E}[X|Y]] = \int_{y=\infty}^{\infty} \mathbb{E}[X|Y = y] f_Y(y) dy$$

- This is similar to law of total probability

$$\Pr(X) = \sum_y \Pr(X|Y = y) \Pr(Y = y)$$

and for this reason it's called law of total expectation

Law of total expectation: proof

- It can be proven through:

$$f_X(x) = \int_{y=-\infty}^{\infty} f_{X,Y}(x,y)dy = \int f_{X|Y}(x,y)f_Y(y)dy$$

Example: random sum of RVs

- Let $W = X_1 + X_2 + \ldots + X_N$ where:
  - X_i are IID with mean μ_X and variance σ_X^2
  - N is a RV independent of X_i
- What is $\mathbb{E}[W]$?
- ***Solution***
- The mean is:

$$
\begin{aligned}
\mathbb{E}[W] &= \mathbb{E}_N[\mathbb{E}[W|N]] && \text{(from law of total expectation)} \\
&= \mathbb{E}_N[\sum_{i=1}^{N} X_i] \\
&= \mathbb{E}_N[N\mu_X] && \text{(from linearity and IID)} \\
&= \mathbb{E}[N]\mu_X && \text{(from linearity)}
\end{aligned}
$$

Corollary of law of total expectation

- If A_i is a partition of the outcome space Ω, i.e., events are mutually exclusive and exhaustive, then

$$\mathbb{E}[X] = \sum_{i=1}^{n} \mathbb{E}[X|A_i] \Pr(A_i)$$

Corollary of law of total expectation: proof

- Consider an indicator variable for each of the events A_i, I_{A_i}
- We can consider the indicator variable A given by the sum of all the I_{A_i}, which is the "certain event"
- By the Law of total expectation

$$\mathbb{E}[X] = \mathbb{E}_A[\mathbb{E}[X|A]] = \int_a \mathbb{E}[X|A = a] f_{A_i}(a) da = \sum_i \mathbb{E}[X|A_i] \int_{A_i} f_{A_i}(a) da = \sum_i$$

  since the expected value of an indicator variable is its probability
- TODO: not super clear

Theorem of mean for joint RVs

- Given two RVs X and Y and a function $g(x, y)$ then $g(X, Y)$ is a random variable and:

$$\mathbb{E}[g(X, Y)] = \int_x \int_y g(x, y) f_{X,Y}(x, y) dx$$

Theorem of conditional mean of a function of RV

- Given two RVs X and Y and a function $g(x)$
- The definition of conditional mean is:

$$\mathbb{E}[X|Y=y] \triangleq \int_x x f_{X|Y}(x,y)dx$$

which is a function of y

- If we transform X through $g(x)$ then the theorem of the mean applies:

$$\mathbb{E}[g(X)|Y=y] = \int_x g(x) f_{X|Y}(x,y)dx$$

Theorem of conditional mean of a function of RVs

- Given two RVs X and Y and a function $g(x, y)$ then $g(X, Y)$ is a random variable and:

$$\mathbb{E}[g(X, Y)|Y = y] = \int_{x=-\infty}^{\infty} g(x, y) f_{X|Y}(x, y) dx$$

- This is equivalent to theorem of the mean but applied to the conditional mean

Theorem of conditional mean of a function of RVs: proof

- By definition of conditional mean $\mathbb{E}[X|Y] \triangleq \int x f_{X|Y}(x) dx$
- The conditional mean is just a mean of a special RV $X|Y$
- The theorem of the mean still applies to $X|Y$

Variance and covariance

- Probability
- Random variables
- Mathematical expectation of RVs
  - Mean
  - **Variance and covariance**
  - Statistics of RVs
- Probability inequalities
- Statistical Inference

Variance of a RV

- Let X be a RV with mean $\mathbb{E}[X] < +\infty$
- The variance of X is defined as:

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

- The variance is also indicated as σ_X^2
- Note that the variance does not have the same unit of measure of the mean, but squared

Computing variance using theorem of mean

- Using the theorem of mean of RV:

$$\mathbb{V}[X] = \int_{x=-\infty}^{\infty} (x - \mathbb{E}[X])^2 f_X(x) dx$$

Standard deviation of a RV

- $=$ the positive square root of the variance:

$$\sigma_X = \sqrt{\mathbb{V}[X]}$$

- The standard deviation has the same unit of measure of the mean, while the variance has the squared dimension

Meaning of variance

- It represents the dispersion (or scatter) of the PDF / PMF of the RV around the mean

Variance of a die toss

- Using the definition:

$$\mathbb{V}[X] \triangleq \mathbb{E}[(X-\mu)^2] = (1-3.5)^2 \cdot \frac{1}{6} + ... + (6-3.5)^2 \cdot \frac{1}{6} = 2.92$$

Variance of a biased coin

- Using the definition:

$$\mathbb{V}[X] \triangleq \mathbb{E}[(X-\mu)^2] = (0-p)^2 \cdot (1-p) + (1-p)^2 \cdot p = (1-p)p$$

Alternative expression for variance

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

- Using the definition of variance and property of mean
$\mathbb{E}[(X-\mu)^2] = \mathbb{E}[X^2 + \mu^2 - 2X\mu]$

Variance of linear combination of 2 RV

- If a and b are constants

$$\mathbb{V}[aX + b] = a^2\mathbb{V}[X]$$

- Variance is not linear with respect to constants

Variance of independent RV

- If $X_1, \ldots, X_n$ are independent RVs and $a_1, \ldots, a_n$ are constants:

$$\mathbb{V}[\sum a_i X] = \sum a_i^2 \mathbb{V}[X]$$

Variance of the difference of RVs

- If X and Y are independent RVs then:

$$\mathbb{V}[X - Y] = \mathbb{V}[X] + \mathbb{V}[Y]$$

- Note that the variance of the difference of independent RVs is the sum of the variances, and not the difference

Law of total variance

- Aka Conditional variance identity, Eve's Law
- If X and Y are two RVs:

$$\mathbb{V}[X] = \mathbb{E}_Y[\mathbb{V}[X|Y]] + \mathbb{V}_Y(\mathbb{E}[X|Y])$$

- MEM: EVVE = Expected Variance + Variance of Expected

Law of total variance: proof

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

(variance property)

$= \mathbb{E}_Y[\mathbb{E}[X^2|Y]] - (\mathbb{E}_Y[\mathbb{E}[X|Y]])^2$

(law of total expectation to both sides)

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

(variance property, i.e., summing and subtracting $\mathbb{E}_Y[(\mathbb{E}[X|Y])^2]$

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

(linearity of mean)

$= \mathbb{E}_Y[\mathbb{V}[X|Y]] + \mathbb{V}_Y[\mathbb{E}[X|Y]]$ (variance property)

- MEM: It's about applying back and forth the alternative variance definition + law of total expectation

Law of total variance: example

- Let $W = X_1 + X_2 + ... + X_N$ where:
  - X_i are IID with mean μ_X and variance σ_X^2
  - N is a RV independent of X_i
- What is $\mathbb{V}[W]$?

Law of total variance: example solution

- Given the law of total variance:

$$\begin{aligned}\mathbb{V}[W] &= \mathbb{V}_N[\mathbb{E}[W|N]] + \mathbb{E}_N[\mathbb{V}[W|N]] \\ &= \mathbb{V}_N[\mathbb{E}[\sum X_i|N]] + \mathbb{E}_N[\mathbb{V}[\sum X_i|N] \\ &= \mathbb{V}_N[N\mu_X] + \mathbb{E}_N[N\sigma_X^2] \\ &= \mathbb{V}[N]\mu_X^2 + \mathbb{E}[N]\sigma_X^2\end{aligned}$$

- By using just the law of total expectation:

$$\begin{aligned}\mathbb{V}[W] &= \mathbb{E}[W^2] - \mathbb{E}[W]^2 \text{ (from alternative expression of variance)} \\ &= \mathbb{E}[(\sum X_i)^2] - (\mathbb{E}[N]\mu_x)^2 \text{ (from previous expression)} \\ &= \mathbb{E}_N[\mathbb{E}[(\sum X_i)^2|N] - ... \text{ (from law of iterated expectations)} \\ &= \mathbb{E}_N[\sum_{i=1}^{N} \mathbb{E}[X_i^2]] - ... \\ &= \mathbb{E}[N](\sigma_X^2 - \mu_X^2) - \mathbb{E}[N]^2\mu_X^2\end{aligned}$$

- TODO: Find the issue

Covariance of RV

- Given two RVs X and Y, the covariance is defined as:

$$\mathrm{Cov}[X, Y] \triangleq \mathbb{E}[(X - \mu_X)(Y - \mu_Y)]$$

- It is also indicated as $\sigma_{X,Y}$

Compute covariance using theorem of mean

- Using the theorem of the mean, the covariance can be written in terms of the joint PDF:

$$\text{Cov}[X, Y] = \int_y \int_x (x - \mu_X)(y - \mu_Y) f_{X,Y}(x, y) dxdy$$

Intuition of covariance

- It measures the strength of the *linear* relationship between random variables X and Y

Covariance in terms of mean

$$\mathrm{Cov}[X, Y] = \mathbb{E}[X \cdot Y] - \mathbb{E}[X] \cdot \mathbb{E}[Y]$$

- This is a generalization of the expression of variance in terms of mean

Covariance of independent RV

- If X and Y are independent RVs, then $\text{Cov}[X, Y] = 0$ (i.e., uncorrelated)
- Note that the converse is not true
  - Uncorrelated variables are not necessarily independent, simply there is no linear association

Variance for sum / difference of RV

- In case of two general RVs:

$$\mathbb{V}[X \pm Y] = \mathbb{V}[X] + \mathbb{V}[Y] \pm 2 \cdot \mathrm{Cov}[X, Y]$$

Relationship between covariance of RV and variance of RV

$$|\mathrm{Cov}[X, Y]| \leq \sqrt{\mathbb{V}[X]\mathbb{V}[Y]}$$

- In other symbols: $|\sigma_{X,Y}| \leq \sigma_X \sigma_Y$

Correlation coefficient of RVs

- Aka Pearson correlation, Pearson rho
- Given two RVs X and Y, the correlation coefficient is defined as:

$$\rho(X, Y) = \frac{\mathrm{Cov}[X, Y]}{\sqrt{\mathbb{V}[X]\mathbb{V}[Y]}} = \frac{\sigma_{X,Y}}{\sigma_X \sigma_Y}$$

- It is also indicated as $\rho_{X,Y}$
- Note that $-1 \leq \rho_{X,Y} \leq 1$

Meaning of correlation coefficient

- The coefficient of correlation is a *normalized* measure of *linear* dependence between RVs
- In fact:
    - If X and Y are independent (or at least uncorrelated) $\rho(X, Y) = 0$
    - If they are equal (or proportional) $\rho(X, Y) = 1$
    - If they are opposite $\rho(X, Y) = -1$

Rank of an array of numbers

- Consider an array of numbers $\underline{\boldsymbol{x}}$ (e.g., realizations of a RV)
- The rank of the numbers $\underline{\boldsymbol{x}}$ is the vector where each number x_i is replaced with its index in the sorted array $\underline{\boldsymbol{x}}$

$$r_X(x_i) = sort(\underline{\boldsymbol{x}}).idx(x_i)$$

Rank of an array of numbers: example

- $X = (7, 1, 9, 5)$
- Order the values in increasing order $1, 5, 7, 9,$
- Assign to $r_X(x_i)$ the index in the array corresponding to the value of x_i

$$r_X = (3, 1, 4, 2)$$

Rank of an array of numbers: interpretation

- Ranking removes the magnitude of the values and retains only information about the order of the values and their mutual relationship

Spearman rho: definition

- Aka rank correlation
- Consider two RVs X, Y and their realizations $\underline{\boldsymbol{x}}, \underline{\boldsymbol{y}}$
- Compute the rank variables $r_{\underline{\boldsymbol{x}}}, r_{\underline{\boldsymbol{y}}}$ corresponding to the realizations of $\underline{\boldsymbol{x}}$ and $\underline{\boldsymbol{y}}$
- Spearman rho is defined as the (Pearson) correlation coefficients between the ranks of the realizations of two RVs X and Y

$$\rho_S(X, Y) = \rho_P(r_{\underline{\boldsymbol{X}}}, r_{\underline{\boldsymbol{Y}}}) = \frac{\mathrm{Cov}[r_{\underline{\boldsymbol{X}}}, r_{\underline{\boldsymbol{Y}}}]}{\sqrt{\mathbb{V}[r_{\underline{\boldsymbol{X}}}] \cdot \mathbb{V}[r_{\underline{\boldsymbol{Y}}}]}}$$

Spearman rho: interpretation

- It is a non-parametric (i.e., there is no underlying model) measure of correlation
- It assesses how well the relationship between two variables can be described by a monotonic function

Pearson vs Spearman rho

- Pearson rho measures *linear* relationship
- Spearman rho measures *monotonic non-linear* relationship

Statistics of RVs

- Probability
- Random variables
- Mathematical expectation of RVs
  - Mean
  - Variance and covariance
  - **Statistics of RVs**
- Probability inequalities
- Statistical Inference

Summarizing statistics

- $=$ function of a PDF that generates a single number (e.g., mean)
- Summarizing statistics can be deceiving, since they hide information

Mode of a RV

- = the value of the RV that occurs most often (i.e., where PDF or PMF have a maximum)
- Note that a RV can have multiple modes and be multimodal (e.g., bimodal, trimodal)
- The mode provides a measure of central tendency, like the mean

Mean vs mode of a RV

- They are both measures of central tendency
- The mean is the average value of the RV when doing infinite IID draws (LLN)
- The mode is the most common value
- The mean can be a value that the RV does not assume
- The mode is a value of a RV
- A RV has a single mean, but can have many modes

Quantile of a RV

- The α-th quantile (with $0 \leq \alpha \leq 1$) of a RV X is the value $x_\alpha \in \mathbb{R}$ of the RV such that:
  - In terms of probability: $\Pr(X \leq x_\alpha) = \alpha$
  - In terms of CDF: $F_X(x_\alpha) = \alpha$ (i.e., the inverse of the CDF)
  - In terms of PDF: the portion of the PDF on the left of x_α is equal to α
- MEM: $x_\alpha = F_X^{-1}(\alpha)$

Quantile of a RV: more general definition

- For discrete RVs X the quantile value x_α might be not unique or be undefined
- In this case the definition is:

$$q_\alpha = \inf_x \{x : \Pr(x) \geq \alpha\}$$

Percentile of a RV

- $=$ the same as quantile where α is expressed as a percent (i.e., in [0%, 100%]) instead of [0, 1]
- E.g., 10th percentile (also called first decile) corresponds to the $\alpha = 0.1$ quantile
    - I.e., it gives an area under of the PDF to the left of it equal to 0.1

Median of a RV

- Aka 50th percentile, 0.5 quantile, or the fifth decile
- The median of a RV X is the value $x_{0.5}$ of the RV such that:
  - In terms of probability: $\Pr(X \leq x_{0.5}) = 0.5$, i.e., there is a 50-50 chance of getting a smaller or larger value of X than $x_{0.5}$
  - In terms of CDF: $F_X(x_{0.5}) = 0.5$
  - For continuous RV the median separates the PDF into 2 parts with equal underlying area 0.5
  - For discrete RV, the median might not exist or might not be unique, due to the discreteness of the CDF / PMF
- It is a measure of central tendency (like mean and mode)

Median is more robust than mean

- One outlier can affect the mean, since its effect is squared
- Outliers have a smaller effect on the median, since only the order (and not the magnitude) is considered

Geometric mean

- Given N RVs or values with $X_i \geq 0$

$$GM = \sqrt[N]{\prod_{i=1}^{N} X_i}$$

- MEM: $AM \geq GM$
- MEM: AM overestimates the true return, which is the GM

Geometric mean in terms of arithmetic mean

- The geometric can be written in terms of arithmetic mean:

$$
\begin{aligned}
GM &= \sqrt[N]{\prod_{i=1}^{N} X_i} \\
&= \exp\left(\frac{\sum_{i=1}^{N} \log(X_i)}{N}\right) \\
&= \exp(\text{avg}(\log(X_1), ..., \log(X_n)))
\end{aligned}
$$

- In words, the geometric mean is the exponential of the arithmetic mean of the logarithm of the values
- MEM: log, average, exp

Harmonic mean

$$HM = 1/\text{avg}(\frac{1}{X_1}, ..., \frac{1}{X_n}) = \frac{1}{(\frac{1}{N}\sum \frac{1}{X_i})} = \frac{n}{\sum \frac{1}{X_i}}$$

- In words, the harmonic mean is the reciprocal of the arithmetic mean of the reciprocals - MEM: $HM \geq GM$

Interquartile range of a RV

- $=$ the difference between the 75 and 25 percentile, i.e., $x_{0.75} - x_{0.25}$
- It measures how big is the x range that contains 50% of the mass around the median
- It is a measure of dispersion of a RV, like the variance

Mean absolute deviation

- Aka MAD
- It is defined as:
$$MAD \triangleq \mathbb{E}[|X - \mu|]$$
- It is a measure of dispersion of a RV, like the variance, but it weights the outliers less heavily than variance
- It is not differentiable

Semi-variance

- Sometimes we want to differentiate between upward and downward deviation
  - E.g., in case of returns for an asset, we are more concerned in downward deviations
- Downward semi-variance is defined:

$$\frac{\sum_{X_i < \mu} (X_i - \mu)^2}{\sum_{X_i < \mu} 1}$$

Skewness

- Skewness measures which side of the distribution is "heavier", and it is defined as:

$$\mathbb{E}[\left(\frac{X-\mu}{\sigma}\right)^3]$$

- MEM: It is the mean of the cube of the z-score of the RV
- ***Notes***
- For symmetric distributions skewness $= 0$
- Positive skewness means that the distribution has a longer tail to the right, and the peak is towards left
  - MEM: The skewness $>$ or <0 points to where is the heavier tail

Skewness: interpretation

- A distribution can be not symmetric and have one side with more mass than the other
- MEM: Think of a Gaussian, keep it centered, then move part of the peak towards the left, so the extra mass goes in the right tail (the mass needs to go somewhere)

Kurtosis

- Kurtosis measures the peaked-ness of the distribution, and it is defined as:

$$\mathbb{E}[\left(\frac{X-\mu}{\sigma}\right)^4]$$

- MEM: It is the mean of the 4th power of the z-score of the RV
- ***Notes***
- High kurtosis means sharper peak and fatter tails
    - MEM: High kurtosis is bad!
- Low kurthosis means rounder peak and thinner tails
- MEM: High kurtosis means sharp peak (Kurt is very thin)

Excess kurtosis

- A Gaussian has kurtosis $= 3$
- For this reason excess kurtosis refers to a Gaussian as baseline:

$$\text{excess kurtosis} = \text{kurtosis} - 3$$

Kurtosis: interpretation

- A distribution can have values concentrated near the mean or on the tails so that it has
  - Thick peak and shallow tails, or
  - Thin peak and fat tails
- MEM: One can start with a Gaussian and then make the peak thinner, the mass needs to go somewhere, and it goes in the tails

Probability inequalities

- Probability
- Random variables
- Mathematical expectation of RVs
- **Probability inequalities**
- Statistical Inference

PAC statements

- = Probably Approximately Correct statement
- In practice there is an approximation that holds with a certain probability
- Many probability inequalities are PAC statements

Markov inequality

- ***Hypothesis***
- Given X discrete or continuous RV
- X is a non-negative RV (i.e., $X \geq 0$, PDF is all after 0)
- X has finite mean: $\mathbb{E}[X] < \infty$
- ***Thesis***
- The probability that X is larger than a certain value is bounded by the mean
$$\Pr(X \geq x) \leq \frac{\mathbb{E}[X]}{x}$$

Markov inequality: geometric interpretation

- Given a RV $X \geq 0$ with a finite mean
- The "flipped CDF" $1 - F_X(x)$ is dominated by an hyperbole passing by $(y, x) = (\mathbb{E}[X], 1)$
- This is also related to the fact that a PDF needs to sum to 1 and thus needs to decrease at least like $1/n$

Proof of Markov inequality

- TODO: Add

Chebyshev inequality

- ***Hypothesis***
- Given X discrete or continuous RV
- X with finite mean μ and variance σ^2
- ***Thesis***
- The probability that X is far from the mean is bound by the variance:

$$\Pr(|X-\mu| \geq \varepsilon) \leq \frac{\sigma^2}{\varepsilon^2}$$

Chebyshev inequality in terms of z-scores

- ***Hypothesis***
- Given X discrete or continuous RV
- X with finite mean μ and variance σ^2
- ***Thesis***
- Expressing the distance from the mean in terms of standard deviation $\varepsilon = k\sigma$:

$$\Pr\left(\frac{|X-\mu|}{\sigma} \geq k\right) \leq \frac{1}{k^2}$$

- The probability that the z-score of a RV is far away from 0 at least a certain number k is bounded by $\frac{1}{k^2}$

Proof of Chebyshev inequality

- TODO: Add

Comparing Markov and Chebyshev inequalities

- Markov assumes $X \geq 0$
- Chebyshev makes no assumptions
- Both inequalities have a similar form:

$$\Pr(X \geq x) \leq \frac{\mu}{x}$$

$$\Pr(|X - \mu| \geq x) \leq \frac{\sigma^2}{x^2}$$

Hoeffding inequality

- Given a Bernoulli RV with probability of success μ
- We want to estimate μ using N samples:

$$\nu = \frac{1}{N} \sum_{i=1}^{N} X_i$$

- Then

$$\Pr(|\nu - \mu| > \varepsilon) \leq 2e^{-2\varepsilon^2 N}$$

- Since ν is bound in $[\mu - \varepsilon, \mu + \varepsilon]$, we want a small ε with a large probability

Statistical Inference

- Probability
- Random variables
- Mathematical expectation of RVs
- Probability inequalities
- **Statistical Inference**
    - Definitions
    - Sample mean
    - Sample variance
    - Asymptotics
    - Confidence intervals
    - Hypothesis testing
    - Multiple hypothesis testing
    - Estimating CDF and statistical functional
    - Bootstrap

Definitions

- Probability
- Random variables
- Mathematical expectation of RVs
- Probability inequalities
- Statistical Inference
    - **Definitions**
    - Sample mean
    - Sample variance
    - Asymptotics
    - Confidence intervals
    - Hypothesis testing
    - Multiple hypothesis testing
    - Estimating CDF and statistical functional
    - Bootstrap

Statistical inference

- $=$ process of generating a conclusion on a large population of objects from a small sample of the population

Population vs sample

- Population = the entire group of objects
- Sample = small part of the population

Examples of statistical inference

- Draw a conclusion about:
  - The fairness of a coin by tossing it repeatedly
  - The weights (or heights) of 12,000 students, selecting only 100 students
  - Defective bolts produced in a factory, by looking at 20 bolts manufactured during each day in a 6 day week (sample size = 120)

Sampling with / without replacement

- If we draw an element from a set, we have the choice of replacing it or not before drawing again

IID samples

- Given a RV $X \sim F$, we can draw N times from its distribution (i.e., without replacement) getting N samples $X_i \sim F$
- These samples X_i are Independent Identically Distributed (IID)

IID samples as idealized condition

- We can sample without replacement from a distribution F
  - F is a distribution so there are infinite samples and not a finite collection of objects
- If sampling was done with replacement, X_i and X_j could be the same and thus X_i could not be independent

Sample statistics

- A sample statistics Y is a deterministic function of given samples $X_1, ..., X_N$ of a population:

$$Y = g(X_1, ..., X_N)$$

- In general we are interested in functions that "summarize" properties of the samples
  - E.g., $g()$ can be mean, variance

Sample statistics is a RV

- A sample statistics $Y = g(X_1, ..., X_N)$ is a RV, since it is a function of RVs X_i
- In other words we can draw samples $x_i^{(k)}$ of X_i and get a different realization $y^{(k)}$ of the sample statistic Y

$$y^{(k)} = g(x_1^{(k)}, ..., x_N^{(k)})$$

Example of sample statistics

- X is a RV modeling the height of a student population P
- Pick 100 students randomly and have $X_1, ..., X_{100}$ RVs from the population P
- In one sample we have a realization for each student height $x_1, ..., x_{100}$
- Then we compute a sample statistic $h_1 = g(x_1, ..., x_{100})$: this is a realization of the RV sample statistics H

Example of sample statistics: OLS beta

- Assume $Y = \alpha + \beta X + \varepsilon$
- Estimate β through OLS, thus $\hat{\beta}$ is
  - A sample statistics
  - A RV, since it is function of the specific samples of x_i and y_i

$$\hat{\beta} = \frac{\overline{\text{Cov}}(Y, X)}{\overline{\mathbb{V}}[X]} = \frac{\frac{1}{N}\sum(y_i - \overline{y})(x_i - \overline{x})}{\frac{1}{N}\sum(x_i - \overline{x})^2}$$

Sampling distribution of a sample statistics

- Since a sample statistics is a RV, it has a probability distribution
- This distribution is called “sampling distribution of the (sample) statistics”

How to evaluate sampling distribution?

1. Closed form
   - Sometimes one knows the distribution of the sample statistics, e.g.,
     - Average of Gaussians is Gaussian
2. Enumeration
   - Consider all the possible samples, e.g., 100 samples from a population of 1,000, i.e., $\binom{1000}{100}$
   - Compute the probability distribution of the sample statistic
3. Approximation
   - Estimate the distribution of the sample statistics by sampling, e.g.,
     - Empirical distribution
     - Bootstrap

Estimator properties

- One can estimate different statistics of a RV (e.g., mean, variance, skewness, PDF) with an estimator, which is a sample statistics
- The estimator is a RV which has:
  - A mean (ideally equal to the mean of the estimated, aka un-biased-ness)
  - A std dev (std err of sample statistics)

Estimator properties: examples

- An estimator of:
  - The mean of X has a std dev, which is not the std dev of X although it is related to it
  - The std dev of X has a std dev in turn

Selection bias

- $=$ difference between the distribution of data sampled in a study vs the distribution of the underlying population

Self-selection bias

- Besides "selection bias" (who is selected to respond in a survey), there is also a "self-selection bias" from who decides to respond
- E.g., determining public opinion from letters or calls made to politicians, people who write / call are typically the ones with largest grievances

Publication bias

- = scientific journals prefer to publish studies that found an effect, rather than no effect

Small sample effect

- In small samples there is a higher probability of finding an effect, rather than in large studies

Meta-analysis

- $=$ analyzes results from several studies on the same topic
- "Funnel plot" to compare effect size to certainty of results

Anthropic selection bias

- Humans can only exist in universe that is capable of supporting human life
- E.g., when physics studies effect of different cosmological constants on multi-verse

Sample mean

- Probability
- Random variables
- Mathematical expectation of RVs
- Probability inequalities
- Statistical Inference
    - Definitions
    - **Sample mean**
    - Sample variance
    - Asymptotics
    - Confidence intervals
    - Hypothesis testing
    - Multiple hypothesis testing
    - Estimating CDF and statistical functional
    - Bootstrap

Sample mean

- Draw n IID samples $X_1, ..., X_n$ from a population
- The sample mean (or "mean of the sample") is the RV:

$$\overline{X} = \overline{X}_n \triangleq \frac{1}{n} \sum_{i=1}^{n} X_i = \overline{\mathbb{E}}[X_i]$$

Standard error of a sample statistics

- $=$ standard deviation of the sample distribution of a sample statistic (e.g., mean, variance, ...), e.g.,
  - Standard error of the mean
  - Standard error of the variance
  - Standard error of OLS regression coefficients

Standard error of the mean

- $=$ standard deviation of the sample mean of n IID samples
- Indicated with $\sigma_{\overline{X}}$, $SE_{\overline{X}}$, SEM

Sample mean is an unbiased estimator

- Assume that we want to estimate the mean μ of a population
- We take n IID samples of the population $X_1, ..., X_n$ (n RVs)
- The sample mean is defined as:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

- Different n samples drawn from the same population will give different values of the sample mean, so the sample mean is a RV
- The sample mean is an unbiased estimator of the population mean, since

$$\mathbb{E}[\overline{X}] = \mathbb{E}[\frac{1}{n}\sum_{i=1}^{n} X_i] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[X_i] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[X] = \mathbb{E}[X]$$

Standard error of the mean

- Assume that we want to estimate the SEM of a population
- We take n IID samples of the population $X_1, \ldots, X_n$ (n RVs),
- The sample mean is defined as:

$$\overline{X} = \frac{1}{n} \sum X_i$$

  and it is a RV
- The standard error of the mean is the standard deviation of $\overline{X}$ and it is equal to

$$\begin{aligned}
\mathbb{V}[\overline{X}] &\triangleq \mathbb{E}[(\overline{X} - \mathbb{E}[\overline{X}]] = \mathbb{E}[(\frac{1}{n} \sum_i X_i - \mu)^2] \\
&= \mathbb{E}[(\frac{1}{n}(\sum_i X_i - \mu))^2] = (\frac{1}{n^2} \mathbb{E}[(\sum_i X_i - \mu))^2] \\
&= (\frac{1}{n^2} \mathbb{E}[(\sum_i (X_i - \mu))^2] = (\frac{1}{n^2} n \mathbb{E}[(X - \mu))^2] \\
&= \frac{1}{n^2} n \mathbb{V}[X] = \frac{1}{n} \mathbb{V}[X]
\end{aligned}$$

Interpretation of the formula for std err of the mean

- This formula makes sense since:
  - If the variation of the underlying distribution being estimated (i.e., σ_X) is larger, the SEM is also larger
  - If the number of samples n is larger, the SEM is smaller

Estimate of standard error of the mean

- We know that
$$\mathbb{V}[\overline{X}] = \frac{\mathbb{V}[X]}{n} = \frac{\sigma_X^2}{n}$$
but we might not know σ_X
- In many formulas, if we don't know the std dev of the population σ_X, we can use the sample standard deviation S
$$\mathbb{V}[\overline{X}] \approx \frac{S^2}{n}$$

Summary of properties for sample mean

- Draw n IID samples $X_1, \ldots, X_n$ from a population
- Compute sample mean $\overline{X}$ from the samples
- What is the relationship between the probability distribution of the sample mean $\overline{X}$ and X?
- ***Expected value***
- The population mean $\mathbb{E}[X]$ is the center of mass of the population distribution
- The sample mean $\overline{X}$ is the center of mass of the observed data distribution
- The sample mean is an unbiased estimate of the population mean, i.e., $\mathbb{E}[\overline{X}] = \mathbb{E}[X]$
- ***Variance***
- The more data n is used to compute the sample mean $\overline{X}$, the more concentrated is the PDF / PMF of the sample mean around the population mean

Sample variance

- Probability
- Random variables
- Mathematical expectation of RVs
- Probability inequalities
- Statistical Inference
    - Definitions
    - Sample mean
    - **Sample variance**
    - Asymptotics
    - Confidence intervals
    - Hypothesis testing
    - Multiple hypothesis testing
    - Estimating CDF and statistical functional
    - Bootstrap

Unbiased estimator of population variance knowing population mean μ_X

- If we know the mean μ_X of the underlying population X, then

$$S^2 = \frac{1}{n}\sum_i (X_i - \mu_X)^2$$

is an unbiased estimation of the population variance $\mathbb{V}[X]$, i.e.,

$$\mathbb{E}[S^2] = \mathbb{V}[X]$$

- Aka sample variance
- ***Proof***

$$\mathbb{E}[S^2] \triangleq \mathbb{E}[\frac{1}{n}\sum_i (X_i - \mu_X)^2] = \frac{1}{n}\sum_i \mathbb{E}[(X_i - \mu)^2] = \frac{1}{n}\sum_i \mathbb{V}[X] = \mathbb{V}[X]$$

Unbiased estimator of population variance not knowing μ_X

- If we need to estimate the mean of the underlying population from the data, then the sample variance:

$$S^2 = \frac{1}{n-1} \sum_i (X_i - (\frac{1}{n} \sum_j X_j))^2$$

is an unbiased estimate of the variance of X, i.e., $\mathbb{E}[S^2] = \mathbb{V}[X]$

- Note that we need to divide by $n - 1$, instead of n to get an unbiased estimate
- This is because the mean is also estimate from the data and it is using a degree of freedom

Sample variance as RV

- Since the sample variance S^2 is a function of the data, then S^2
  - Is a RV
  - Has a population distribution
- The expected value of the population distribution of the sample variance $\mathbb{E}[S^2]$ is the variance of the population that we are estimating $\mathbb{V}[X]$ (unbiased estimate)
- The more data n we have
  - The more concentrated the distribution of S^2 is
  - We don't have a relationship for it in general (it is function of higher moments), so we can use numerical techniques (e.g., bootstrap)

Asymptotics

- Probability
- Random variables
- Mathematical expectation of RVs
- Probability inequalities
- Statistical Inference
    - Definitions
    - Sample mean
    - Sample variance
    - **Asymptotics**
    - Confidence intervals
    - Hypothesis testing
    - Multiple hypothesis testing
    - Estimating CDF and statistical functional
    - Bootstrap

Asymptotics

- $=$ behavior of sample statistics as the sample size n goes to infinity
- E.g., Law of Large Numbers (LLN) and Central Limit Theorem (CLT)

LLN vs CLT

- Both are statements about the sample mean
  - LLN: sample mean is consistent
  - CLT: sample mean is asymptotically Gaussian
- Both are about asymptotic behaviors $n \to \infty$
- Both apply to continuous and discrete RVs

Consistent estimator

- An estimator is consistent $\iff$ its value converges to what should estimate as the amount of collected data goes to infinity

Consistent vs unbiased estimator

- An *unbiased* estimate refers to averaging an infinite number $\mathbb{E}$ of times a fixed number n of samples

$$\mathbb{E}[g(X_1, ..., X_n)]$$

- A *consistent* estimate refers to doing a single average of a diverging number of samples n

$$\lim_{n \to \infty} g(X_1, ..., X_n)$$

Law of Large Numbers (LLN) in few words

- The LLN is about consistency of the sample mean, i.e., it tells us what happens to the sample mean when we collect an infinite amount of samples

Law of Large Numbers (LLN)

- The sample mean of IID samples:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

  is a consistent estimator for the population mean, i.e.,

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} X_i = \mathbb{E}[X]$$

- The convergence is in probability

LLN for other estimators

- LLN applies also to other estimators that rely on the mean, e.g., variance and std dev:

$$\frac{1}{n}\sum_{j} h(X_i) \to \int g(y)dF_X(y) = \mathbb{E}[g(X)]$$

  where the convergence is in probability
- TODO: who is h vs g?

LLN for variance

- For the variance:

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

- Using LLN:

$$\lim_{n\to\infty} \mathbb{V}[\overline{X}] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = \mathbb{V}[X]$$

  where we used $g(x) = x^2$ in the first term and the previous result, and applying the LLN for the mean to the second term

Example of LLN: Bernoulli distribution

- Consider IID draws X_i from a Bernoulli variable with a certain p
- From LLN

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} X_i = \mathbb{E}[X] = p$$

- In words the fraction of times that the coin comes up heads (i.e., the relative frequency) approximates the probability of success for $n \to \infty$
    - This is the basis for the frequentist interpretation of probability

Central Limit Theorem (CLT) in few words

- The CLT tells that the shape of the distribution of the sample mean $\overline{X}$ for large n is Gaussian, independently from the PDF of the sampled distribution X

Central Limit Theorem (CLT)

- Consider the sample mean of IID samples

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

then for large n $\overline{X}$ is Gaussian, independently from the PDF of the sampled distribution X

- Note that CLT does not tell anything about the rate of convergence to a Gaussian in terms of the value of n

CLT + LLN

- Combining all the asymptotic results for $\overline{X}$: a) CLT about the shape of $\overline{X}$ b) LLN for the mean of $\overline{X}$ c) variance of $\overline{X} = \mathbb{V}[x]/n$
- We obtain that for large n
  1. Sample mean is Gaussian centered on the population mean and with a variance related to the population variance
  $$\overline{X} \sim N(\mathbb{E}[X], \frac{\mathbb{V}[X]}{n})$$
  2. Z-scoring 1)
  $$\frac{\overline{X} - \mu_X}{\sigma_X/\sqrt{n}} \sim N(0,1)$$
  3. Using the sample estimates of the unknown quantities
  $$\frac{\overline{X} - \mu_{\overline{X}}}{\sigma_{\overline{X}}} \sim N(0,1)$$
  4. In words 3)
  $$\frac{\text{estimate} - \text{mean of estimate}}{\text{std err of estimate}} \sim N(0,1)$$

Example of CLT for fair dice

- Let X_i be the outcome of rolling a fair die
  - We know that $\mathbb{E}[X] = 3.5$ and $\mathbb{V}[X] = 2.92$, so we don't have to estimate anything from the data
- We roll n dice and take the average $\overline{X}$, i.e., compute the sample mean
- The sample mean is a discrete RV
- The CLT tells that:

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \to Z$$

  tends to a standard normal for large n
- We can verify this numerically by:
  - Performing many experiments of averaging n die outcomes
  - Standardizing the resulting distribution
  - Plotting the histogram against a standard normal

Example of CLT for biased coin flip

- Let X_i be the outcome of a biased coin with unknown probability of success p
- We know that $\mathbb{E}[X] = p$ and $\mathbb{V}[X] = p(1-p)$
- How to estimate p?
- ***Solution***
- Let's call $\hat{p}$ the sample proportion of successes:

$$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

- The CLT tells us that:

$$\frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}}$$

  tends to a standard normal for large n, so we can estimate $\hat{p}$
- Note that the approximation gets better for larger n, but there is no information on the rate of convergence, e.g., for different values of the params the rate of convergence can be different

Confidence intervals

- Probability
- Random variables
- Mathematical expectation of RVs
- Probability inequalities
- Statistical Inference
  - Definitions
  - Sample mean
  - Sample variance
  - Asymptotics
  - **Confidence intervals**
  - Hypothesis testing
  - Multiple hypothesis testing
  - Estimating CDF and statistical functional
  - Bootstrap

Confidence intervals for a statistic

- Given a statistic Y (e.g., mean, median) of a RV X, the α-confidence interval I is the interval that contains the true value of the statistic with a certain probability α:

$$\Pr(Y \in I) = \alpha$$

Confidence intervals for the sample mean

- The α-confidence interval for the sample mean $\overline{X}$ is the interval I around μ_X such that $\Pr(\mu \in I) = \alpha$

Confidence intervals for the mean using sample mean

- Every time we know the distribution of the sample mean $\overline{X}$, we can compute confidence intervals for the mean of the underlying population μ_X

Correct interpretation of confidence intervals

- Assume we estimate the confidence interval of average height of female population in US using the sample mean and standard error
- We claim: “the national mean female height is between 63 and 65 inches with 95% probability”
- ***Incorrect interpretation***
- We have no way to assess the probability of the confidence interval to contain the population mean, since it’s unknown
- This is a misinterpretation of the meaning of confidence intervals
- The statement is not a Bayesian statement
- ***Correct interpretation***
- The statement needs to be interpreted in a frequentist sense
- If we compute the confidence intervals many times (e.g., extracting different data sets), in 95% of the cases the confidence interval will capture the true population mean

z-confidence intervals for the mean

- X is not a Gaussian RV
- We use a large number n of IID samples from X
- We know σ_X variance of the underlying population
- ***Thesis***
- We want to use the realization of $\overline{X}$ we have to estimate the unknown $\mu_X = \mathbb{E}[X]$
- ***Algorithm***
- We know that:
  - $\overline{X}$ is Gaussian (because of CLT)
  - $\overline{X}$ has mean $\mathbb{E}[X]$ (because the sample mean is unbiased estimator)
  - The std err of $\overline{X}$ is $\frac{\sigma_X}{\sqrt{n}}$
- We can build α (e.g., 95%) confidence interval for $\overline{X}$ in the form:

$$\Pr(\overline{X} \text{ inside } \mu_X \pm Z_\alpha \frac{\sigma_X}{\sqrt{n}}) = \alpha$$

where Z_α is a two-sided standard normal quantile (e.g.,

t-confidence intervals for the mean

- X is Gaussian
- We use a small number of n of IID samples from X
- We don't know σ_X
- ***Thesis***
- We have a realization of $\overline{X}$ and we want to use this information to estimate the unknown $\mu_X = \mathbb{E}[X]$
- ***Algorithm***
- Take n IID samples X_i of a Gaussian and compute:
  - The (unbiased) sample mean: $\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$
  - The (unbiased) sample variance: $S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2$
- We know that:
  - $\overline{X}$ is Gaussian, since it is a linear combination of Gaussians
  - S^2 is chi-square with $n-1$ degrees of freedom (multiplied by a constant), since it is sum of squared IID standard Gaussians
  - $T = \frac{\overline{X} - \mathbb{E}[\overline{X}]}{S/\sqrt{n}}$ has a t-distribution with $\nu = n-1$ degrees of freedom after

z- vs t-confidence intervals for the mean

- ***z-confidence intervals***
- The hypotheses for z-confidence intervals for the mean are:
  - X with any distribution
  - $\mathbb{V}[X]$ is known
  - Large sample size n
- The sample mean $\overline{X}$ is Gaussian
- We can build z-confidence intervals for μ_X in the form:

$$\mu_X \in \overline{X} \pm Z_\alpha \times \frac{\sigma_X}{\sqrt{n}}$$

  where:
  - Z_α is a 2-sided z-quantile
- ***t-confidence intervals***
- The hypotheses for t-confidence intervals for the mean are:
  - X is Gaussian
  - $\mathbb{V}[X]$ is unknown
  - Small sample size n
- The sample mean $\overline{X}$ is a student t-distribution with $n-1$ degrees of freedom
- We can build t-confidence intervals for μ_X in the form:

$$\in \overline{X} \pm T \quad s$$

When to use t-confidence intervals

- In general it is always better to use t-intervals when using numerical methods
  - For back of the envelope calculations z-intervals can be easier to compute
- The t-confidence intervals assume that data is from IID normal distribution
  - It still works as long as distribution is roughly symmetric and mound-shaped

Confidence intervals for asymmetric distributions

- For skewed distributions:
  - Cannot use t-distributions (in fact the confidence intervals will not be symmetric around the mean)
  - Take logs of the observations to make distributions more symmetrical
  - Use bootstrap

T-confidence intervals for paired observations

- For paired observations X and Y
  - Take the difference of paired observations to get a new RV $X - Y$
  - Use t-intervals for the mean of difference μ_{X-Y}

T-confidence intervals for groups in randomized trial (A/B test)

- We want to compare the measures from two groups in a randomized trial, also known as A/B test
- E.g., receiving a medicine vs a placebo
- We randomize the trials before assigning to A and B to balance covariates in the two groups, that might contaminate the results
- We cannot use paired observations since the groups are independent
- We assume that:
  - The variance in the two groups is the same
  - The number of samples for the groups are n_x and n_y
- The α (e.g., 95%) confidence interval for $\mu_Y - \mu_X$ is:

$$\overline{Y} - \overline{X} \pm t_{\nu,\alpha} S_p \left(\frac{1}{n_x} + \frac{1}{n_y} \right)^{1/2}$$

where:

  - The degrees of freedom of the t-distribution are $\nu = n_x + n_y - 2$

Hypothesis testing

- Probability
- Random variables
- Mathematical expectation of RVs
- Probability inequalities
- Statistical Inference
  - Definitions
  - Sample mean
  - Sample variance
  - Asymptotics
  - Confidence intervals
  - **Hypothesis testing**
  - Multiple hypothesis testing
  - Estimating CDF and statistical functional
  - Bootstrap

What is hypothesis testing?

- We have a statement H_a about a phenomenon
- We want to quantify the statistical evidence supporting it

Hypothesis testing set-up

- We consider two hypotheses to explain the phenomenon under study:
    1. The alternative hypothesis H_a: what we want to test
    2. The null hypothesis H_0 (read "h-nought"): the phenomenon is just the result of random fluctuations
- We assume H_0 is true unless the evidence strongly suggests that H_a is true and that H_0 should be rejected
- MEM: It's like a legal trial where one is assumed innocent (H_0) until proved guilty (H_a) beyond reasonable doubt (statistical evidence)

Test statistic, rejection region, and decision

- We compute the distribution of the test statistics under H_0
- Compute rejection region for null hypothesis from confidence level α
- Observed data are used to compute a test statistics
- One reject either the null or alternative hypothesis based on the value of the test statistics compared to the rejection region of the null hypothesis

Accepting null / alternative hypothesis

- In statistics / physics we should always use the term "rejecting an hypothesis", instead of "proving an hypothesis" since:
  1. We cannot find evidence that proves a theory right
  2. We can only find evidence that falsifies an hypothesis
- E.g., we cannot prove the statement "all swans are white", since we should examine all swans and make sure they are all white
- Instead if we find a single non-white swan we can reject the statement "all swans are white"

Type I and II errors

- There are 4 possible outcomes of our statistical decision process
    1. True negative: correctly accept null hypothesis
    2. True positive: correctly accept alternative hypothesis
    3. Type I error (false positive): accept H_a when H_0 is true
    4. Type II error (false negative): accept H_0 when H_a is true
- MEM: P in FP has 1 downstroke -$>$ type 1, N in FN has 2 down-strokes -$>$ type 2

Probabilities of type I and II in hypothesis testing

- The probability of
  - Type I error (i.e., false positive) is called α, i.e., the confidence level of the test
  - Type II error (i.e., false negative) is called β (related to power of test)

Confidence level vs confidence interval

- There is a little confusion between α in the context of:
  1. Confidence intervals
     - E.g., “an interval with $\alpha = 95\%$ confidence interval”
  2. Hypothesis testing
     - E.g., “a test with $\alpha = 5\%$ confidence level”

One-sided vs two-sided test

- One-sided test: $H_0 : \Theta = \Theta_0$ vs $H_a : \Theta > \Theta_0$
- Two-sided test: $H_0 : \Theta = \Theta_0$ vs $H_a : \Theta \neq \Theta_0 \iff H_a : \Theta > \Theta_0$ or $\Theta < \Theta_0$
- Note that this implies different rejection regions for the same confidence level α

One-sided hypothesis test: example of reasoning for sample mean

- We have a RV X with a mean assumed to be μ_0
- We sample X and get a $\overline{x} > \mu_0$ (i.e., a realization of the sample mean $\overline{X}$)
- Did we get the value $\overline{x}$ because the mean of X is:
  - Truly μ_0 and there are random fluctuations due to the stochastic nature of X (null hypothesis); or
  - Larger than μ_0 (alternative hypothesis)?
- We don't know the answer, but we can require that the probability to reject the null hypothesis by mistake (i.e., false positive) is a certain confidence level α
  - We assume H_0
  - We compute the interval of values C that would make us reject H_0 (rejection region)
  - Check whether $\overline{x} > C$ or not
- Note that if $\overline{x} > C$
  - We still don't know if we just witness a rare event or H_0 is false
  - We only know that if there was a rare event, it had a probability less than α to happen
- To perform the test we need to know the sample statistics under the null hypothesis (e.g., normal or t-distribution)

Hypothesis testing algorithm

- Assume H_0
- Set a confidence level α
- Compute the rejection region under H_0 for the test statistic at a given confidence level α
- Compare the test statistics computed from the data with the rejection region
- Report the binary outcome "reject / retain H_0"

One-sided hypothesis test: example

- Consider testing the hypothesis about the mean μ_X of X:

$$H_0 : \mu_X = \mu_0$$
$$H_a : \mu_X > \mu_0$$

- The idea is to reject the null hypothesis if $\overline{X}$ is larger than a constant C chosen so that the probability of a type I error (i.e., false positive) is α:

$$\Pr(\overline{X} > C \mid H_0) = \alpha$$

  where $\overline{X}$ has a certain sample statistics (e.g., normal or t-distribution)

- In other terms from α we come up with the constant C and then we verify if $\overline{x}$ is $> C$ or not
- ***Numerical example***
- We need to find $C : \Pr(\overline{X} > C | H_0) = \alpha$
- Assume $\overline{X} \sim N(\mu_0, \sigma_X^2/n)$ and $\alpha = 0.05$ then

$$C = \mu_0 + 1.645 \times \sigma_X/\sqrt{n}$$

- Often we prefer to express the previous equation in terms of Z-scores:

Two-sided hypothesis test: example

- Consider testing the hypothesis about the mean μ_X of X:

$$H_0 : \mu_X = \mu_0$$

$$H_a : \mu_X \neq \mu_0$$

- The idea is still to find an interval so that one would mistakenly reject H_0 with a probability α
- In this case we reject the null hypothesis if the test statistic is either too large or too small

$$\Pr(|\frac{\overline{X} - \mu_0}{\sigma_X/\sqrt{n}}| > C) = \alpha$$

  so we need to consider the area under both tails of the PDF
- E.g., for $\alpha = 0.05$ we need $C = 2$, which is a more stringent check than $C = 1.645$ needed for a 1-sided test, since the prior is weaker

P-value

- $=$ probability under the null hypothesis of obtaining evidence as or more extreme than what observed

$$\text{p-value} = \Pr(\text{seeing evidence } \geq H_a | H_0)$$

- It can be one-sided or two-sided

Interpretation of p-values

- P-values answer the question: "suppose nothing is going on: how unusual it is to see the estimate we got?"
- If the p-value is small, then either "H_0 is true and we have observed a rare event" or "H_0 is false"

Example of p-value

- Testing $H_0 : \mu = \mu_0$ vs $H_a : \mu > \mu_0$, we get a test statistic (t-score in this case) of 2.5 for 15 df
- What's the probability of getting a t-score ≥ 2.5 by chance?
- `pt(2.5, 15, lower.tail=FALSE) = 0.01225 = 1\%`

P-value vs hypothesis testing

- P-value and hypothesis testing are related but look at the problem in different ways
  1. In hypothesis testing
     - We compute the rejection region for H_0 that gives the desired significance level α
     - We compare the test statistic with the rejection region
     - The answer is binary, i.e., reject / accept null hypothesis
  2. With p-values
     - The result is a probability, i.e., the probability of getting the evidence under the null hypothesis

P-value in terms of confidence level of hypothesis testing

- We can think of the p-value as the smallest value of confidence level α for which we would still reject the null-hypothesis
- The rejection region is bounded by the value that has a p-value equal to the confidence level α, e.g.,
  - If p-value is $3\% = 0.03$ we can reject the null hypothesis up to a confidence level of 0.03
  - We reject H_0 at $\alpha = 0.05, 0.04, 0.03$ but not at $\alpha = 0.02$

Example of p-value (7 girls)

- A friend has 8 children, 7 of which are girls
- We wonder if the probability of having a girl p is 0.5: $H_0 : p = 0.5$ vs $H_a : p > 0.5$
- Under H_0 the test statistic is binomial, and compute the probability of seeing the data under H_0: $\Pr(\text{Binomial}(0.5, 8) \geq 7)$

```
choose(8, 7) * 0.5^8 + choose(8, 8) * 0.5^8
= pbinom(6, size=8, prob=0.5, lower.tail=FALSE) = 0.03516
```

(in R for discrete probability we need to decrease the count by 1, since R considers $>$)

- If we were testing this hypothesis we would reject H_0 at 5% level, at 4% level, until the p-value of 3.516%

Example of p-value (infection rate)

- An hospital has an infection rate of 10 infections per 100 person / days, i.e., rate $= 0.1$ person per unit of time
- Assume that an infection rate of 0.05 is an important benchmark (e.g., above that threshold some expensive quality control procedure is in place, or shut down the hospital)
- We don't want to raise an alarm due to just random fluctuations, so we test formally the hypothesis modeling the uncertainty as Poisson:

$$H_0 : \lambda = 0.05 \text{ vs } H_a : \lambda > 0.05$$

- We need to compute the probability of the evidence i.e., obtaining 10 or more infections in the monitoring period of 100 days, assuming that H_0 (i.e., the rate is 5):

```
ppois(9, 5, lower.tail=FALSE) = 0.03183
```

  (R does $>$ so we need to decrease the count by 1)

- If we want confidence level of $\alpha = 0.01$ then we should not execute the quality control procedure, for $\alpha = 0.05$ we should execute the procedures

Trade-off between confidence level and power of a test

- We want to avoid false positives
  - Thus we limit the false positive rate $\Pr(H_a|H_0)$ using a low significance level α
- On one hand, if all we cared was to not make mistakes
  - We could set α to a very low level
  - Then the test would not detect any positives at all
- On the other hand we are also interested in rejecting H_0 when it is false
  - This is related to power of a test $\Pr(H_a|H_a)$

Power of a test

- The power of a test is the probability of rejecting the null-hypothesis when it is false

$$\text{power} \triangleq \Pr(\text{Reject } H_0 \mid H_0 \text{ is false}) = \Pr(H_a|H_a)$$

- One wants tests to have tests with high power
- Typically one designs an experiment (e.g., needed number of samples) so that it is possible to reject the null-hypothesis if it is false

Power of a test as function of β

- $\beta = \Pr(H_0|H_a)$ is the probability of type II error (i.e., false negative)
- Power of a test is defined as $\Pr(H_a|H_a)$
- Thus the power of a test is equal to $1 - \beta$

Example of calculating power for z-test

- Assume that we know σ and we use a z-test
- $H_0 : \overline{X} \sim N(\mu_0, \sigma^2/n)$ vs $H_a : \overline{X} \sim N(\mu_a, \sigma^2/n)$
- The power of the test is defined as $\Pr(\text{Reject } H_0 \mid H_a \text{ is true})$
- Since H_a is assumed true, then there is a distribution for $\overline{X}$ under H_a
- In hypothesis testing we reject H_0 at confidence level α if $\frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}} > Z_{1-\alpha}$
- The formula for power of z-test is:

$$\Pr\left(\frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}} > Z_{1-\alpha} \mid \overline{X} \sim N(\mu_a, \sigma/\sqrt{n})\right)$$

- Note that the power is function of:
  - The value of the test statistic that we want to detect in H_a (μ_a)
  - The value of the test statistic assumed in H_0 (μ_0)
  - The significance level α
  - σ and n through the sample variance
- The power of the test to detect the real μ_a is:

```
z <- qnorm(1 - alpha)
```

T-test power

- If we don't know σ we need to use a t-test
- We always prefer to use t-test instead of z-test, since it is more accurate:

$$\Pr\left(\frac{\overline{X} - \mu_0}{S/\sqrt{n}} > t_{n-1,1-\alpha} \mid \overline{X} \text{ has } \mu = \mu_a\right)$$

- Note that the t-distribution for $\mu = \mu_a$ is a non-central t-distribution (i.e., it is not centered around 0)
- In R there is a function `power.t.test` to compute the power

Multiple hypothesis testing

- Probability
- Random variables
- Mathematical expectation of RVs
- Probability inequalities
- Statistical Inference
    - Definitions
    - Sample mean
    - Sample variance
    - Asymptotics
    - Confidence intervals
    - Hypothesis testing
    - **Multiple hypothesis testing**
    - Estimating CDF and statistical functional
    - Bootstrap

Multiple tests and false discoveries

- Current era of statistics is characterized by:
  - Huge data sets (since data is cheap)
  - Performing thousands of hypothesis tests to answer questions
- Performing multiple comparisons leads to false positives / discoveries

P-hacking

- $=$ running many experiments and then reporting the one with smallest p-value

Example of data mining (jelly beans and acne)

- One believes that jelly beans cause acne
  - Get data about consumption of jelly beans and occurrence of acne
  - This relationship is tested and nothing is found at 5% significance
  - Then one might start testing jelly bean of each color one at the time
  - After 20 attempts one finds out that pink jelly beans correlate with acne with a p-value of 5%
- By running 20 experiments each with 5% probability of being incorrect by chance there is almost certainty to find something
- ***Correct approach***
- Come up with hypotheses ahead of time
- Adjust the p-value to account for data mining bias / multiple hypothesis testing
- Hold out data to verify the relationship we have found

Example of multiple-comparison problem using coins

- A procedure to determine if a coin is unfair consists in
  - Flipping a coin 10 times
  - Checking if it lands heads 10 times
- The null hypothesis is that the coin is fair
- ***Single test***
- Assume the coin is fair
- The p-value of the test under H_0 is $1/2^{10} = 1/1024 \approx 0.001$, which results in rejecting H_0 with a p-value < 0.05
- ***Multiple tests***
- The probability that at least one coin out of 1000 is not fair (by luck) is almost 1, since it is 1 - probability that are all fair: $1 - (1 - 0.001)^{1000} \approx 1$
- A multiple comparison problem arises if we want to use this test to check the unfairness of many coins

Nomenclature for multiple testing

- Consider the 4 possible scenarios for:
  - Decision: accepting H_0 or H_a
  - Ground truth: H_0 or H_a is true

	H_0 true	H_a true	
Accept H_0	TN	FN	
Accept H_a	FP	TP	

- These scenarios are mutually exclusive and cover all the possibilities, so their sum is equal to the number of experiments m
- Let's call according to standard nomenclature:
  - m: the total number of hypotheses tested
  - m_0: the number of true null hypotheses
  - V: the number of rejected null hypotheses when the null was true (i.e., false positives)
  - R: the total number of rejected null hypotheses (i.e., discoveries)
    - MEM: R is discoveRy
- The 4 quantities m, m_0, R, V allow to compute the entire confusion matrix

	H_0 true	H_a true	
Accept H_0			

Probability of false positive

- The probability of a false positive is defined as:

$$\Pr(FP) = \Pr(\text{Accept } H_a | H_0 \text{ is true}) = \frac{\Pr(\text{Accept } H_a \text{ and } H_0 \text{ is true})}{\Pr(H_0 \text{ is true})} = \lim_{m \to \infty}$$

- In words, we do infinite experiments and then compute the probability of false positive
- The problem is that it is not observable quantity

False Positive Rate

- Aka FPR
- The false positive rate is defined as

$$FPR = \mathbb{E}_m[\frac{V}{m_0}]$$

- Expectation is over repeating the m experiments
- Note that probability of false positive and expectation of the ratio is different, since in first case we consider a single experiment, in the second we consider the ensemble of experiments (???)

Controlling FPR in a single experiment

- Discarding all discoveries with p-value $< \alpha$ controls the false positive rate at level α *on average* for a *single* experiment
  - On average: in the sense that we could do the same experiments over and over
  - For a single experiment: this might be not enough when doing lots of tests (e.g., 10,000) because of the multiple-comparison problem

Family Wise Error Rate

- Aka FWER
- Defined as:

$$FWER \triangleq \Pr(V \geq 1)$$

  i.e., the probability of at least one false positive running m experiments
- Family-wise seems to refer to a "family" of experiments (i.e., a multiple comparison)

Bonferroni correction to control FWER

- Suppose you do m tests
- You want to control FWER so that $\Pr(V \geq 1) \leq \alpha$
- Calculate p-values
- Call significant any experiment for which p-value $< \alpha/m$, i.e., $\Pr(H_a^{(i)} | H_0^{(i)}) < \alpha/m$

Bonferroni correction to control FWER: proof

- We can use the union bound to show that the p-value over multiple tests is less than α

$$
\begin{aligned}
\Pr(V \geq 1) &= \Pr(\text{reject falsely at least one } H_0) && \text{(by def)} \\
&= \Pr(FP_1 \cup FP_2 \cup ... \cup FP_n) && \text{(expanding "at least")} \\
&\leq \sum_i \Pr(FP_i) && \text{(union bound)} \\
&\leq \sum \frac{\alpha}{m} && \text{(confidence level)} \\
&= \alpha
\end{aligned}
$$

- Note that no assumption of independence between tests is made
- ??? If there is independence the bound can be improved since

$$\Pr(V \geq 1) = 1 - \Pr(V = 0) = 1 - \Pr(\neg FP_1 \wedge \neg FP_2 ... \wedge \neg FP_n) = 1 - (\Pr(\neg FP))^n$$

Bonferroni correction: pros and cons

- Pros
  - Easy to calculate
- Cons
  - May be very conservative

False Discovery Rate

- Aka FDR
- Defined as:

$$FDR = \mathbb{E}\left[\frac{V}{R}\right]$$

  i.e., the average fraction of false positives V with respect to the discoveries R
- This is an interesting metric since we know how many discoveries we made
  - Number of discoveries R are an observable variable

FDR-controlling procedures

- FDR-controlling procedures, when conducting multiple comparison, are designed to *control* the *expected* proportion of:
  - False discoveries (which is an observable variable)
  - Not false positives (which is a variable we cannot observe)

FPR vs FWER vs FDR

- False-Positive Rate is:

$$FPR = \mathbb{E}_m[\frac{V}{m_0}]$$

  the expected fraction of false discoveries (V) with respect to to the number of null hypotheses to reject (m_0)

  - It is the ratio of two not observable quantities

- Family-Wise Error Rate is:

$$FWER = \Pr(V \geq 1)$$

  the probability of having at least a false discovery (V)

- Family Discovery Rate is:

$$FDR = \mathbb{E}\left[\frac{V}{R}\right]$$

  the expected fraction of false discoveries (V) with respect to the number of discoveries (R)

  - It is the ratio of an observable and an unobservable quantities

BH to control FDR

- Benjamini-Hochberg is one of the most popular correction methods
- ***Algorithm***
- Suppose you do m (independent) tests
- You want to control FDR so that $\mathbb{E}[V/R] < \alpha$
- Calculate p-values of all experiments
- Order the p-values from smallest to largest $p_1, p_2, ..., p_m$
- Find the smallest index corresponding to the p-value that falls under the sloped line $\alpha \frac{i}{m}$
  - This p-value p_T is called the BH rejection threshold
- Call significant any experiment with $p_i < p_T$

BH correction: pros and cons

- Pros
  - Easy to calculate
- Cons
  - Allows for more false positives than Bonferroni correction
  - Might not work well when the hypotheses are not independent

BY to control FDR

- This is an extension of BH method when tests are dependent
- The sloped line is:

$$l_i = \alpha \frac{i}{m} \frac{1}{\sum_{j=1}^{m} \frac{1}{j}}$$

- Since we are dividing for a value that is larger than 1, the sloped line becomes lower and the threshold is more stringent

Multiple testing in python

- `statsmodels.multipletests`

Using normal distribution of z-scores of test statistics

- Make a normal quantile plot of the z-scored test statistics
- The null hypothesis is that the distribution is Gaussian
- If the observed quantiles are more dispersed than the normal quantiles, this is evidence that some of the significant results may be true positives */

Estimating CDF and statistical functional

- Probability
- Random variables
- Mathematical expectation of RVs
- Probability inequalities
- Statistical Inference
  - Definitions
  - Sample mean
  - Sample variance
  - Asymptotics
  - Confidence intervals
  - Hypothesis testing
  - Multiple hypothesis testing
  - **Estimating CDF and statistical functional**
  - Bootstrap

Empirical CDF

- **Problem**
  - Consider X with an unknown CDF $F(x)$
  - We want to estimate $F(x)$ from n samples $X_1, ..., X_n$
- **Solution**
  - Using the frequentist interpretation:

$$F(x) = \Pr(X \leq x) \approx \frac{\#(X \leq x)}{\#\text{attempts}}$$

  - The empirical CDF is defined as:

$$\hat{F}_n(x) = \frac{\sum_{i=1}^{n} I(X_i \leq x)}{n}$$

  - In words, the empirical CDF is a discrete RV putting mass $\frac{1}{n}$ at each value X_i

Convergence of empirical CDF

- For any value x it holds that the empirical CDF is unbiased estimator

$$\mathbb{E}[\hat{F}_n(x)] = F(x)$$
$$\mathbb{V}[\hat{F}_n(x)] = \frac{F(x)(1 - F(x))}{n}$$

- This implies that the empirical CDF $\hat{F}_n$ converges in probability to the true CDF F

Proof of mean of empirical CDF

- Consider the mean of the empirical CDF $\mathbb{E}[\hat{F}_n(x)]$

$$
\begin{aligned}
\mathbb{E}[\hat{F}_n(x)] &= \mathbb{E}[\frac{1}{n}\sum_i I(X_i \leq x)] \text{ (because of def of empirical CDF)} \\
&= \frac{1}{n}\sum_i \mathbb{E}[I(X_i \leq x)] \text{ (by linearity of mean)} \\
&= \frac{1}{n}\sum_i \Pr(X_i \leq x) \text{ (because of mean of indicator var } \mathbb{E}[I(A)] = \Pr(\\
&= \frac{1}{n} \cdot n \Pr(X_i \leq x) \text{ (by definition of CDF)} \\
&= F(x)
\end{aligned}
$$

Proof of variance of empirical CDF

- Consider the variance of the empirical CDF $\mathbb{V}[\hat{F}_n(x)]$

$$
\begin{aligned}
\mathbb{V}[\hat{F}_n(x)] &= \mathbb{E}[(\hat{F}_n(x) - F(x))^2] \text{ (by def of variance and unbias-ndess of empir}\\
&= \mathbb{E}[(\frac{1}{n}\sum I(X_i \le x) - F(x))^2] \text{ (by def of empirical CDF)}\\
&= \frac{1}{n^2}\mathbb{E}[\sum I((X_i \le x) - F(x))^2]\\
&= \frac{1}{n^2}\mathbb{E}[\sum (I(X_i \le x) - F(x))^2] \text{ (since all } X_i \text{ are independent and th}\\
&= \frac{1}{n^2} \cdot n\mathbb{E}[(I(X_i \le x) - F(x))^2]\\
&= \frac{1}{n}\mathbb{E}[(I(X_i \le x))^2 + (F(x))^2 - 2I(X_i \le x)F(x)] \text{ (developing the sq}\\
&= \frac{1}{n}(\mathbb{E}[(I(X_i \le x))^2] + (F(x))^2 - 2\mathbb{E}[(I(X_i \le x))^2]F(x))\\
&= \frac{1}{n}(\mathbb{E}[I(X_i \le x)] + \ldots) \text{ (since the square of indicator var } I^2 = I)\\
&= \frac{1}{n}(F(x) + (F(x))^2 - 2(F(x))^2) \text{ (since } \mathbb{E}[I(X_i \le x)] = \Pr(X \le x)\\
&= \frac{1}{n}(F(x) - (F(x))^2)
\end{aligned}
$$

Statistical functional

- Given a RV F, a statistical functional $T(F)$ is any function of the CDF of F

Statistical functional: number or RV

- The statistical functional $T(F)$:
    - Is a number (e.g., the mean, the median, the variance) if we use a distribution F
    - Is a RV if we use a sample distribution F (since different samples will give different values)

Statistical functional: example

- The following quantities are statistical functionals since they are function of the CDF of F
  - Mean since $\mu = \int x dF$
  - Variance since $\sigma^2 = \int (x - \mu)^2 dF$
  - Median since it is $= F^{-1}(\frac{1}{2})$

Plug-in principle

- Given a statistical functional $T(F)$, the plug-in estimator of $T(F)$ is defined by:
$$\hat{T}_n = T(\hat{F}_n)$$
- In words, to estimate a functional through its samples, we "plug in" the empirical CDF $\hat{F}_n$ into the functional
- Note that it is not a theorem but simply a common sense guideline

Linear statistical functional

- A statistical functional $T(F)$ is linear $\iff$ it is in the form:

$$T(F) = \int r(x)dF(x)$$

  where $r(x)$ is a weighting function
- In words T is a linear combination of values from the PDF dF
- MEM: It is the same form as the theorem of the lazy statistician but using the CDF

Linear statistical functional: examples and non-examples

- Examples:
  - Mean
  - Variance
  - Skewness
- Non-examples:
  - Median
  - Trimmed mean (i.e., the mean without a percent of extreme values)

Plug-in estimator for linear statistical functional

- Applying the plug-in principle for linear statistical functional:

$$
\begin{aligned}
\hat{T}_n &= T(\hat{F}_n) && \text{(def of plug-in principle)} \\
&= \int r(x) d\hat{F}_n(x) && \text{(def of linear statistical functional)} \\
&= \frac{1}{n} \sum_i r(x_i) \hat{F}_n(x_i) && \text{(PMF has mass } \frac{1}{n} F_n(x_i) \text{ in each point)}
\end{aligned}
$$

Bootstrap

- Probability
- Random variables
- Mathematical expectation of RVs
- Probability inequalities
- Statistical Inference
  - Definitions
  - Sample mean
  - Sample variance
  - Asymptotics
  - Confidence intervals
  - Hypothesis testing
  - Multiple hypothesis testing
  - Estimating CDF and statistical functional
  - **Bootstrap**

Bootstrap in brief

- Bootstrap is used to estimate the distribution of a sampling statistic $T = g(X_1, ..., X_n)$ given a finite amount of samples $X_i \sim F$
  - E.g., $g(\cdot)$ can be the mean, median, standard deviation, OLS coefficient, etc.
  - Bootstrap is a generalization of the plug-in principle
- Useful when:
  - The theoretical distribution of the statistic is unknown; or
  - The sample size is too small for traditional parametric methods
- Bootstrap is a **non-parametric** method
  - We don't make any assumption on the distribution we need to estimate
  - In a parametric method we assume that there is a model and we only need to estimate some parameters of the model
- Applications
  - Estimating distribution of sample statistics: e.g., $F_X(x), f_x(x)$
  - Constructing confidence intervals: e.g., $\mu \pm \epsilon$
  - Calculating standard errors: e.g., $\sigma(\hat{T}_n)$

Bootstrap procedure

- Given $X \sim F$
  - Draw n IID samples X_i from X
  - Consider a statistic $T = g(X_1, ..., X_n)$ of the data
  - We want to approximate the distribution of T or estimate a statistic of T (e.g., mean, std err)

Algorithm

- Use the observed data $X_1, ..., X_n$ to construct an estimated population distribution $\hat{F}_n$
  - We pretend that the empirical distribution is the real one
- Repeat B times:
  - Draw n samples with replacement from $\hat{F}_n$
  - Compute the sample statistics from the n samples: $T^{(i)} = g(X_1^{(i)}, ..., X_n^{(i)})$
- Use B samples of $T^{(i)}$ to estimate its empirical distribution $\hat{T}$
- Compute statistics (e.g., confidence interval, standard error) of the statistics T from the empirical distribution of $\hat{T}$

Sample with replacement as computational shortcut in bootstrap

- In theory we need to
  - compute the empirical PMF of $\hat{F}$
  - draw from it
- In practice drawing an observation from $\hat{F}$ is equivalent to drawing one point at random from $X_1, ..., X_n$ (i.e., sample with replacement)

Bootstrap: pros

- Tremendously useful tool
- Fewer assumptions
  - It does not need simplifying assumptions required to get closed formulas
  - E.g., the underlying data does not need to be Gaussian
- Greater accuracy
  - It does not rely on large sample sizes, in contrast with asymptotics from CLT / LLN
- Generality
  - The same method applies to any sample statistics, even difficult non-linear ones (e.g., median)
- Simulation vs math
  - Bootstrap liberated data scientists from performing lots of complex mathematics, approximations, and asymptotics

Bootstrap: example of die rolls

- We want to compute the distribution of the sum of rolling a die 50 times

$$T = \sum_{i=1}^{50} X_i = g(X_1, ..., X_{50})$$

- We know the PMF of the die, including the probability of head p

1. By math
   - We can compute the distribution using mathematics
     - Compute PDF or use theorem of lazy statistician
2. By sampling (real or simulated)
   - Repeat the procedure enough to get convergence of the PDF
     - Roll the die 50 times
     - Compute the sample statistic T
   - Plot the approximate distribution of T

- What if we don't know anything else than just 50 samples of the die?
- Bootstrap
  - Sample from the empirical distribution with replacement
  - Build the distribution of T

Pseudo-code for bootstrap of the median

```
def bootstrap_median(x, n_boot):
    # Compute n_boot sample statistics.
    median_boot = [0.0] * n_boot
    for i in range(n_boot):
        # Sample with replacement.
        x_star = sample with replacements from x
        # Compute median for bootstrapped samples.
        median_boot[i] = median(x_star)
    # Compute mean and std err from approximation of
        sample statistics.
    m_median = numpy.mean(median_boot)
    se_median = numpy.std(median_boot)
    return m_median, se_median
```

Bootstrap for variance of sample statistics: explanation

- Assume:
  - $X \sim F$
  - n IID samples X_i from F
  - Compute a statistic of the data $T = g(X_1, ..., X_n)$
- We want to compute $\mathbb{V}_F[T]$ (variance of sample statistics), where subscript F indicates that it depends on the unknown distribution F
- There are 2 approximations to compute $\mathbb{V}_F[T]$
- First approximation
  - We don't have F, but only samples drawn from it
  - We can approximate the distribution of F with the distribution of $\hat{F}$, since we know that the empirical CDF $\hat{F}$ converges to the true CDF F (plug-in principle)

$$\mathbb{V}_F[T] \approx \mathbb{V}_{\hat{F}}[T]$$

  - This approximation
    - Is not so small
    - Depends on the number of samples and shape of F

Bootstrap for variance of sample statistics (1/3)

- Under the hypotheses of bootstrap:
  - $X \sim F$
  - n IID samples X_i from F
  - Compute a statistic of the data: $T = g(X_1, ..., X_n)$
- We want to estimate $\mathbb{V}_F[T]$, the variance of the statistic under the true (unknown) distribution F.
- There are two approximations used to estimate $\mathbb{V}_F[T]$:
  - First approximation (Plug-in Principle)
    - We approximate the true CDF with empirical CDF
  - Second approximation (Monte Carlo Simulation)

Bootstrap for variance of sample statistics (2/3)

First approximation (Plug-in Principle)

- We don't know F, but we observe samples drawn from it
- Approximate F with the empirical distribution $\hat{F}$:
  - $\hat{F}$ assigns probability mass $1/n$ to each observed X_i
  - The empirical CDF $\hat{F} \to F$ converges to the true CDF F uniformly almost surely
- Using the plug-in principle:

$$\mathbb{V}_F[T] \approx \mathbb{V}_{\hat{F}}[T]$$

- This approximation error depends on:
  - The number of samples n
  - The shape of the true distribution F

Bootstrap for variance of sample statistics (3/3)

Second approximation (Monte Carlo Simulation)

- Even with $\hat{F}$ known, $\mathbb{V}_{\hat{F}}[T]$ might not have a closed-form expression
- Use bootstrap resampling: draw B samples $T_1, ..., T_B$ by resampling with replacement from the data
- Then compute:

$$\overline{T} = \frac{1}{B}\sum_{i=1}^{B} T_i$$

$$v_{\text{boot}} = \frac{1}{B}\sum_{i=1}^{B}(T_i - \overline{T})^2$$

- This estimate converges:

$$v_{\text{boot}} \xrightarrow{P} \mathbb{V}_{\hat{F}}[T] \quad \text{as } B \to \infty$$

- The accuracy of v_{boot} improves with larger B

Bootstrap for variance of sample statistics: analytical formula

- Assume:
  - $X \sim F$
  - n IID samples X_i from F
  - Compute a statistic of the data: $T = g(X_1, ..., X_n)$
- We want to compute $\mathbb{V}_F[T]$ (variance of sample statistics), where subscript F indicates that it depends on the unknown distribution F
- In some special cases we have a formula for $\mathbb{V}_F[T]$ using F
  - E.g., for sample statistic $T = \frac{1}{n}\sum_i X_i$ (sample mean) and X Gaussian:

$$\mathbb{V}_F[T] = \frac{\mathbb{V}[F]}{n}$$

- If we don't have $\mathbb{V}[F]$ we can use $\hat{F}$ to compute an approximation of $\mathbb{V}_{\hat{F}}[F]$

$$S^2 = \frac{1}{n}\sum_i (X_i - \overline{X})^2$$

- This is a case where we replace the first approximation of bootstrap with

Bootstrap confidence intervals

- In the general case:
  - We compute the empirical distribution of T by bootstrapping
  - Estimate the confidence intervals from the CDF using percentiles
- In some cases by CLT we know that $T = g(X_1, ..., X_n)$ tends to a Gaussian
- We can assume that T is Gaussian and thus confidence intervals are:

$$\overline{T} \pm Z_{a/2} se_{boot}$$

where se_{boot} is the sqrt of variance estimated through bootstrap

Bootstrap hypothesis testing: example

- We have two samples of data A and B of different lengths
- Check if a sample statistics (e.g., the median) of A and B are different
- ***Algorithm***
- We cannot do a paired test since A and B are not paired
- We can test if the difference of sample statistics is different enough from 0
- In other words the test statistic is the difference of medians
- We bootstrap the difference of the medians from the data