Theory of Probability to Bayesian Inference

2022/02/18

Sigma Algebra and Measure Spaces

[From: Schuller QM Lecture 5]

Defn: Let M be a non-empty set. Then a collection of subsets $\sigma \subseteq \mathcal{P}(M)$ [where $\mathcal{P}(M)$ is the Power Set of M, i.e., the set of all subsets of M] is called a σ-algebra of M if:

(i) $M \in \sigma$

(ii) $A \in M \Rightarrow M \setminus A \in \sigma$ ("M without A" i.e., the complement of A in M, A^c)

(iii) $\bigcup_{n \in \mathbb{N}} A_n \in \sigma$ for any sequence $\{A_n\}_{n \in \mathbb{N}}$ with $A_n \in \sigma \ \forall n$.

Can show that the "de Morgan Rule" of set theory: $\bigcap_{n \geq 1} M \setminus A_n = M \setminus \bigcup_{n \geq 1} A_n$ implies that countable intersections are measurable: $A = \bigcap_{n \geq 1} A_n \in \sigma$

- $A \in \sigma$ is called a Measurable Set in M
- The pair (M, σ) is a Measurable Space.

Defn: Given a measureable space (M, σ), a Measure $\mu : \sigma \to (\overline{\mathbb{R}})_0^+$ is a map satisfying:

($\overline{\mathbb{R}} = (\mathbb{R} \cup \{-\infty, \infty\})$ "Extended Real Number Line")

(i) $\mu(\emptyset) = 0$

(ii) For a sequence of pairwise disjoint measurable sets $A_1, A_2, \ldots \in \sigma$ (i.e., $(A_i \cap A_j = \emptyset \ \forall \ i \neq j)$):

$$\mu\left(\bigcup_{n \geq 1} A_n\right) = \sum_{n \geq 1} \mu(A_n)$$ "σ-additivity"

- The triple (M, σ, μ) is a Measure Space
- Properties of a measure:

(i) Monotony: $A_1 \subseteq A_2 \Rightarrow \mu(A_1) \leq \mu(A_2)$

(ii) Sub-additivity:

$$\mu\left(\bigcup_{n \geq 1} A_n\right) \leq \sum_{n \geq 1} \mu(A_n)$$ (No Pairwise-disjoint here)

(iii) Continuity from below:

$A_1 \subseteq A_2 \subseteq \cdots$ w/ $A := \bigcup_{n \geq 1} A_n \Rightarrow \lim_{n \to \infty} \mu(A_n) = \mu(A)$

(iv) Continuity from above

$A_1 \supseteq A_2 \supseteq \cdots$ w/ $A := \bigcap_{n \geq 1} A_n$ and $\mu(A_1) < \infty$

Axioms of Probability and their consequences

- **Defn**: A measure space $(M, \mathcal{E}, P)$ is called a **probability space** with M the **sample space**, $\mathcal{E}$ the **event space**, and $P(E)$ the **probability** of event $E \in \mathcal{E}$, when the **probability measure**, $P: \mathcal{E} \to (\overline{\mathbb{R}})^+_0$ satisfies:

$$P(M) = 1$$

[Thus, apart from the definition of a measure space, the only axiom is "total measure one"]

- The so-called **Axioms of Probability** follow from the above-defined properties of a measure space:

1. $P(E) \geq 0 \quad \forall \; E \in \mathcal{E}$ (defn of a measure)
2. $P(M) = 1$ (the only real axiom here)
3. $P\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} P(E_i)$, for pairwise disjoint $\{E_1, \ldots E_n\}$ (σ-additivity of a measure)

- **Consequences**

  - **Monotonicity**: $A \subseteq B \Rightarrow P(A) \leq P(B)$

$$\begin{aligned} B &= B \cap M \\ &= B \cap (A \cup A^c) \\ &= (B \cap A) \cup (B \cap A^c) \quad \text{Now a disjoint union!} \\ &= A \cup (B \cap A^c) \quad \begin{matrix} \swarrow A \subseteq B \\ \searrow \sigma\text{-additivity} \end{matrix} \end{aligned}$$

$$\Rightarrow P(B) = P(A) + P(B \cap A^c) \geq P(A) \; \checkmark$$

  - **Probability of the complement**: $P(E^c) = 1 - P(E)$

$$M = E \cup E^c \quad \text{disjoint union}$$

$$P(M) = P(E) + P(E^c) \Rightarrow 1 = P(E) + P(E^c) \; \checkmark$$

  - **Probability of the empty set**: $P(\phi) = P(M^c) = 1 - P(M) = 0$

Empty set need not be the only event w/ $P(E) = 0$

  - **Target of the probability measure**: $P: \mathcal{E} \to [0, 1]$

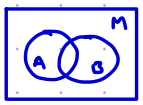
$$P(E^c) = 1 - P(E) \geq 0 \Rightarrow P(E) \leq 1$$

$$\Rightarrow 0 \leq P(E) \leq 1 \quad \forall \; E \in \mathcal{E}$$

"Addition Law of Probability"

$$\boxed{P(A \cup B) = P(A) + P(B) - P(A \cap B)}$$

We can write



$$\rightsquigarrow \quad (A \cup B) = (A \cap B) \cup (B^c \cap A) \cup (A^c \cap B)$$

$$A = (A \cap B) \cup (A \cap B^c)$$

$$B = (B \cap A) \cup (B \cap A^c)$$

} All disjoint unions

$$P(A \cup B) = P(A \cap B) + P(B^c \cap A) + P(A^c \cap B)$$

$$P(A) = P(A \cap B) + P(A \cap B^c)$$

$$P(B) = P(B \cap A) + P(B \cap A^c)$$

$$\Longrightarrow \quad P(A \cup B) = P(A) + [P(B) - P(B \cap A)] \quad \checkmark$$

Science of Probability: The probability space is associated to a well-defined "experiment" with $o \in M$ each of the possible elementary outcomes of that experiment. An event $E \in \mathcal{E}$ is an elementary or derived/composite outcome. The probability of an event occurring is $P(E)$.

Ross's First Course in Probability, and Kolmogorov's Found. of Probability

Examples:

Experiment: Flipping a coin twice

Sample space: $M = \{(H,H), (H,T), (T,H), (T,T)\}$
$= \{o_1, o_2, o_3, o_4\}$

Example events and probabilities

- First coin is heads: $E = \{o_1, o_2\}$
  $P(E) = \underbrace{P(o_1) + P(o_2)}_{\text{disjoint events}} = \underbrace{\frac{1}{4} + \frac{1}{4}}_{\text{assuming fair \& indep.}} = \frac{1}{2}$
- One coin is tails: $E = \{o_2, o_3, o_4\}$, $P(E) = \frac{3}{4}$

These discrete examples assume that $o \in M \Rightarrow o \in \mathcal{E}$ i.e., the elementary outcomes are measurable, and that they are equally-probable

Experiment: Placing in a horse race among 7 horses

Sample space: $M = \{$the $7!$ permutations of $1234567\}$

Events:

- Horse 4 places 1st: $E = \{6!$ perm of $\boxed{4}123567\}$, $P(E) = \frac{1}{7}$

EXPERIMENT: ROLLING TWO DICE IN SEQUENCE

SAMPLE SPACE: $M = \{(1,1), (1,2) \dots \langle 36 \text{ ELEMENTS} \rangle\}$

EVENTS:

- A TOTAL OF 7: $E = \{(1,6),(6,1),(3,4),(4,3),(2,5),(5,2)\}$

$P(E) = 6/36 = 1/6$

THIS TYPE OF CONTINUOS EXAMPLE WAS UNIFIED W/ THE DISCRETE CASE BY KOLMOGOROV. THERE IS NO MEASURE ASSOCIATED TO $x \in M$, I.E., FOR $x \in \mathbb{R}$ $x \notin \mathcal{E}$. POINTS ARE NOT MEASURABLE, BUT: $(a,b), (a,b], [a,b] \in \mathcal{E}$

EXPERIMENT: MEASURING THE LIFETIME OF A LIGHTBULB

SAMPLE SPACE: $M = \mathbb{R}_0^+$ (HOURS)

EVENTS:

- BULB DOES NOT LAST 12h: $E = [0, 12)$, $P(E) = ?^*$

THIS IS DEFINITELY WRONG: $x = (a,x] \cap [x,b)$ (COUNTABLE INTERSECTIONS ARE MEASUREABLE). BUT I'LL BET WE ASSIGN: $\mu(x) = 0$

TERMINOLOGY:

THEORY OF SETS	EVENTS IN PROBABILITY EXPT.
$A \cap B = \phi$ OR $A_1 \cap A_2 \cap \dots \cap A_n$	EVENTS ARE INCOMPATIBLE (MUTUALLY-EXCLUSIVE, INDEPENDENT?)
$E = A_1 \cap A_2 \cap \dots \cap A_n$	EVENT E IS THE SIMULTANEOUS OCCURANCE OF THE EVENTS $\{A_i\}$
$E = A_1 \cup A_2 \cup \dots \cup A_n$	EVENT E IS THE OCCURANCE OF AT LEAST ONE OF $\{A_i\}$
COMPLEMENT OF A, A^c	THE NON-OCCURANCE OF A
$A = \phi$	EVENT A IS IMPOSSIBLE
$A = M$	EVENT A MUST OCCUR
THE PARTITION* OF M BY $\mathcal{A} = \{A_1, A_2, \dots, A_n\} \subseteq \mathcal{E}$	ONE AND ONLY ONE EVENT A_i WILL OCCUR
$B \subset A$	THE OCCURANCE OF B IMPLIES THE OCCURANCE OF A.

* — THE SETS ARE PAIRWISE DISJOINT, $A_i \cap A_j = \phi \ \forall \ i,j$ AND $\bigcup_i A_i = M$.

ALGEBRA OF SETS (FOR $A, B, C \subseteq M$, WITH ϕ THE EMPTY SET)

SINCE A PROBABILITY SPACE HAS A SET AS ITS BASE STRUCTURE.., LET US REVIEW (WIKIPEDIA)

- UNION AND INTERSECTION ARE COMMUTATIVE:

$$A \cup B = B \cup A \quad \text{AND} \quad A \cap B = B \cap A$$

- UNION AND INTERSECTION ARE ASSOCIATIVE

$$A \cup (B \cup C) = (A \cup B) \cup C \quad \text{AND} \quad \langle \text{SAME FOR } \cap \rangle$$

WARNING: THIS "MAKES SENSE" W/ $a \times (b+c) = a \times b + a \times c$ BUT THIS DOES NOT! $a + (b \times c) \neq (a+b) \times (a+c)$

- DISTRIBUTIVE PROPERTIES

$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$$
$$A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$$

- IDENTITIES

$$A \cup \phi = A \quad \text{AND} \quad A \cap M = A$$

- COMPLEMENTS

$$A \cup A^c = M \quad \text{AND} \quad A \cap A^c = \phi$$

WE COULD ALSO PROVE DEMORGAN'S LAWS:

$$\left(\bigcup_{i=1}^{n} A_i\right)^c = \bigcap_{i=1}^{n} A_i^c \quad \text{AND} \quad \left(\bigcap_{i=1}^{n} A_i\right)^c = \bigcup_{i=1}^{n} A_i^c$$

CONDITIONAL PROBABILITIES AND BAYES' THEOREM

2022/02/19

- DEFN: THE CONDITIONAL PROBABILITY OF EVENT A UNDER CONDITION B IS DEFINED (WHEN $P(B) > 0$):

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

(KOLMOGOROV USES: $P_B(A) = P(A|B)$ WHICH IS NICER BECAUSE $P_B(A)$ IS THE "PROBABILITY OF A".)

IN EXPERIMENTAL TERMS, THIS IS THE PROBABILITY THAT EVENT A OCCURS UNDER THE CONDITION THAT EVENT B ALSO OCCURS.

- THE MULTIPLICATION THEOREM FOLLOWS FROM INDUCTION OF $P(A \cap B) = P(A) \cdot P(B|A)$:

NOTE, E.G., $P(A \cap B) = P(A)\,P(B|A)$; $P(A \cap B \cap C) = P((A \cap B) \cap C) = P(A \cap B) \cdot P(C|A \cap B) = P(A) \cdot P(B|A) \cdot P(C|A \cap B)$

$$P(A_1 \cap A_2 \cap \cdots \cap A_n) = P(A_1)\,P(A_2|A_1)\,P(A_3|A_1 \cap A_2) \cdots \times P(A_n|A_1 \cap A_2 \cap \cdots \cap A_{n-1})$$

- FOR FIXED EVENT B, $(M, \mathcal{E}, P(\cdot|B))$ IS A PROBABILITY SPACE SINCE:

$P(\cdot|B)$ SATISFIES σ-ADDITIVITY AND $P(M|B) = 1$ ✓

$$P(M|B) = P(M \cap B)/P(B) = P(B)/P(B) = 1$$

AND FOR DISJOINT $\{A_1, \ldots, A_n\}$

$$P\left(\bigcup_{i=1}^{n} A_i \,\middle|\, B\right) = \frac{P((\cup A_i) \cap B)}{P(B)} = \frac{P((A_1 \cap B) \cup \cdots \cup (A_n \cap B))}{P(B)}$$
$$= \sum \frac{P(A_i \cap B)}{P(B)} = \sum_{i=1}^{n} P(A_i|B)$$

THE THEOREM OF BAYES: THE DEFINITIONS OF $P(A|B)$ AND $P(B|A)$ ALLOW US TO WRITE (USING $P(A \cap B) = P(B \cap A)$):

$$P(A|B)\,P(B) = P(B|A)\,P(A)$$

OR

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

IF WE THEN CONSIDER A PARTITION OF M, $\mathcal{A} = \{A_1, \ldots, A_n\}$, WE CAN WRITE:

WE COULD USE A PARTITION OF B HERE, RIGHT? $B = B \cap B$

$$B = B \cap M = B \cap \left(\bigcup_{i=1} A_i\right)$$
$$= (B \cap A_1) \cup (B \cap A_2) \cup \cdots \cup (B \cup A_n)$$

AND THEREFORE: (UNION OF DISJOINT SETS)

$$P(B) = \sum_{i=1}^{n} P(B \cap A_i) = \sum_{i=1}^{n} P(B|A_i)\,P(A_i)$$

AND SO, FOR ANY A_i WE CAN WRITE:

$$P(A_i|B) = \frac{P(B|A_i)\,P(A_i)}{P(B|A_1)\,P(A_1) + \cdots + P(B|A_n)\,P(A_n)}$$

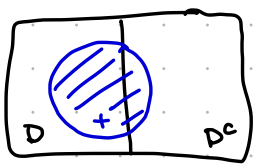
[FOR THE PARTITION OF M, $\mathcal{A} = (A_1, \ldots, A_n)$]

WHICH IS CALLED BAYES' THEOREM.

- THE EVENTS $A_i \in \mathcal{A}$ ARE USUALLY CALLED HYPOTHESES, AND THE ABOVE EQUATION GIVES THE PROBABILITY OF THE A_i HYPOTHESIS GIVEN THAT EVENT B HAS OCCURED.

- MORE TERMINOLOGY:
  - $P(A_i)$ IS THE PRIOR PROBABILITY OF HYPOTHESIS A_i (INITIAL ASSUMPTIONS ARE MODIFIED GIVEN NEW EVIDENCE)
  - $P(A_i|B)$ IS THE POSTERIOR PROBABILITY (ONCE B HAS OCCURED)
  - DENOMINATOR $\sum_i P(B|A_i)\,P(A_i)$ IS A NORMALIZING CONST (SUM OF THE TERMS IN THE NUMERATOR)

- EXAMPLE: PROBABILITY OF HAVING DISEASE (D) GIVEN A POSITIVE TEST ($+$). $\mathcal{A} = \{D, D^c\}$, $M = D \cup D^c = \{+\} \cup \{+^c\}$



$$P(D|+) = \frac{P(+|D) \cdot P(D)}{P(+|D)\,P(D) + P(+|D^c)\,(1 - P(D))}$$

WHERE $P(D)$ IS THE PREVALENCE OF DISEASE IN THE COMMUNITY, $P(+|D)$ IS THE SENSITIVITY OF THE TEST AND $(1 - P(+|D^c))$ IS THE SPECIFICITY OF THE TEST.

As a connection to application, call the partition the <u>hypotheses</u> $\mathcal{H} = \{H_1, H_2, \ldots, H_3\}$ call an observed event <u>evidence</u>, E. Given a <u>prior probability of hypothesis H_i</u>, $P(H_i)$, we find the <u>updated posterior probability of H_i, given new evidence</u>,

$$P(H_i \mid E) = \frac{P(E \mid H_i) \cdot P(H_i)}{\sum_j P(E \mid H_j) P(H_j)}$$

Where $P(E \mid H_i)$ is sometimes called the <u>likelihood</u> of the hypothesis.

As we will see later, $P(E|H)$ is also called the <u>likelihood function</u>, i.e., when viewed as a function of H, with fixed E:

$$L(H) = P(E|H)$$

<u>Example</u>: A detective is 50% confident of a particular suspect's guilt ($P(H) = 0.5$). New evidence shows that the perpetrator is left-handed ($P(E|H) = 1$). If the particular suspect is left-handed (like 19% of the population, $P(E|H^c) = 0.19$):

H = Guilty (G)

H^c = Not Guilty (NG)

E = Left-Handed (L)

$$P(G \mid L) = \frac{P(L|G) \cdot P(G)}{P(L|G)P(G) + P(L|NG)P(NG)}$$

$$= \frac{1 \cdot 0.5}{1 \cdot 0.5 + 0.19 \cdot 0.4} = 87\%$$

Experiment here involves the properties of the <u>suspect</u>, so:

$P(L) \neq P(L|NG)$

(left-handedness of population)

<u>Example</u>: 2-child family moves into neighborhood. You meet mother & daughter. What is probability other child is a girl?

$M = \{GG, GB, BB\}$

g = met one girl

$$P(GG \mid g) = \frac{P(g|GG) \cdot P(GG)}{P(g|GG) \cdot P(GG) + P(g|GB) \cdot P(GB) + P(g|BB) \cdot P(BB)}$$

Assuming that it is equally likely for the mother to be w/ any child and that the couple's <u>sequence</u> of offspring is equally likely to be $\{GG, GB, BG, BB\}$:

$$P(GG|g) = \frac{1 \cdot 1/4}{1/4 + 1/2 \cdot 1/2 + 0 \cdot 1/4} = \frac{1/4}{1/4 + 1/4} = \frac{1}{2}$$

There are 6 equally likely outcomes here:

$R_F(RR)$, $R_B(RR)$, $R(RG)$

$B(RG)$, $B_F(BB)$, $B_B(BB)$

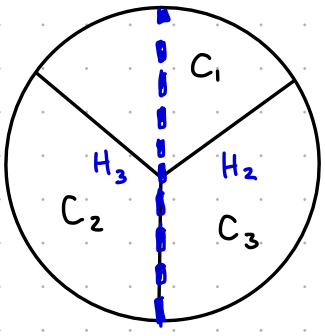
So 2 of 3 "R" options are from the RR child.

<u>Example</u>: 3 two-sided cards w/ RR (red-red), RB (red-black) BB (black-black) on front-back. Given a red is shown, what chance is it RR?

$$P(RR|R) = \frac{P(R|RR) \cdot P(RR)}{P(R|RR) \cdot P(RR) + P(R|RB) P(RB) + P(R|BB) \cdot P(BB)} = \frac{1 \cdot 1/3}{1/3 + 1/2 \cdot 1/3} = \frac{2}{3}$$

THIS PROBLEM BECAME FAMOUS AFTER APPEARING IN THE "ASK MARILYN" ADVICE COLUMN (PARADE MAGAZINE, 1990/9/9). THE SOLUTION GIVEN WAS, ESSENTIALLY:

THE PROBABILITY OF YOUR SELECTION DID NOT CHANGE (FROM 1/3) JUST BECAUSE THE HOST OPENS A DOOR. BUT THE PROBABILITY OF THE OTHER TWO DOORS (2/3) ALSO DIDN'T CHANGE ... IT IS JUST ALL IN THE UNOPENED.



EXAMPLE (MONTY HALL): BEHIND THREE DOORS ARE A CAR AND THREE GOATS. YOU CHOOSE DOOR 1. THE GAME SHOW HOST OPENS ANOTHER DOOR, SHOWS YOU A GOAT AND OFFERS YOU THE CHANCE TO CHANGE YOUR CHOICE. SHOULD YOU?

$(C_1 \cup C_2 \cup C_3 = M)$

$$P(\text{car in } 1 \mid \text{Host opens } 2) = P(C_1 \mid H_2)$$

$$= \frac{P(H_2|C_1)\cdot P(C_1)}{P(H_2|C_1)P(C_1) + P(H_2|C_2)P(C_2) + P(H_2|C_3)P(C_3)}$$

$$= \frac{1/2 \cdot 1/3}{1/2 \cdot 1/3 + 0 \cdot 1/3 + 1 \cdot 1/3}$$

$$= \frac{1/6}{1/6 + 2/6} = \frac{1}{3}$$

YES, YOU SHOULD CHANGE! CONSIDER ALSO:

$$P(C_3 \mid H_2) = \frac{P(H_2|C_3)P(C_3)}{P(H_2|C_3)P(C_3) + P(H_2|C_2)P(C_2) + P(H_2|C_1)P(C_1)}$$

$$= \frac{1 \cdot 1/3}{1/3 + 0 + 1/2 \cdot 1/3} = \frac{2}{3}$$

AND:

$$P(C_1 \mid H_3) = \frac{P(H_3|C_1)P(C_1)}{P(H_3|C_1)P(C_1) + P(H_3|C_2)P(C_2) + 0}$$

$$= \frac{1/2 \cdot 1/3}{1/2 \cdot 1/3 + 1 \cdot 1/3} = \frac{1}{3}$$

SO, WHICHEVER DOOR THE HOST OPENS, YOUR CURRENT DOOR ONLY HAS 1/3 PROBABILITY OF CONTAINING THE CAR, WHILE THE OTHER CLOSED DOOR HAS 2/3 PROB.

DEFN: THE ODDS RATIO OF A PARTICULAR HYPOTHESIS H IS:

$$OR(H) := \frac{P(H)}{P(H^c)} = \frac{P(H)}{1 - P(H)}$$

IF, FOR EXAMPLE $P(H) = 3/4$ THEN $OR(H) = 3$. ONE SAYS "THE ODDS ARE 3 TO 1 IN FAVOR".

THE INTRODUCTION OF NEW EVIDENCE CAN CHANGE THE ODDS RATIO VIA BAYES THEOREM:

$$OR(H|E) = \frac{P(H|E)}{P(H^c|E)} = \frac{P(E|H)}{P(E|H^c)} \left(\frac{P(H)}{P(H^c)}\right)$$

$OR(H)$

THIS WILL BE CALLED THE BAYES FACTOR

SO THE RATIO OF THE LIKELIHOODS OF THE EVIDENCE GIVEN THE HYPOTHESIS OR ITS COMPLEMENT YIELDS THE FACTOR BY WHICH THE ODDS RATIO CHANGES.

- EX: IN THE MONTY HALL PROBLEM:

"ODDS ARE 1 TO 2 AGAINST FINDING THE CAR IN 1" (ALTHOUGH SAME FOR ANY DOOR)

$$OR(\text{CAR IN } 1) = \frac{P(1)}{P(1^c)} = \frac{1/3}{2/3} = \frac{1}{2}$$

BUT

$$OR\left(\text{CAR IN } 1 \,\middle|\, \begin{matrix}\text{HOST OPENS}\\ \text{DOOR 2}\end{matrix}\right) = \frac{P(H_2|1)}{P(H_2|1^c)} \cdot OR(1)$$

$$OR(1|H_2) = \frac{1/2}{1} \cdot OR(1) = \frac{1}{2}\cdot\frac{1}{2} = \frac{1}{4}$$

"ODDS ARE 1 TO 4 AGAINST"

SO THE ODDS HAVE DECREASED!

A MODEL OF A SYSTEM, m, COULD BE CONSIDERED A HYPOTHESIS, AND DATA ACCUMULATED, D, COULD BE CONSIDERED EVIDENCE IN FAVOR (OR AGAINST) SUCH A MODEL:

POSTERIOR → $P(m|D) = \dfrac{P(D|m)\cdot P(m)}{P(D)}$ ← PRIOR

WHERE WE AGAIN CALL $P(D|m)$ THE LIKELIHOOD OF MODEL GIVEN THE DATA. IF THE MODEL DEPENDS ON PARAMETERS, $\vec{\theta}$, WE CAN INCLUDE DEPENDENCE ON THE VALUES OF THOSE PARAMETERS:

$$P(m\cap(\vec{\theta}=\vec{\theta}^*)|D) = \frac{P(D|m\cap(\vec{\theta}=\vec{\theta}^*))\cdot P(m\cap(\vec{\theta}=\vec{\theta}^*))}{P(D)}$$

DEFN: IN THE COMPARISON OF TWO MODELS, THE BAYES FACTOR IS THE RATIO OF THEIR LIKELIHOODS:

$$K(m_1,m_2) := \frac{P(D|m_1)}{P(D|m_2)} = \frac{P(m_1|D)}{P(m_2|D)}\cdot\frac{P(m_2)}{P(m_1)}$$

WHICH IS THE RATIO OF THEIR POSTERIOR PROBABILITIES IF THE MODELS ARE INITIALLY EQUALLY LIKELY.

INTERPRETATION OF BAYES FACTOR (JEFFREYS, 1961)

K	SIGNIFICANCE
$< 10^0$	NEGATIVE (SUPPORTS m_2)
10^0–$10^{1/2}$ (1–5)	BARELY ANYTHING
$10^{1/2}$–10^1 (5–10)	SUBSTANTIAL
10^1–$10^{3/2}$ (10–15)	STRONG
$10^{3/2}$–10^2 (15–20)	VERY STRONG
$>10^2$ (>20)	DECISIVE

IF $P(D \mid m \cap (\vec{\theta} = \hat{\theta}))$ IS USED, WHERE $\hat{\theta}$ IS THE MAXIMUM LIKELIHOOD ESTIMATE OF THE PARAMETER, THEN THIS BECOMES THE LIKELIHOOD TEST OF CLASSIC (FREQUENTIST) STATISTICS. BUT FOR THE BAYES FACTOR, IT IS ASSUMED THAT, FOR A MODEL THAT DEPENDS ON PARAMETERS:

$$P(D \mid m) = \sum P(D \mid m \cap \theta_i) \cdot P(\theta_i \mid m)$$

WHERE THE EVENT $\theta_i := (\vec{\theta} = \vec{\theta}_i)$ AND THE SUM IS OVER ALL POSSIBLE VALUES OF THE PARAMETER VECTOR (WHICH IS A PARTITION).

WE KNOW:
$P(E) = \sum P(E \mid A_i) P(A_i)$
FOR A PARTITION $\{A_i\}$. BUT $P(D \mid m)$ FOR FIXED m IS A PROBABILITY ITSELF... CALL IT (A LA KOLMOGOROV):
$P_m(D) := P(D \mid m)$
THEN:
$P_m(D) = \sum P_m(D \mid A_i) P_m(A_i)$
WHICH IS, UNWRAPPING...
$P(D \mid M) = \sum P(D \mid M \cap A_i) \cdot P(A_i \mid M)$

INDEPENDENT TRIALS AND INDEPENDENT EVENTS

3/10-18/2022

- DEFN: GIVEN A PROBABILITY SPACE $(M, \mathcal{E}, P)$, A PARTITION $\mathcal{A} = \{A_i\}$ OF M INTO DISJOINT EVENTS $\{A_i\}$ IS CALLED A TRIAL. THIS CAN GENERALLY BE THOUGHT OF AS A SPECIFIC QUESTION FOR WHICH THE DISJOINT EVENTS PROVIDE ALL POSSIBLE ANSWERS.

I'M CALLING KOLMOGOROV'S "EXPERIMENTS" TRIALS. SEEMS THAT THE MODERN EQUIVALENT IS "CLASSES" SEE math.stackexchange.com/questions/4402669

EXAMPLES

- TWO FAIR COINS TOSSED: FIRST COIN HEADS?

A = FIRST COIN IS HEADS

$$\mathcal{A} = \{A, A^c\}$$

- TWO FAIR COINS TOSSED: HOW MANY HEADS?

A_i = i HEADS APPEARED

$$\mathcal{A} = \{A_0, A_1, A_2\}$$

- FIVE CARDS DEALT: WHAT KIND OF POKER HAND?

$$\mathcal{A} = \{(\text{HIGH CARD}), (\text{ONE PAIR}), (\text{TWO PAIR}), \ldots, (\text{RSF})\}$$

FOR A SET OF n TRIALS, $(\mathcal{A}^{(1)}, \mathcal{A}^{(2)}, \ldots, \mathcal{A}^{(n)})$, WITH

$$\mathcal{A}^{(i)} = \{A_1^{(i)}, A_2^{(i)}, \ldots, A_{r_i}^{(i)}\}$$

THERE ARE $r = r_1 r_2 \cdots r_n$ PROBABILITIES FOR THE OUTCOME OF THOSE TRIALS (FOR THE n ANSWERS TO THE n QUESTIONS ASKED):

$$P_{q_1 \cdots q_n} := P\left(A^{(1)}_{q_1} \cap A^{(2)}_{q_2} \cap \cdots \cap A^{(n)}_{q_n}\right) \quad ; \quad \sum_{\vec{q}} P_{\vec{q}} = 1$$

- DEF'N: A SET OF n TRIALS $(\mathcal{A}^{(1)}, \ldots, \mathcal{A}^{(n)})$ ARE MUTUALLY INDEPENDENT TRIALS IF FOR ANY ALLOWED $(q_1, \ldots, q_n)$ THE FOLLOWING EQUATION HOLDS TRUE:

$$P_{q_1 q_2 \cdots q_n} = P\left(A^{(1)}_{q_1}\right) \cdot P\left(A^{(2)}_{q_2}\right) \cdot \cdots \cdot P\left(A^{(n)}_{q_n}\right).$$

I DON'T SEE THIS BUT HAVEN'T THOUGHT TOO LONG...

AMONG THESE $r = r_1 \cdot r_2 \cdot \cdots \cdot r_n$ EQUATIONS THERE ARE ONLY $\left(r - \sum_i r_i + (n-1)\right)$ INDEPENDENT EQUATIONS.

- THEOREM: IF n TRIALS ARE INDEPENDENT, THEN ANY $m < n$ OF THEM ARE ALSO INDEPENDENT.

- DEF'N: A SET OF n EVENTS $(A_1, \ldots, A_n)$ ARE MUTUALLY INDEPENDENT EVENTS IF THE TRIALS:

$$\mathcal{A}^{(i)} := \left\{A_i, A_i^c\right\} \qquad (i = 1, \ldots, n)$$

ARE MUTUALLY INDEPENDENT TRIALS.

- FROM THE ABOVE WE SEE THAT $r_i = 2 \Rightarrow r = 2^n$ AND THEREFORE THERE ARE $2^n - 2n + n - 1 = 2^n - n - 1$ INDEPENDENT EQUATIONS WHICH CAN BE WRITTEN:

I'M NOT SURE HOW TO COUNT THESE TO GET $2^n - n - 1$ BUT HAVEN'T THOUGHT MUCH

$$P(A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_m}) = P(A_{i_1}) \cdot P(A_{i_2}) \cdot \cdots \cdot P(A_{i_m})$$

$$\text{w/} \quad 1 \le i_1 < i_2 < \cdots < i_m \le n \qquad (m = 1, 2, \ldots, n)$$

- FOR TWO INDEPENDENT EVENTS THERE IS ONE EQN:

$$P(A_1 \cap A_2) = P(A_1) \cdot P(A_2)$$

WHICH IS AN INDEPENDENT EQN OF THE FOUR:

$\mathcal{A}_1 = \{A_1, A_1^c\}$

$\mathcal{A}_2 = \{A_2, A_2^c\}$

$$P(A_1 \cap A_2) = P(A_1) \cdot P(A_2)$$
$$P(A_1^c \cap A_2) = P(A_1^c) \cdot P(A_2)$$
$$P(A_1 \cap A_2^c) = P(A_1) \cdot P(A_2^c)$$
$$P(A_1^c \cap A_2^c) = P(A_1^c) \cdot P(A_2^c)$$

CONSIDER, FOR EXAMPLE, THE LHS OF THE 3RD EQN:

$$\begin{aligned} P(A_1 \cap A_2^c) &= P(A_1 \cap \underbrace{(A_2 \cup A_2^c)}_{M}) - P(A_1 \cap A_2) \\ &= P(A_1) - P(A_1 \cap A_2) \quad \text{(USING 1ST EQN)} \\ &= P(A_1) - P(A_1) \cdot P(A_2) \\ &= P(A_1)\,[1 - P(A_2)] \\ &= P(A_1)\,P(A_2^c) \end{aligned}$$

SO WE SEE THAT EQN (1) IMPLIES THE OTHERS.

- RELATIONSHIP TO CONDITIONAL PROBABILITY: THE BASIC EQUATION REGARDING INDEPENDENCE OF n EXPERIMENTS, AND, IN PARTICULAR, ITS RESTRICTION TO ANY SUBSET OF $m \leq n$ DISTINCT EXPERIMENTS:

(SEE THEOREM AFTER DEFN)

$$P\left(A_{q_1}^{(i_1)} \cap \cdots \cap A_{q_m}^{(i_m)}\right) = P\left(A_{q_1}^{(i_1)}\right) \cdot \ldots \cdot P\left(A_{q_m}^{(i_m)}\right),$$

TOGETHER W/ THE "MULTIPLICATION THEOREM," IMPLIES THAT:

$$P\left(A_{q_m}^{(i_m)} \,\middle|\, A_{q_1}^{(i_1)} \cap \cdots \cap A_{q_{m-1}}^{(i_{m-1})}\right) = P\left(A_{q_m}^{(i_m)}\right)$$

I.E., CONDITIONING A TRIAL'S EVENT ON THE OCCURANCE OF EVENTS IN OTHER, INDEPENDENT, TRIALS DOES NOT CHANGE THAT EVENT'S PROBABILITY. FOR TWO INDEPENDENT EVENTS:

(OBVIOUS FROM DEFN OF CONDITIONAL PROB)

$$P(A \cap B) = P(A) \cdot P(B) \implies P(A|B) = P(A)$$

(AND $P(B|A) = P(B)$)

- **THEOREM**: A NECESSARY AND SUFFICIENT CONDITION FOR THE INDEPENDENCE OF TRIALS $A^{(1)}, A^{(2)}, \ldots, A^{(n)}$ IS THAT THE CONDITIONAL PROBABILITY OF SOME $A_q^{(i)}$ UNDER THE ASSUMPTION THAT SEVERAL OTHER TRIALS, $A^{(i_1)}, A^{(i_2)}, \ldots, A^{(i_k)}$ HAVE HAD DEFINITE RESULTS, $A_{q_1}^{(i_1)}, A_{q_2}^{(i_2)}, \ldots, A_{q_k}^{(i_k)}$ IS EQUAL TO THE (UNCONDITIONAL / ABSOLUTE) PROBABILITY OF $A_q^{(i)}$ (WHEN $P(A_q^{(i)}) > 0$)
- **THEOREM**: SAME AS ABOVE BUT FOR EVENTS.

- **DEFN**: EVENTS A AND B ARE CONDITIONALLY INDEPENDENT GIVEN C IF

$$P(A \cap B \mid C) = P(A \mid C) \cdot P(B \mid C)$$

- **EXAMPLES**
  - (6 ON EACH ROLL OF DIE) VS. (6 ON FIRST, SUM IS 8)
  - (DRAWING TWO CARDS W/ REPLACEMENT) VS. (W/O REPLACE.)
  - A MARKOV CHAIN (AKA, MARKOV PROCESS) DESCRIBES A SEQUENCE OF EVENTS FOR WHICH THE PROBABILITY OF THE NEXT EVENT DEPENDS ONLY ON THE CURRENT STATE. THUS, CONDITIONED ON THE CURRENT STATE, THE FUTURE AND PAST STATES ARE INDEPENDENT

MARKOV:
$P(E_{n+1} \cap E_{n-1} \mid E_n)$
$= P(E_{n+1} \mid E_n) \cdot P(E_{n-1} \mid E_n)$
MORE COMPREHENSIVELY:
$P(E_{n+1} \mid E_n, E_{n-1}, \ldots) = P(E_{n+1} \mid E_n)$

RANDOM VARIABLES

2022/02/26

- RECALL THAT A MEASUREABLE MAP FROM ONE MEASUREABLE SPACE $(M, \mathcal{E})$ TO ANOTHER $(N, \mathcal{F})$ IS A MAP, $g: M \to N$, FOR WHICH $\text{PREIM}_g(f \in \mathcal{F}) \in \mathcal{E}$. IF $(M, \mathcal{E}, P)$ IS A MEASURE SPACE, THEN ONE CAN DEFINE A MEASURE ON $(N, \mathcal{F})$ AS THE PUSH-FORWARD OF P:

$$g_* P : \mathcal{F} \to \overline{\mathbb{R}}^+, \quad (g_* P)(f) := P\left(\text{PREIM}_g(f)\right)$$

SCHULLER QM 05 (MEASURE)

MEASUREABLE MAPS ARE (WITH MINOR ADDITIONS) THOSE THAT ARE INTEGRABLE.

- **DEF'N**: GIVEN A PROBABILITY SPACE $(M, \mathcal{E}, P)$ AND A MEASURABLE SPACE $(N, \mathcal{F})$, A MEASURABLE MAP, $g: M \to N$, IS CALLED AN $(N, \mathcal{F})$-VALUED RANDOM VARIABLE.

- RECALL THAT, GIVEN A TOPOLOGICAL SPACE (M, Θ_M), THE BOREL σ-ALGEBRA OF (M, Θ_M) IS:

$$\sigma(\Theta_M) = \{A \in M \mid \forall\ \sigma\text{-ALGEBRAS}, \sigma, \text{ OF } M \text{ WITH } \Theta_M \subseteq \sigma : A \in \sigma\}$$

I.E., IT IS THE σ-ALGEBRA GENERATED BY THE TOPOLOGY. FOR $(\mathbb{R}^d, \Theta_{STD})$ WE CALL $\sigma(\Theta_{STD})$ THE BOREL σ-ALGEBRA OF $\mathbb{R}^d$

GENERATOR: $\sigma(E)$ FOR $E \subseteq \mathcal{P}(M)$ IS $\sigma(E) = \bigcap_{\text{ALL } \sigma \text{ CONTAINING } E} \sigma$ AND IS THE SMALLEST σ-ALGEBRA CONTAINING E.

- **PROPOSITION**: GIVEN A PROBABILITY SPACE $(M, \mathcal{E}, P)$, A MAP $X: M \to \mathbb{R}$ IS A REAL-VALUED RANDOM VARIABLE IF

$$\text{PREIM}_X\left((-\infty, a)\right) \in \mathcal{E} \qquad \forall\ a \in \mathbb{R}$$

- **PROOF**: WE ASSUME THE TARGET IS THE MEASURABLE SPACE $(\mathbb{R}, \sigma(\Theta_{STD}))$. THE ABOVE DEFINITION GUARANTEES THAT X IS A MEASURABLE MAP FOR A TARGET $(\mathbb{R}, \mathcal{E}_{LOR})$ WHERE

$$\mathcal{E}_{LOR} = \sigma\left(E_{LOR}\right) \text{ w/ } E_{LOR} = \{(-\infty, a)\}$$

IS THE σ-ALGEBRA GENERATED BY THE SET OF "LEFT-OPEN RAYS". THUS WE MUST PROVE THAT

$$\sigma\left(E_{LOR}\right) = \sigma\left(\Theta_{STD}\right)$$

(SCHILLER SAYS:) LEMMA: TO SHOW THAT $g: M \to N$ IS MEASURABLE FOR $(M, \mathcal{E})$, $(N, \sigma(E))$, IT SUFFICES TO CHECK THE ELEMENTS OF THE GENERATING SET: $\text{preim}_g(e \in E) \in \mathcal{E}$

WE WILL USE THE FACT THAT, FOR A SUBSET OF A σ-ALGEBRA: $A \subseteq \sigma \Rightarrow \sigma(A) \subseteq \sigma$. NOW, KNOWING THAT AN OPEN INTERVAL IS OPEN, WE CAN WRITE A HALF-CLOSED INTERVAL AS THE COUNTABLE INTERSECTION OF OPEN INTERVALS:

SEE PROP. 1.9 OF people.clas.ufl.edu/pascoej/files/6616notes01dec2017.pdf

$$[a,b) = \bigcap_{n=1}^{\infty} \left(a - \tfrac{1}{n}, b\right) \in E_{OI}$$

SEE: mpaldridge.github.io/teaching/ma40042-notes-02.pdf

FURTHER, WE CAN WRITE:

$$(-\infty, a) = \bigcup_{n=1}^{\infty} [n, a)$$

THUS, AS SETS,

$$E_{LOR} \subseteq E_{HCI} \subseteq E_{OI} \subseteq \Theta_{STD}$$

AND, THEREFORE

$$\sigma(E_{LOR}) \subseteq \sigma(\Theta_{STD})$$

NOT AT ALL SURE HOW TO DO THIS ... REF. ABOVE (mpaldridge) IS NOT SATISFYING

SO THEN IT REMAINS TO SHOW:

$$\sigma(\Theta_{STD}) \subseteq \sigma(E_{LOR})$$

WHICH COULD BE DONE BY SHOWING ... ??!

- DEF'N: THE PROBABILITY FUNCTION OF A REAL-VALUED RANDOM VARIABLE ON $(M, \mathcal{E}, P)$ IS THE PUSH-FORWARD OF P:

$\sigma(\Theta_{STD})$ THE BOREL σ-ALGEBRA OF $\mathbb{R}$.

$$P^{(X)} : \sigma(\Theta_{STD}) \to \overline{\mathbb{R}}^{+}, \quad P^{(X)}(u) := P\left(\text{PREIM}_X(u)\right)$$

THIS ALLOWS US TO VIEW $\left(\mathbb{R}, \sigma(\Theta_{STD}), P^{(X)}\right)$ AS A PROBABILITY SPACE.

- DEF'N: THE DISTRIBUTION FUNCTION (AKA THE CUMULATIVE DISTRIBUTION FUNCTION) OF A RANDOM VARIABLE X IS THE FUNCTION

I.E., A MAP: $F^X : \mathbb{R} \to [0,1]$

$$F_X(a) = P^{(X)}\left((-\infty, a)\right) = \text{“}P(X < a)\text{”}^{*}$$

* math.stackexchange.com/questions/2142162
LET'S ALLOW OURSELVES TO WRITE:
$P^{(X)}((-\infty, a)) = \text{“}P(X \leq a)\text{”}$
WHERE
$\text{“}P(X \leq a)\text{”} = P(\text{“}\{X \leq a\}\text{”})$
AND
$\text{“}\{X \leq a\}\text{”} = \{\omega \in \mathcal{E} \mid X(\omega) \leq a\}$

AND $F_X(a)$ CAN BE STATED, AS WRITTEN ABOVE, AS "THE PROBABILITY THAT THE RANDOM VARIABLE IS LESS THAN a".

- CLAIM,

$$F_X(-\infty) = 0 \quad \text{AND} \quad F_X(+\infty) = 1$$

- AND

$$P^{(X)}([a,b)) = P^{(X)}\left((-\infty, a)^{c} \cap (-\infty, b]\right)$$

Recall the Addition Law:
$P(A \cup B) = P(A) + P(B) - P(A \cap B)$

$$= 1 - P^{(X)}\left((-\infty, a)\right) + P\left((-\infty, b)\right) - P^{(X)}\left((-\infty, a)^c \cup (-\infty, b)\right)$$

$$= 1 - F_X(a) + F_X(b) - \cancel{P^{(X)}\left((-\infty, +\infty)\right)}^{\,1}$$

$$= F_X(b) - F_X(a)$$

- Which implies that

$$F_X(a) \leq F_X(b) \quad \text{for} \quad a \leq b$$

showing that $F_X(x)$ is a nondecreasing fcn.

(Can show F^X is continuous (see Kolmogorov III-2))

- Defn: When the distribution function $F^X(a)$ is differentiable we define the probability density of X at a:

$$f_X(a) = \frac{d}{dx} F_X(x)\Big|_{x=a}$$

And if the following holds $\forall x \in \mathbb{R}$ then the distribution is called continuous:

$$F_X(x) = \int_{-\infty}^{x} f_X(a)\, da$$

And then the probability of any Borel set $A \subseteq \mathbb{R}$ is

$$P^X(A) = \int_A f_X(a)\, da$$

- The distribution and density functions can be generalized for conditional probabilities. Define:

$$P^{(X)}(A \mid B) := "P(X \subset A \mid B)"$$

(Note again: different P maps)

$$= P\left(\{E \in \mathcal{E} \mid X(E) \subset A\} \mid B\right)$$

Then:

$$F_X(a \mid B) = P(X < a \mid B)$$

And

$$f_X(a \mid B) = \frac{d}{dx} F_X(x \mid B)\Big|_{x=a}$$

Multi-Dimensional Distribution Functions: Given N random variables $X = (X_1, X_2, \ldots, X_N)$ we have

$$X : M \longrightarrow \mathbb{R}^N$$

(As above, one can extend the target of X from intervals to the entire Borel σ-algebra.)

And a probability function (Borel σ-algebra of $\mathbb{R}^N$)

$$P^{(X_1, X_2, \ldots, X_N)} : \sigma(\Theta_{STD}) \longrightarrow \overline{\mathbb{R}}^+$$

Defined in the same manner as $P^{(X)}$. A distribution function is defined

$$F_{X_1 X_2 \cdots X_N}(a_1, a_2, \ldots, a_N) := P^{(X_1, \ldots, X_N)}\left(L_{a_1 a_2 \cdots a_N}\right)$$

where $L_{a_1 a_2 \cdots a_n} = \{x \in \mathbb{R}^N \mid x_1 < a_1, x_2 < a_2, \ldots\}$

Although F gives directly only the values $P^{(X)}$ on the "L" sets, $P^{(X_1, \ldots, X_N)}$ is uniquely determined for all $A \in \sigma(\Theta_{STD})$ by knowledge of $F_{X_1 X_2 \cdots X_N}$.

(Key Point! Kolmogorov pg 26 → 2.3-IV.)

And... if the derivative exists, we have the N-dimensional probability density of $X = (X_1, X_2, \ldots, X_N)$:

$$f_X(a_1, a_2, \ldots, a_N) = \frac{\partial^{(N)}}{\partial x_1 \, \partial x_2 \cdots \partial x_N} F_{X_1 \cdots X_N}(x_1, \ldots, x_N)\Big|_{\vec{a}}$$

(Or, more precisely: $f_{X_1 X_2 \cdots X_N}(a_1, \ldots, a_N)$)

Which for continuous F_X (see above) implies:

$$P^{(X)}(A) = \int \cdots \int_A f(a_1, \ldots, a_N)\, d^N a$$

For all Borel sets $A \subseteq \mathbb{R}^N$.

Expectation of Discrete Random Variables: Consider a

(AKA, a Discrete Random Var.)

random variable $X : M \to \mathbb{R}$ that takes only finitely-many values $\{a_n\}$ and thus partitions the space M into $\{A_n = \text{preim}_X(a_n)\}$. The mathematical expectation of X is:

(Recall: $P(\text{preim}_X(a_n))$)

$$E[X] = \sum_n a_n P(A_n) = \sum_n a_n P^{(X)}(a_n) =: \sum_n a_n p_X(a_n)$$

where p_X is the push-forward of P restricted to the set $\{a_n\}$, and is called the probability mass function.

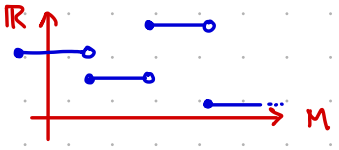
Lebesgue Integration

2022/03/07 – 10

From Schuller QM Lecture 06

Defn: Given the measurable spaces (M,Σ) and $(\mathbb{R},\sigma(\mathcal{O}_{std}))$, a measurable function $S: M \to \mathbb{R}^+_0$ is called a <u>simple function</u> (or, <u>step map</u>) if it takes M into only finitely-many values, i.e.,

$$S(M) = \{s_1, s_2, \dots, s_N\}$$



The function thus partitions M into the finite partition $\{A_n\}$ and can therefore be expressed as

$$S = \sum_{n=1}^{N} s_n \chi_{A_n} \quad ; \quad \chi_A : M \to \mathbb{R}, \; m \mapsto \begin{cases} 1 & m \in A \\ 0 & m \notin A \end{cases}$$

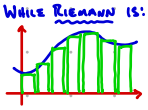
Indicator Fcn = Characteristic Fcn

where χ_A is an <u>indicator function</u> and $A_n = \text{preim}_S(s_n)$.

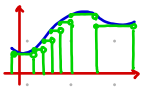
Defn: Let (M,Σ,μ) be a measure space and $(\overline{\mathbb{R}},\overline{\sigma})$ a measurable space. The <u>integral of a non-negative measurable map</u>, $f: M \to \overline{\mathbb{R}}$ $(f(m) \geq 0 \;\forall\, m \in M)$, is:

($\overline{\sigma}$ = the extended $\sigma(\mathcal{O}_{std})$ for $\overline{\mathbb{R}}$)

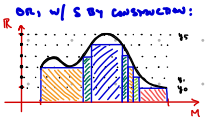
While Riemann is:



Lebesgue is:



Or, w/ S by construction:



Where partition of range is used to define S and thus partition the domain.

$$\int f\, d\mu := \sup\left\{ \sum_{z \in S(M)} \underbrace{z}_{\text{"height"}} \cdot \underbrace{\mu\left(\text{preim}_S(z)\right)}_{\text{"width"}} \;\middle|\; \begin{array}{l} S \text{ is a} \\ \text{simple fcn} \\ \text{w/ } S \leq f \end{array} \right\}$$

For the integral over a subset of the domain, $(A \subseteq M) \in \Sigma$, we define the <u>integral of a non-negative measurable map f over A</u> as:

$$\int_A f\, d\mu := \int f \cdot \chi_A \, d\mu$$

Recall: Product of maps $(f \cdot \chi)(m) := f(m) \cdot \chi(m)$ (in $\mathbb{R}$)

An <u>alternative notation</u> is to express the integral in terms of the values of the function (aka, the dummy variable(s) of integration). E.g., For mult: $\mathbb{R}\times\mathbb{R} \to \mathbb{R}$, we write:

$$\int_A \text{mult}\, d\mu = \int_A \text{mult}(x,y)\, d\mu(x,y) = \int_A x \cdot y \, d\mu(x,y) = \int_A x\, y\, \mu(d(x,y))$$

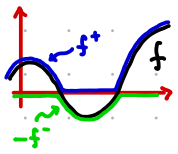
Common in Probability Th.

Fubini's Thm? One needs to do something w/ "product measurable spaces" μ on $(M\times N, \Sigma_M \otimes \Sigma_N)$ w/ $\mu = \mu_M \cdot \mu_N$ to get $\mu(d(x,y)) = \mu(dx)\,\mu(dy) = dx\,dy$

- **DEFN**: THE INTEGRAL OF A MEASURABLE MAP (NOT NECESSARILY POSITIVE), $f : M \to \overline{\mathbb{R}}$, FOR WHICH $\int |f| \, d\mu < \infty$, IS DEFINED AS

THIS CONDITION MAKES THE MAP "INTEGRABLE"

$$\int f \, d\mu := \int f^+ \, d\mu - \int f^- \, d\mu$$



WHERE

$$f^+ := \max(f, 0) \quad \text{AND} \quad f^- = \max(-f, 0)$$

2022/03/11

From Billingsley's Probability & Measure (2nd Ed) Section 16

- **CHANGE OF VARIABLE**: LET $T : M \to M'$ BE A MEASURABLE MAP FROM $(M, \mathcal{F}, \mu)$ TO $(M', \mathcal{F}')$. WE CAN DEFINE A MEASURE ON $(M', \mathcal{F}')$ AS USUAL VIA THE PUSHFORWARD:

$$\mu'_T : \mathcal{F}' \to \overline{\mathbb{R}}^+_0,$$

$$A' \mapsto \mu'_T(A') := \mu\left(\text{Preim}_T(A')\right)$$

IF $f : M' \to \mathbb{R}$ IS A MEASURABLE REAL FUNCTION ON $(M', \mathcal{F}')$ THEN $(f \circ T) : M \to \mathbb{R}$ IS A MEASURABLE REAL FUNCTION ON $(M, \mathcal{F})$.

THEOREM: IF f IS NONNEGATIVE, THEN

$$(*) \qquad \int_M (f \circ T) \, d\mu = \int_{M'} f \, d\mu'_T$$

AND

$$(**) \qquad \int_{\text{Preim}_T(A')} (f \circ T) \, d\mu = \int_{A'} f \, d\mu'_T$$

FURTHER, A FUNCTION f (NOT NECESSARILY NONNEGATIVE) IS INTEGRABLE WITH RESPECT TO μ_T IF AND ONLY IF $(f \circ T)$ IS INTEGRABLE WITH RESPECT TO μ.

Proof: If $f = \chi_{A'}$ (indicator function) then:

$$f \circ T = \chi_{A'} \circ T = \chi_{\text{Preim}_T(A')}$$

So eqn (*) becomes

2022/03/12

$$\int_M \chi_{\text{Preim}_T(A')} \, d\mu = \int_{M'} \chi_{A'} \, d\mu_T'$$

Indicator fcn is a 2-valued simple function w/ value 1 over A' and 0 elsewhere. So the "sup" is calculated trivially.

$$\mu\left(\text{Preim}_T(A')\right) = \mu_T'(A')$$

which is simply the definition of μ_T'. By the linearity of the integral, eqn (*) then holds for all nonnegative simple functions. And, if $\{f_n\}$ is a sequence of simple functions $f_n \to f$, then $(f_n \circ T) \to (f \circ T)$ and (*) follows by the monotone convergence theorem. Application of (*) to $|f|$ "establishes the assertion about integrability", and (*) follows for general f by decomposition into f^+ & f^-. Finally, (**) follows from (*) via $f \to f \cdot \chi_{A'}$ (multiply not composition).

Not sure but don't care right now.

Defn: Given a measurable map $f: [a,b] \to \mathbb{R}$, and a function $g: [a,b] \to \mathbb{R}$ of bounded variation, the Lebesgue–Stieltjes integral is defined as the following Lebesgue integral:

There is a generalization to: $f, g: \mathbb{R}^d \to \mathbb{R}$

$$\int_a^b f(x) \, dg(x) := \int_{[a,b]} f \, d\mu_g$$

where μ_g is the uniquely-defined Lebesgue–Stieltjes measure on $[a,b]$ that has value on any interval:

$$\mu\left((s,t]\right) = g(t) - g(s)$$

See: Billingsley End of Section 17

The importance of this rewriting of a Lebesgue integral is that, for continuous f, the L.S. integral is Riemannian.

Mathematical Expectation

- **Defn**: Given a random variable $X: M \to \mathbb{R}$ on the probability space $(M, \mathcal{F}, P)$, the **mathematical expectation of X** is the Lebesgue integral:

$$E[X] := \int_M X \, dP = \int_M X(m) \, P(dm)$$

(Just notation)

Note that our prior definition of the same quantity — valid for a discrete, finitely-valued random variable — coincides with this definition since such a random variable *is* a simple function.

- Recall that we defined the probability function of X as:

$$P^{(X)} : \sigma(\theta_{std}) \to \overline{\mathbb{R}}^+, \quad u \mapsto P^{(X)}(u) := P\left(\text{preim}_X(u)\right)$$

This is precisely the measure μ_T' defined in the change of variable formula, with $\mu = P$, $T = X$, and $\mu_T' = P^{(X)}$. Therefore we can write

$$E[X] = \int_M X \, dP$$

($T = X$; $f = Id_{\mathbb{R}}$ (identity fcn))

$$= \int_M Id_{\mathbb{R}} \circ X \, dP$$

(change of variables theorem)

$$= \int_{\mathbb{R}} Id_{\mathbb{R}} \, dP^{(X)}$$

(alternative notation)

$$= \int_{\mathbb{R}} x \, dP^{(X)}(x) = \int_{\mathbb{R}} x \, P^{(X)}(dx)$$

- Using the Lebesgue–Stieltjes notation, we can write

$$\int Id_{\mathbb{R}} \, dP^{(X)} =: \int_{-\infty}^{\infty} x \, dF_X(x)$$

Compared to L-S defn: $f = Id_{\mathbb{R}}$, $g = F_X$, $\mu_g = P^{(X)}$

WHERE WE HAVE IDENTIFIED $P^{(X)}$ AS THE UNIQUELY-DEFINED MEASURE ASSOCIATED TO $F_X : \mathbb{R} \to \mathbb{R}$ THAT, FOR ANY INTERVAL $(a,b] \subseteq \mathbb{R}$, HAS VALUE:

$$\mu_{F_X}\left((a,b]\right) = F_X(b) - F_X(a) := P^{(X)}\left((a,b]\right).$$

* – IS IT?

THE LAST EQUALITY IS (EQUIVALENT TO *) HOW WE DEFINED THE DISTRIBUTION FUNCTION ORIGINALLY. FOR CONTINUOUS $F_X(x)$ THIS STIELTJES INTEGRAL IS RIEMANNIAN. AND, IF $F_X(x)$ IS DIFFERENTIABLE, THEN

(SHORTHAND RIEMANNIAN INTEGRATION)

$$\int_a^b x\, dF(x) = \lim_{N\to\infty} \sum_{n=0}^{\infty} x_n \cdot \left(F(x_{n+1}) - F(x_n)\right)$$

$$= \lim_{N\to\infty} \sum_{n=0}^{\infty} x_n \frac{F(x_n+\Delta x) - F(x_n)}{\Delta x} \Delta x$$

RECALL: $f(x)\big|_a = \frac{dF(x)}{dx}\Big|_a$

$$= \int_a^b x\, f_X(x)\, dx$$

THUS, WE CAN FURTHER WRITE THE MATHEMATICAL EXPECTATION (WHEN F_X AND f_X ARE WELL-BEHAVED) AS:

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

WHERE f_X IS THE PROBABILITY DENSITY OF X.

- CONDITIONAL EXPECTATION OF A RV WITH RESPECT TO AN EVENT

WE HAVE ALREADY SEEN THAT A CONDITIONAL PROBABILITY, $P(A|B)$, IS A PROBABILITY MEASURE FOR THE SPACE $(M, \mathcal{E}, P(\cdot|B))$. WE COULD WRITE THIS PROBABILITY USING KOLMOGOROV'S NOTATION

$$P_B(A) := P(A|B)$$

WE CAN THEN DEFINE THE CONDITIONAL EXPECTATION OF A RANDOM VARIABLE X WITH RESPECT TO THE

EVENT B:

$$E_B[X] := \int_M X \, dP_B \qquad \left(P_B(\cdot) = P(\cdot \mid B)\right)$$

$$= \int_B X \, dP_B$$

SINCE $P_B(A) = 0$ FOR $A \cap B = \phi$

$$E_B[X] = \frac{1}{P(B)} \int_B X \, dP$$

$\left(P(B)\, E_B[X] = \int_B X \, dP\right) \iff$

SINCE $P_B(A) = \frac{P(A)}{P(B)}$ WHEN $A \subset B$ (SINCE $A \cap B = A$)

WE COULD CONSIDER TWO CONDITIONAL PROBABILITIES

$$P(A)\, E_A[X] + P(B)\, E_B[X] = \int_A X \, dP + \int_B X \, dP$$

$$= \int_{A+B} X \, dP$$

SUCH THAT

$$E_{A+B}[X] = \frac{P(A)\, E_A[X] + P(B)\, E_B[X]}{P(A+B)}$$

AND FOR $A+B = M$:

$$E[X] = P(A)\, E_A[X] + P(A^c)\, E_{A^c}[X]$$

MORE GENERALLY, FOR A PARTITION $M = A_1 \cup A_2 \cup \dots \cup A_n$

$E[X] = \sum_{i=1}^{n} P(A_i)\, E_{A_i}[X]$

CONDITIONAL PROBABILITIES WRT TRIALS AND RANDOM VARIABLES

- DEFN: THE CONDITIONAL PROBABILITY OF AN EVENT, B, AFTER A TRIAL $\mathcal{A}$ IS A RANDOM VARIABLE THAT, FOR EVERY EVENT $A_i \in \mathcal{A}$, TAKES THE VALUE $P(B \mid A_i)$. WE CAN WRITE THIS AS

THESE ARE KOLMOGOROV'S WORDS BUT THEY DON'T MAKE SENSE B.C. $X: M \to \mathbb{R}$ $=$ IT'S NOT KNOWN THAT $A_i \in M$

2022/03/13 – 20

math.stackexchange.com/questions/4402669

DEFN: FOR A TRIAL $\mathcal{A}$ WITH COUNTABLY-MANY OUTCOMES $\{A_i\}$, WE CAN ASSOCIATE AN INTEGER-VALUED RANDOM VARIABLE:

$$X_{\mathcal{A}} : M \to \mathbb{N}, \quad m \longmapsto i \quad \text{WHEN} \quad m \in A_i$$

SO THE $\text{PREIM}_{X_{\mathcal{A}}}(i) = A_i$ $(i = 1, 2, \ldots, k)$. WE THEN DEFINE THE CONDITIONAL PROBABILITY OF THE EVENT B AFTER THE TRIAL $\mathcal{A}$ AS THE RANDOM VARIABLE:

JUST NOTATION: $B \in \mathcal{E}$: EVENT; $\mathcal{A}$: PARTITION OF M; $[P(B|\mathcal{A})]$: RANDOM VARIABLE

$$[P(B|\mathcal{A})] : M \to \mathbb{R}$$

$$m \longmapsto [P(B|\mathcal{A})](m) = P(B \mid A_{X_{\mathcal{A}}(m)})$$

NOTATIONALLY, WE COULD WRITE

$$P(B \mid \mathcal{A} = A_{X_{\mathcal{A}}(\cdot)}) := [P(B|\mathcal{A})](\cdot)$$

OR, SINCE THE VALUES OF $X_{\mathcal{A}}$ ARE 1-1 WITH THE ELEMENTS OF THE PARTITION, WE COULD JUST USE:

$$P(B \mid X_{\mathcal{A}} = X_{\mathcal{A}}(\cdot)) := [P(B|\mathcal{A})](\cdot)$$

- THE EXPECTATION OF THIS DISCRETE RANDOM VARIABLE IS:

$$E\left[[P(B|\mathcal{A})]\right] = \sum_{i=1}^{k} P(B \mid \mathcal{A} = A_i)\, P(A_i)$$

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

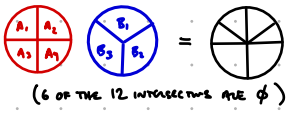
AND THE CONDITIONAL EXPECTATION OF $[P(B|\mathcal{A})]$ WITH RESPECT TO (FIXED EVENT) C IS:

$$E_C\left[[P(B|\mathcal{A})]\right] = \frac{1}{P(C)} \int_C [P(B|\mathcal{A})]\, dP$$

$$= \frac{1}{P(C)} \sum_{\{i \mid A_i \subseteq C\}} P(B|A_i)\, P(A_i)$$

Given two partitions of M, $\mathcal{P}^{(1)}$ & $\mathcal{P}^{(2)}$, let their product (or meet), $\mathcal{P}^{(1)}\mathcal{P}^{(2)}$, be the new partition that is the set of all nonempty intersections:

$\{P^{(1)}_{i_1} \cap P^{(2)}_{i_2}\}$ $(P^{(k)}_{i_k} \in \mathcal{P}^{(k)})$



(6 of the 12 intersections are ϕ)

See theorem near end of "Independent Trials & Events"

Given multiple trials $A^{(1)}, A^{(2)}, \ldots, A^{(n)}$, we can define

$$\left[P\left(B \mid A^{(1)}A^{(2)}\cdots A^{(n)}\right)\right] : M \longrightarrow \mathbb{R}$$

as the random variable that takes the value:

$$P\left(B \mid A^{(1)}_{i_1} \cap A^{(2)}_{i_2} \cap \cdots \cap A^{(n)}_{i_n}\right)$$

when acting on an element, m, for which $m \in A^{(1)}_{i_1} \cap A^{(2)}_{i_2} \cap \cdots \cap A^{(n)}_{i_n}$. The condition for independent trials can then be written as:

$$\left[P\left(A^{(k)}_q \mid A^{(1)}A^{(2)}\cdots A^{(k-1)}\right)\right] = P\left(A^{(k)}_q\right)$$

for arbitrary k, q.

⑤ Conditional Probability with respect to a Random Variable

A random variable $X : M \to \mathbb{R}$ can be thought of as partitioning M through its preimage, and thus X can be considered a trial. In the case of a discrete random variable, the prior discussions regarding trials and conditioning with respect to trials can be immediately applies to random variables

(For a Discrete Random Variable)

$$\left[P(B \mid X)\right] : M \longrightarrow \mathbb{R}$$

$$m \longmapsto P\left(B \mid \text{Preim}_X(X(m))\right)$$

where we could use the notation asove to write:

$$P(B \mid X = X(m)) := \left[P(B \mid X)\right](m)$$

For general random variables, which can take an uncountable number of outcomes, the set

$$\text{“}X = X(m)\text{”} = \text{Preim}_X(X(m))$$

KEY POINT/MOTIVATION OF THIS ENTIRE ANALYSIS

OFTEN HAS PROBABILITY ZERO. THEREFORE THE ABOVE RESULT CANNOT BE USED SINCE $P(B|A) = P(B \cap A)/P(A)$ $(P(A) > 0)$. WE MUST THEREFORE DEFINE THE CONDITIONAL PROBABILITY WITH RESPECT TO A RANDOM VARIABLE INDIRECTLY.

DEF'N: FOR A GENERIC RANDOM VARIABLE, $X : M \to \mathbb{R}$, FROM $(M, \mathcal{E}, P)$ TO $(\mathbb{R}, \sigma(\theta_{STD}), P^{(X)})$, THE CONDITIONAL PROBABILITY OF AN EVENT $B \in \mathcal{E}$ WITH RESPECT TO X IS THE UNIQUE RANDOM VARIABLE $[P(B|X)]$ SATISFYING:

$$P(B \mid \text{PREIM}_X(u)) = E_{\text{PREIM}_X(u)}\left[[P(B|X)]\right]$$

FOR ALL $u \in \sigma(\theta_{STD})$ WITH $P^{(X)}(u) > 0$.

PROOF: WE MUST SHOW: ① THAT THIS AGREES WITH THE ABOVE DEFINITION FOR DISCRETE RANDOM VARIABLES AND, ② THAT IT UNIQUELY* DEFINES $[P(B|X)]$ FOR GENERIC RANDOM VARIABLES.

* - UP TO THE EQUIVALENCE DEFINED ON NEXT PAGE.

① FOR A DISCRETE X WITH TARGET VALUES $\{x_1, x_2, \ldots, x_n\}$, LET $A_i = \text{PREIM}_X(x_i)$, AND LET $A_u = \text{PREIM}_X(u)$ FOR $u \subseteq \{x_1, \ldots, x_n\}$:

$\mathcal{A} = \{A_i\}$

CONDITIONAL EXPECTATION WAS DEFINED AS A LEBESGUE INTEGRAL, BUT THAT WORKS FOR A DISCRETE RV (I.E., A STEP FCN)

$$E_{A_u}\left[[P(B|X)]\right] := \frac{1}{P(A_u)} \sum_{\{i \mid A_i \in A_u\}} [P(B|X)]_i \, P(A_i)$$

IF WE LET $[P(B|X)]_i = P(B \mid X = x_i) = P(B|A_i)$, AS DEFINED ABOVE FOR THE DISCRETE CASE, THEN:

$$\frac{1}{P(A_u)} \sum_{\{i \mid A_i \subseteq A_u\}} P(B|A_i) P(A_i) = \frac{1}{P(A_u)} \sum_{\{i \mid A_i \subseteq A_u\}} P(B \cap A_i)$$

$$= \frac{1}{P(A_u)} P\left(B \cap \left[\bigcup_{\{i \mid A_i \subseteq A_u\}} A_i\right]\right) = \frac{P(B \cap A_u)}{P(A_u)} = P(B|A_u)$$

WHERE WE HAVE USED THE FACT THAT SOME SUBSET OF THE $\mathcal{A} = \{A_i\}$ PARTITIONS A_u SINCE $A_u \subseteq \mathcal{A}$.

② WE WILL PROVE UNIQUENESS. FOR THE EXISTENCE PROOF, SEE CHAPTER V IN KOLMOGOROV (USING NIKODYM'S THM.).

STARTING FROM OUR DEFINING EQUATION FOR $[P(B|X)]$:

$$P\left(B \mid \text{PREIM}_X(u)\right) = E_{\text{PREIM}_X(u)}\left[[P(B|X)]\right]$$

MULTIPLY BOTH SIDES BY $P(\text{PREIM}_X(u))$. FIRST THE LHS:

$$P(\text{PREIM}_X(u)) \cdot P\left(B \mid \text{PREIM}_X(u)\right) = P\left(B \cap \text{PREIM}_X(u)\right)$$

AND THEN THE RHS:

$$P(\text{PREIM}_X(u)) \cdot \left[\frac{1}{P(\text{PREIM}_X(u))} \int_{\text{PREIM}_X(u)} [P(B|X)]\, dP\right]$$

$$= \int_{\text{PREIM}_X(u)} [P(B|X)]\, dP$$

Recall: For $T: M \to M'$
$\int_M (f \circ T)\, d\mu = \int_{M'} f\, d\mu_T'$
where
$\mu_T'(u) = \mu(\text{PREIM}_T(u))$

NOW WE WISH TO USE THE CHANGE OF VARIABLES THEOREM USING $T = X : M \to \mathbb{R}$ AND DEFINING $f : \mathbb{R} \to \mathbb{R}$ VIA:

$$f \circ X : M \to \mathbb{R}$$

$$m \longmapsto (f \circ X)(m) = [P(B|X)](m)$$

THUS, THE FUNCTION f MAPS THE TARGET OF THE RANDOM VARIABLE X INTO THE TARGET OF THE RANDOM VARIABLE $[P(B|X)]$ FOR THE SAME $m \in M$:

$$f : \mathbb{R} \longrightarrow \mathbb{R}$$

$$x \longmapsto f(x) = [P(B|X)]\left(\text{PREIM}_X(x)\right)$$

NOTATIONALLY, IT MAKES SENSE TO WRITE THIS AS

$$P(B \mid X = \cdot) \;:=\; [P(B|X)]\left(\text{PREIM}_X(\cdot)\right)$$

SO, WE FIND:

$$\int_{\text{PREIM}_X(u)} [P(B|X)]\, dP = \int_{\text{PREIM}_X(u)} P(B|X=\cdot) \circ X\, dP$$

$$= \int_u P(B|X=\cdot)\, dP^{(X)}$$

(NOW AN INTEGRAL OVER $u \subseteq \mathbb{R}$)

AND SETTING LHS EQUAL TO RHS:

$$P(B \cap \text{Preim}_X(\mathcal{U})) = \int_{\mathcal{U}} P(B \mid X = \cdot)\, dP^{(X)}$$

ONE OF THE FUNDAMENTAL PROPERTIES OF THE LEBESGUE INTEGRAL (PROPERTY IX, SECTION 1, CHAPTER IV) IS THAT IF, FOR A PROBABILITY SPACE $(M, \mathcal{E}, P)$,

$$\int_A X\, dP = \int_A Y\, dP \qquad \forall\ A \in \mathcal{E}$$

THEN X AND Y ARE EQUIVALENT RANDOM VARIABLES $(P(\{x \neq y\}) = 0)$. THEREFORE THE RHS ABOVE UNIQUELY SPECIFIES $P(B \mid X = \cdot)$ ON $(\mathbb{R}, \sigma(\mathcal{O}_{STD}), P^{(X)})$.

- NOTATION: FIRST IS KOLMOGOROV. SECOND IS MINE.

$$P_X(a; B) = P(B \mid X = a) = [P(B \mid X)](\text{Preim}_X(a))$$

- TWO RANDOM VARIABLES X AND Y INDUCE THE SAME PARTITIONING OF M IF THERE EXISTS A BIJECTIVE MAP FROM THE TARGET OF X, $\mathcal{U} \subseteq \mathbb{R}$, TO THE TARGET OF Y, $\mathcal{V} \subseteq \mathbb{R}$, $g : \mathcal{U} \to \mathcal{V}$, SUCH THAT

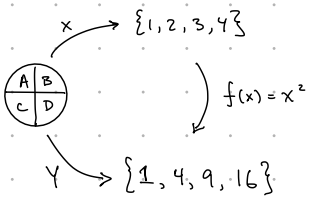
$$g \circ X : M \to \mathcal{V}$$

$$m \mapsto (g \circ X)(m) = Y(m)$$

I.E., $Y = g \circ X$. IN THAT CASE, THE RANDOM VARIABLES FOR CONDITIONAL PROBABILITY, AS DEFINED ABOVE ARE THE SAME*:

$$[P(B \mid X)] = [P(B \mid Y)]$$

THUS WE SAY THAT $[P(B \mid X)]$ IS DEFINED UNIQUELY UP TO EQUIVALENCE.



* I REALLY DON'T SEE THE EQUALITY OF $[P(B|X)]$ AND $[P(B|Y)]$.

- **Construction of Conditional Probability RV**: Although we have shown it exists and is unique (up to some equivalence), we have not yet written down an expression for $[P(B|X)]$. We did transform the defining equation from

$$P(B \mid \text{Preim}_X(u)) = \mathbb{E}_{\text{Preim}_X(u)}\left[[P(B|X)]\right]$$

(for all $u \in \sigma(\theta_{std})$ for which $P(\text{Preim}_X(u)) > 0$) to:

$$P(B \cap \text{Preim}_X(u)) = \int_u P(B \mid X = \cdot)\, dP^{(X)}$$

(for all $u \in \sigma(\theta_{std})$ w/ $P(\text{Preim}_X(u)) > 0$), where: we defined $P(B|X=\cdot): \mathbb{R} \to \mathbb{R}$ via:

$$[P(B|X)] = P(B|X=\cdot) \circ X$$

Now, we can write this integral in terms of the distribution F_X using the Lebesgue–Stieltjes formalism, with $g = F_X$ and therefore $\mu_g = P^{(X)}$. Thus, letting $u = (-\infty, a]$, we have:

Recall: $\int_a^b f(x)\, dg(x) := \int_{(a,b]} f\, d\mu_g$ w/ $\mu_g((s,t]) = g(t) - g(s)$

$$P(B \cap \text{Preim}_X((-\infty, a])) = \int_{-\infty}^{a} P(B|X=x)\, dF_X(x)$$

Kolmogorov notes (section 3, chapter V) that a theorem of Lebesgue provides a solution to the integrand:

$$P(B|X=a) = P(B)\left[\lim_{h\to 0} \frac{F_X(a+h|B) - F_X(a|B)}{F_X(a+h) - F_X(a)}\right]$$

where we define the conditional distribution, $F_X(x|B)$,

VIA:

This is where Kolmogolov's notation $P_B(A) = P(A|B)$ makes more sense. $P_B^{(x)}$ is a probability measure on $(\mathbb{R}, \mathcal{S}(\mathbb{R}))$ that happens to be constructed from a conditional probability P_B on $(M, \mathcal{E})$. My notation cabbies with it some intuition that the A & B in $P(A|B)$ are events in the same probability space but they need not be!

$$P^{(X)}(u) := P\left(\text{PREIM}_X(u)\right)$$

$$\Rightarrow \quad P^{(X)}(u \mid B) := P\left(\text{PREIM}_X(u) \mid B\right)$$

$$\Rightarrow \quad F_X(x \mid B) := P^{(X)}((-\infty, x) \mid B)$$

$$= P\left(\text{PREIM}_X(-\infty, x) \mid B\right)$$

NOW, IF THE DERIVATIVES OF $F_X(x)$ AND $F_X(x|B)$ EXIST, AND IF $f_X(a) = \left.\frac{dF_X(x)}{dx}\right|_{x=a} > 0$, WE CAN FURTHER WRITE:

$$\boxed{P(B \mid X = a) = P(B) \; \frac{f_X(a \mid B)}{f_X(a)}}$$

WHERE

$$f_X(a \mid B) = \left.\frac{dF_X(x \mid B)}{dx}\right|_{x=a}$$

THUS, WE REDUCED THE QUESTION OF THE RANDOM VARIABLE $[P(B|X)]$ TO THE FUNCTION $P(B|X=\cdot) : \mathbb{R} \to \mathbb{R}$ AND HAVE NOW DEFINED IT IN TERMS OF DENSITIES OF THE ORIGINAL RANDOM VARIABLE X.

- MARGINALIZING OVER CONDITIONAL PROBABILITIES: RETURNING TO OUR LAST INTEGRAL DERIVED FROM THE DEFINING EQUATION, BUT LETTING $u \to \mathbb{R}$, WE FIND

$$P(B) = \int_{-\infty}^{+\infty} P(B \mid X = x) \, dF_X(x)$$

$$\boxed{P(B) = \int_{-\infty}^{\infty} P(B \mid X = x) \, f_X(x) \, dx}$$

↓ DEFN ABOVE

$$P(B) = \int_{-\infty}^{\infty} P(B) \, f_X(a \mid B) \, dx \quad \Rightarrow \int f_X(a|0)\,dx = 1$$

- We can rearrange the equation we derived for $P(B|X=\cdot)$ and utilize the above marginalization to write:

$$f_X(a|B) \cdot P(B) = P(B|X=a) \cdot f_X(a)$$

$$f_X(a|B) = \frac{P(B|X=a) \cdot f_X(a)}{P(B)}$$

$$\boxed{f_X(a|B) = \frac{P(B|X=a) \cdot f_X(a)}{\int_{-\infty}^{\infty} P(B|X=x)\, f_X(x)\, dx}}$$

This seems to be a tautology based on the definition of $P(B|X=x)\cdots$ what is gained here?

which Kolmogorov calls <u>Bayes Theorem for Continuous Distributions</u>. I guess the idea here is: if we can deduce $P(B|X=\cdot)$ otherwise (or measure it in a trial/experiment) then its relationship to these other probability densities is this.

Bayes Theorem For Probability Densities

2022/03/23

- In the previous section, we mentioned in passing the conditional probability density but let us specify it explicitly now:
  - <u>Def'n</u>: <u>The probability density of a random variable, X, conditioned on the event B,</u> for which $P(B) > 0$, is:

$$f_X(x|B) = \frac{d}{dx}\left[F_X(x|B)\right]$$

$$= \frac{d}{dx}\left[P^{(X)}((-\infty, x)|B)\right]$$

$$= \frac{d}{dx}\left[P\left(\text{PreIm}_X(-\infty, x)\,|\,B\right)\right]$$

$$= \frac{d}{dx}\left[P\left(\text{PreIm}_X(-\infty, x) \cap B\right) / P(B)\right]$$

We are actually defining both $f_X(x|B)$ and $F_X(x|B)$ here... but the definitions are completely natural

(and in Kolmogorov's notation $f_B^{(X)}(x) \leftarrow F_B^{(X)}(x) \leftarrow P_B^{(X)}$ not really even new quantities.)

WE ARE INTERESTED IN THE PROBABILITY DENSITY OF ONE RANDOM VARIABLE CONDITIONED ON THE VALUE OF A SECOND RANDOM VARIABLE, I.E., CONDITIONED ON THE EVENT THAT THE 2ND RANDOM VARIABLE TAKE SOME VALUE!

$$f_X(x \mid \{Y = y\})$$

THIS IS PERFECTLY WELL DEFINED FOR A DISCRETE RANDOM VARIABLE Y, SO LONG AS $P(\{Y=y\}) > 0$. BUT FOR A CONTINUOUS RANDOM VARIABLE IT IS OFTEN THE CASE THAT $P(\{Y=y\}) = 0 \ \forall \ y$! THUS, WE CANNOT DIRECTLY APPLY THE DEFINITION GIVEN ABOVE. HOWEVER, GIVEN THE RESULTS OF THE PREVIOUS SECTION WE CAN MAKE THE FOLLOWING DEFINITION.

DEF'N: THE CONDITIONAL PROBABILITY DENSITY OF A RANDOM VARIABLE X WITH RESPECT TO THE EVENT THAT A SECOND RANDOM VARIABLE, Y, HAS TAKEN THE VALUE y IS:

$$f_X(x \mid \{Y=y\}) := \frac{d}{dx}\left[\left[P(\text{PREIM}_X(-\infty, x) \mid \{Y=y\})\right]\right]$$

WHERE WE USE THE RANDOM VARIABLE NOTATION $[P(B \mid \{Y=y\})]$ OF THE PREVIOUS SECTION.

NOTE: THE COMMONLY-USED NOTATION: $f_{X|Y}(x|y)$ IS BAD! WE ARE TALKING ABOUT THE DENSITY OF X WITH RESPECT TO A FIXED VALUE OF Y.

GIVEN THE DERIVED EXPRESSION IN THE PREVIOUS SECTION, WE CAN FURTHER CALCULATE:

$$f_X(x \mid \{Y=y\}) = \frac{d}{dx}\left[P(\text{PREIM}_X(-\infty, x)) \times \left(\frac{f_Y(y \mid \text{PREIM}_X(-\infty, x))}{f_Y(y)}\right)\right]$$

$$= \frac{\frac{d}{dx}\left[P(\text{PREIM}_X(-\infty, x) \cdot \frac{d}{dy}\left[P(\text{PREIM}_Y(-\infty, y) \mid \text{PREIM}_X(-\infty, x))\right]\right]}{f_Y(y)}$$

$$= \frac{\frac{\partial^2}{\partial x \partial y}\left[P\left(\text{Preim}_X(-\infty, x)\right) \cdot P\left(\text{Preim}_Y(-\infty, y) \,\middle|\, \text{Preim}_X(-\infty, x)\right)\right]}{f_Y(y)}$$

$$= \frac{\frac{\partial^2}{\partial x \partial y}\left[P\left(\text{Preim}_X(-\infty, x) \cap \text{Preim}_Y(-\infty, y)\right)\right]}{f_Y(y)}$$

This density is commonly written in inference applications as:

"$f(x|y)$"

which is Terrible because it is missing so much information:

- conditioning w/rt an event $\{Y=y\}$
- distribution of the RV "X", $f_X(x|\dots)$
- so x is a dummy var.; y is fixed

$$\boxed{f_X(x \mid \{Y=y\}) = \frac{f_{XY}(x,y)}{f_Y(y)}}$$

- **Bayes Theorem for Densities**: we can similarly find:

$$f_Y(y \mid \{X=x\}) = \frac{f_{XY}(x,y)}{f_X(x)}$$

Therefore:

$$f_X(x \mid \{Y=y\}) \cdot f_Y(y) = f_Y(y \mid \{X=x\}) \cdot f_X(x)$$

Or:

$$\boxed{f_X(x \mid \{Y=y\}) = \frac{f_Y(y \mid \{X=x\}) \cdot f_X(x)}{f_Y(y)}}$$

Bayes' Theorem for Densities

Or, alternatively:

$$\boxed{f_X(x \mid \{Y=y\}) = \frac{f_Y(y \mid \{X=x\} \cdot f_X(x)}{\int_{-\infty}^{\infty} \underbrace{f_Y(y \mid \{X=x'\}) \cdot f_X(x')}_{f_{XY}(x,y)} \, dx'}}$$

Overview of Bayesian Inference

2022/03/25

[Following D. Mackay's lectures 9-10, Inf. Th., Pat. Rec., Neur. Net]

1. **Toy Problem**: We are given a few real numbers $\{x_i\}_{i=1}^N$ and we are told (or we assume, as a **hypothesis model**, H_N) that they are samples from a normal distribution. What can we infer about that distribution from the data?

   - Under the hypothesis, H_N, our **underlying probability space**, M, is the set of all possible N-element data sets generated from all possible Normal distributions (specified by the mean and standard deviation, (μ, σ)). One could imagine expanding this space by removing the Normal distribution hypothesis, H_N, and considering, e.g., a multitude of other data-generating distributions. In that manner one could do **model-hypothesis comparison**. We could also expand to data sets of any size.
   - The particular **data set** that we have observed could be written as the event, D. Because there are uncountably-many possibilities, this is a probability-zero event and should be handled with a density, i.e., define a vector of random variables $\vec{D} = (D_1, D_2, \ldots, D_N)$ with joint density $f_{\vec{D}}(\vec{d})$, and write the event as $D = \{\vec{D} = \vec{d}\}$ where $\vec{d} = \{x_i\}$ is our observed data set.
   - The **parameters of the Normal distribution** are also random variables, with joint distribution $f_{M\Sigma}(\mu, \sigma)$. Another part of our hypothesis is its specification (**the prior**) on M.

Interesting: as soon as you call the parameters random variables it requires you to define the prior, i.e., the joint distribution.

ANSWER: THE BAYESIAN ANSWER TO THE QUESTION (WHAT CAN WE INFER ABOUT $N(\mu,\sigma)$ FROM THE DATA?) IS THE ("POSTERIOR") JOINT PROBABILITY DENSITY OF (M,Σ) CONDITIONED ON THE DATA EVENT, $D = \{\vec{D} = \vec{d}\}$, I.E.,

$$f_{M\Sigma}\left(\mu,\sigma \,\middle|\, \{\vec{D}=\vec{d}\}\right) = \frac{f_{\vec{D}}\left(\vec{d} \,\middle|\, \{(M,\Sigma)=(\mu,\sigma)\}\right) \cdot f_{M\Sigma}(\mu,\sigma)}{f_{\vec{D}}(\vec{d})}$$

OR, THE MARGINAL LIKELIHOOD OF H

WHERE THE MARGINAL DENSITY OF THE DATA — WHICH INCORPORATES OUR HYPOTHESIS OF THE NORMAL DISTRIBUTION, AND OUR PRIOR DISTRIBUTION OF (M,Σ), AND IS THEREFORE ALSO CALLED THE EVIDENCE FOR THE MODEL GIVEN THE DATA — IS WRITTEN:

(PRIMES TO INDICATE THESE ARE (DUMMY) INTEGRATION VARIABLES NOT THE (μ,σ) IN THE DOMAIN OF: $f_{M\Sigma}(\mu,\sigma|D)$)

$$f_{\vec{D}}(\vec{d}) = \iint f_{\vec{D}}\left(\vec{d} \,\middle|\, \{(M,\Sigma)=(\mu',\sigma')\}\right) f_{M\Sigma}(\mu',\sigma')\, d\mu'\, d\sigma'$$

AND THE CONDITIONAL DENSITY OF $\vec{D}$ — CONDITIONED ON THE VALUES OF (M,Σ), AND EVALUATED AT THE OBSERVED DATA SET, $\vec{d}$ — BUT VIEWED AS A FUNCTION OF THE PARAMETERS (μ,σ), IS CALLED THE LIKELIHOOD FUNCTION (OF THE PARAMETERS, GIVEN THE OBSERVED DATA):

$$\mathcal{L}(\mu,\sigma;\vec{d}) = f_{\vec{D}}\left(\vec{d} \,\middle|\, \{(M,\Sigma)=(\mu,\sigma)\}\right)$$

IN THE BOXED SOLUTION ABOVE, WE ARE ASSUMING A FIXED VALUE OF $\vec{d}$ (THE OBSERVED DATA) BUT THE POSTERIOR DENSITY OF (M,Σ) IS A FUNCTION OF (μ,σ), WHOSE VALUE FOR A PARTICULAR (μ,σ) IS FOUND BY EVALUATING THE RHS AT (μ,σ).

- **General Formalism**: To make explicit that we always have some assumptions about the system in question — e.g., data from a certain type of distribution, number of data points — we will condition distributions on a hypothesis-model event, H. We will write the set of all parameters as a vector of random variables, $\vec{\Theta}$, with (prior) joint distribution $f_{\vec{\Theta}}(\vec{\theta})$. The observed data, $\vec{d}$, will be assumed to be a sample from the joint distribution of a random variable vector $\vec{D}$. Thus Bayes theorem is written:

$$f_{\vec{\Theta}}(\vec{\theta} \mid \{\vec{D} = \vec{d}\} \cap H) = \frac{f_{\vec{D}}(\vec{d} \mid \{\vec{\Theta} = \vec{\theta}\} \cap H) \cdot f_{\vec{\Theta}}(\vec{\theta} \mid H)}{\int f_{\vec{D}}(\vec{d} \mid \{\vec{\Theta} = \vec{\theta}'\} \cap H) \cdot f_{\vec{\Theta}}(\vec{\theta}' \mid H)\, d\vec{\theta}'}$$

Likelihood function, $\mathcal{L}(\vec{\theta})$ — Prior density — Posterior density — $f_D(\vec{d} \mid H)$ — Marginal distribution of the data ... more common: Marginal likelihood of (or evidence for) H

We don't usually make assumptions about the form of $f_{\vec{\Theta}}(\vec{\theta} \mid \cdots)$, but:

- Choose a broad enough prior $f_{\vec{\Theta}}(\vec{\theta} \mid H)$ to allow for all values of interest, e.g., uniform distribution on $\mathcal{U} \subseteq \mathbb{R}^d$
- The assumptions about the distribution of $\vec{D}$ (the likelihood function) will shape $f_{\vec{\Theta}}(\cdots)$

- **Example**: Consider 6 data points drawn from a normal distribution.

[I drew these from $N(\pi, \sqrt{\pi})$]

$$\vec{d} = [3.81, 6.96, 7.02, -0.86, 5.15, 3.87]$$

What can we determine about that distribution? We will assume that the elements of $\vec{D}$ are independent

AND IDENTICALLY DISTRIBUTED RANDOM VARIABLES, SO:

$$f_{\vec{D}}(\vec{d}) = f_{D_1}(d_1) \cdot f_{D_2}(d_2) \cdot \cdots \cdot f_{D_6}(d_6)$$

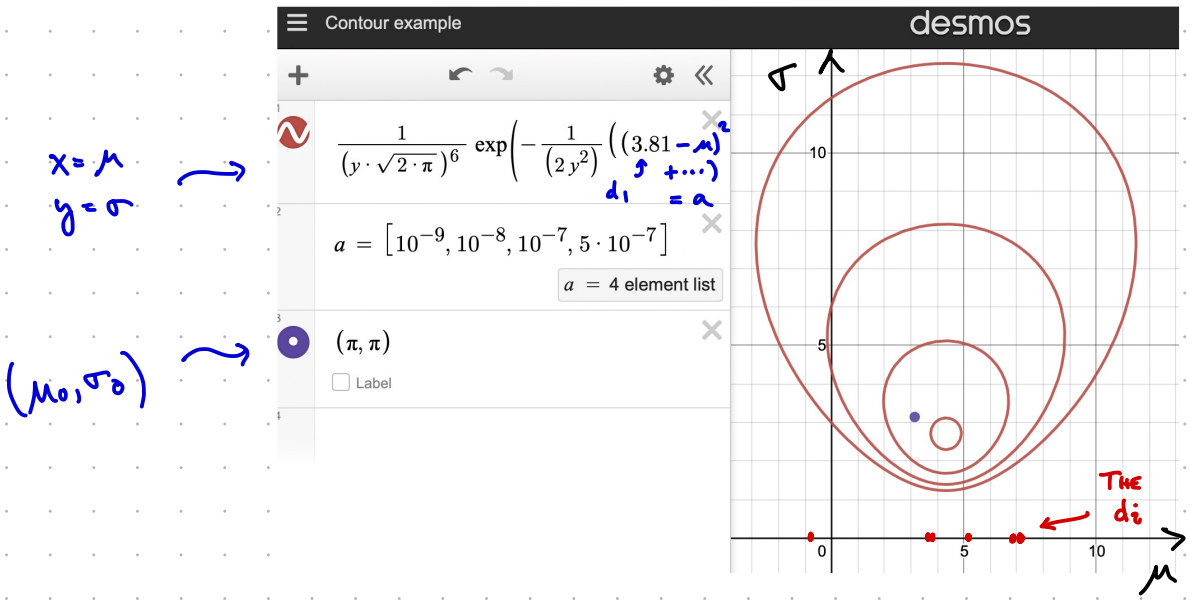
WITH $f_{D_i} \sim \mathcal{N}(\mu, \sigma^2)$ WHERE (THE DISTRIBUTIONS OF) μ AND σ ARE TO BE INFERRED FROM THE DATA. THE LIKELIHOOD FUNCTION IS THEREFORE:

$$\mathcal{L}(\mu,\sigma) = f_{\vec{D}}\left(\vec{d} \mid \{(M,\Sigma) = (\mu,\sigma)\} \frown H\right)$$

$$= \prod_{i=1}^{6} \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(d_i-\mu)^2}{2\sigma^2}\right]$$

$$= \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^6 \exp\left[-\sum_{i=1}^{6} \frac{(d_i-\mu)^2}{2\sigma^2}\right]$$

THE CONTOUR PLOT OF $\mathcal{L}(\mu,\sigma)$ LOOKS LIKE THIS:



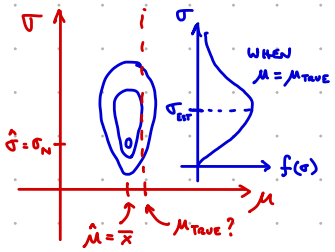
WHERE WE SEE THAT THE ACTUAL $(\mu_0,\sigma_0) = (\pi,\pi)$ WITH WHICH I GENERATED THE 6 DATA POINTS LIES VERY CLOSE TO THE PEAK OF THE LIKELIHOOD. IF WE MULTIPLY THE LIKELIHOOD BY, E.G., A UNIFORM DIST OVER $-10 \leq \mu \leq 20$ AND $0 \leq \sigma \leq 20$, THE QUALITATIVE PICTURE DOES NOT CHANGE. THEN DIVISION BY THE CONSTANT $f_D(\vec{d})$ WOULD ALSO NOT CHANGE THE PICTURE. SO THE LIKELIHOOD BASICALLY TELLS THE WHOLE STORY!

- THE PEAK OF THE LIKELIHOOD FUNCTION WILL GIVE YOU (OBVIOUSLY) THE VALUES FOUND BY MAXIMUM LIKELIHOOD. THE VALUES AT THE PEAK ARE

$$\hat{\mu} = \bar{x} := \frac{1}{N}\sum_{i=1}^{N} d_i$$

$$\hat{\sigma}^2 = \sigma_N^2 := \frac{1}{N}\sum_{i=1}^{N}(d_i - \bar{x})^2$$

- THE SHAPE OF THE LIKELIHOOD FUNCTION IS NOT SYMMETRIC FOR SECTIONS OF CONSTANT σ. IT IS EASY TO SEE WHY: IF THE ACTUAL DISTRIBUTION MEAN IS OFFSET FROM $\hat{\mu}$ THEN THE OBSERVED DATA ARE FURTHER FROM μ_{TRUE}, ON AVERAGE, THAN FROM $\hat{\mu}$, AND THEY IMPLY A LARGER σ^2.



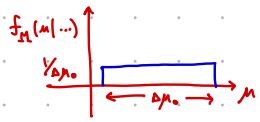
- ONE COULD ASK EITHER:

  (i) <u>I KNOW WHAT σ IS EXACTLY</u>. BUT WHAT DOES THE DATA TELL ME ABOUT μ?

  (ii) I DON'T KNOW μ AND DON'T CARE... TELL ME ABOUT THE <u>NOISE</u>. <u>WHAT DOES THE DATA SAY ABOUT σ?</u>

THE FIRST QUESTION IS ANSWERED BY: ($\{\Sigma = \sigma\}$ — NOW KNOWN)

$$f_M(\mu \mid D \cap \{\Sigma=\sigma\} \cap H) = \frac{f_{\vec{D}}\left(\vec{d} \mid \{(M,\Sigma)=(\mu,\sigma)\} \cap H\right) \cdot f_M(\mu \mid \{\sigma\} \cap H)}{f_{\vec{D}}\left(\vec{d} \mid \{\Sigma=\sigma\} \cap H\right)}$$



CHOOSING SOME BROAD PRIOR $f_M(\mu|\cdots)$, AND CALCULATING THE NORMALIZING CONSTANT:

$$f_{\vec{D}}\left(\vec{d} \mid \{\Sigma=\sigma\} \cap H\right) = \int f_{\vec{D}}\left(\vec{d} \mid \{(M,\Sigma)=(\mu,\sigma)\} \cap H\right) \times f_M(\mu \mid \{\sigma\} \cap H)\, d\mu$$

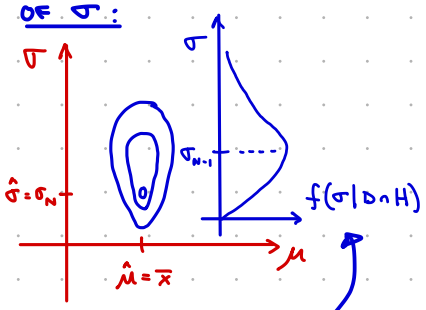
AND THE SECOND QUESTION IS ANSWERED BY:

$$f_{\Sigma}(\sigma \mid D \cap H) = \frac{f_{\vec{D}}(\vec{d} \mid \{\Sigma = \sigma\} \cap H) \cdot f_{\Sigma}(\sigma \mid H)}{f_{\vec{D}}(\vec{d} \mid H)}$$

NOTICE IN THIS SECOND QUESTION THAT THE LIKELIHOOD IS EXACTLY THE NORMALIZING CONSTANT OF THE FIRST QUESTION (THAT "WE DON'T CARE ABOUT") IF YOU DO THE MARGINALIZATION INTEGRATION OVER μ TO OBTAIN $f_{\vec{D}}(\vec{d} \mid \{\Sigma = \sigma\} \cap H)$ THEN THE PEAK OF THIS MARGINAL LIKELIHOOD FOR σ IS:

$$\max\left(f(\sigma \mid D \cap H)\right) = \sigma_{N-1} := \sqrt{\frac{\sum_{i=1}^{N}(d_i - \mu)^2}{N-1}}$$

MARGINALIZED DISTRIBUTION OF σ:



IT IS THE MARGINAL DISTRIBUTION AFTER MULTIPLYING BY PRIOR AND NORMALIZING. BEFORE THAT IT IS THE MARGINAL LIKELIHOOD.

LASTLY: THE NORMALIZING CONSTANT IN OUR SECOND QUESTION IS CALLED THE MARGINAL LIKELIHOOD OF H (THE HYPOTHESIS - MODEL):

STANDARD TERM

$$f_{\vec{D}}(\vec{d} \mid H) := \int\int \left[f_D(\vec{d} \mid \{M, \Sigma = \mu, \sigma\} \cap H) \times \underbrace{f_M(\mu \mid \{\Sigma = \sigma\} \cap H) \cdot f_{\Sigma}(\sigma \mid H)}_{\text{OR: } f_{M\Sigma}(\mu, \sigma \mid H)} \right] d\mu \, d\sigma$$

NON STANDARD

IT IS ALSO SOMETIMES CALLED THE "EVIDENCE FOR H" (GIVEN THE OBSERVED DATA).

- EXAMPLE: A SOURCE EMITS PARTICLES THAT DECAY AT POSITIONS THAT ARE EXPONENTIALLY DISTRIBUTED WITH LENGTH SCALE λ BUT ARE ONLY OBSERVED IN THE WINDOW:

$$(x_{min}, x_{max}) = (a, b) = (1, 20)$$

GIVEN N OBSERVATIONS $\{x_1, x_2, \ldots, x_N\}$ WHAT IS λ?

WE START BY WRITING DOWN OUR GIVEN ASSUMPTION (THE ASSUMPTIONS OF OUR HYPOTHESIS MODEL). FOR A SINGLE OBSERVED DECAY, THE POSITION IS DRAWN FROM:

$$f_X(x \mid \{\Lambda=\lambda\} \cap H) := \begin{cases} \frac{1}{Z(\lambda)} \exp\left(-\frac{x}{\lambda}\right) & (x \in (a,b)) \\ 0 & \text{OTHERWISE} \end{cases}$$

WHERE THE NORMALIZING CONSTANT IS:

$$Z(\lambda) = \int_a^b \exp\left(\frac{-x}{\lambda}\right) dx = \left[-\lambda e^{-\frac{x}{\lambda}}\right]_a^b = \lambda\left[e^{-a/\lambda} - e^{-b/\lambda}\right]$$

LET $\vec{x} = [x_1, x_2, \ldots, x_N]$ BE THE DATA, AND ASSUME THAT THE RANDOM VARIABLES FOR EACH ARE INDEPENDENT AND IDENTICALLY-DISTRIBUTED, SUCH THAT $f_{\vec{X}} = (f_X)^N$, I.E.

$$f_{\vec{X}}(\vec{x} \mid \{\Lambda=\lambda\} \cap H) = \frac{1}{[Z(\lambda)]^N} \exp\left[-\sum_{i=1}^{N} \frac{x_i}{\lambda}\right]$$

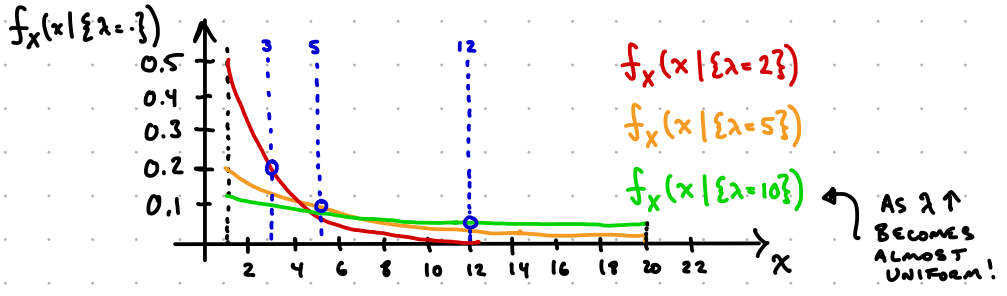
WE CAN THEN WRITE DOWN BAYES THEOREM:

$$f_\Lambda(\lambda \mid \{\vec{X}=\vec{x}\} \cap H) = \frac{f_{\vec{X}}(\vec{x} \mid \{\Lambda=\lambda\} \cap H) \cdot f_\Lambda(\lambda \mid H)}{\int_0^\infty f_{\vec{X}}(\vec{x} \mid \{\Lambda=\lambda\} \cap H) \cdot f_\Lambda(\lambda \mid H)\, d\lambda}$$

THIS CAN VERY EASILY BE TYPED INTO A COMPUTER (SPECIFICALLY JUST THE LIKELIHOOD FUNCTION, WHICH GIVES ALL INTERESTING INFORMATION FOR $f_\Lambda(\lambda \mid \ldots)$ GIVEN A UNIFORM PRIOR). BUT LET'S GET SOME INTUITION FOR THE LIKELIHOOD FOR A FEW DATA POINTS, E.G.,

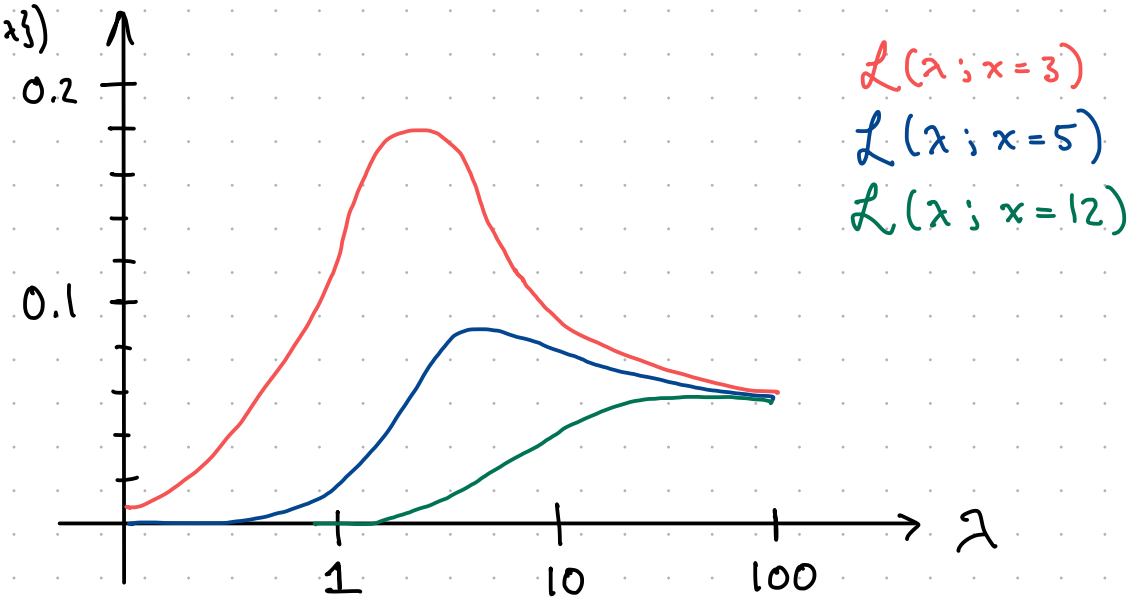
$$N = 3 \quad , \quad \vec{x} = (3, 5, 12)$$

IF WE PLOT SOME TRIAL EXPONENTIALS, THERE ARE SOME λ VALUES THAT ARE GOOD FOR SOME DATA POINTS



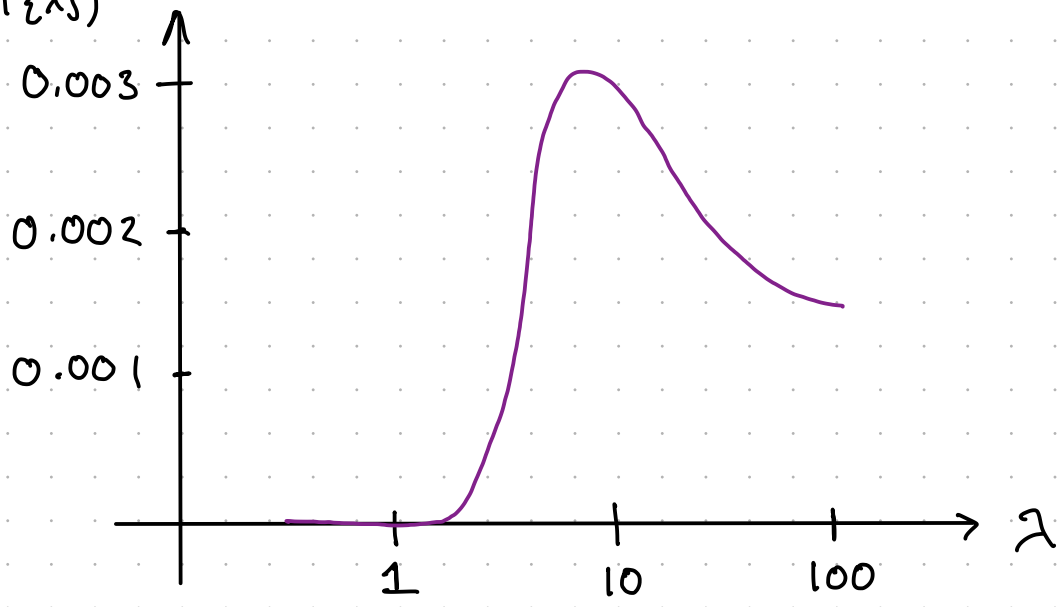
NOW, THE LIKELIHOOD IS JUST THE EXPONENTIAL FUNCTION, BUT VIEWED AS A FUNCTION OF λ (NOTE, THOUGH, THE DEPENDENCE OF Σ ON λ AS WELL). WE CAN PLOT THE LIKELIHOODS FOR EACH DATA POINT INDIVIDUALLY:

$\mathcal{L}(\lambda; x) = f_X(x|\{\lambda\})$



AMAZINGLY, THE SINGLE DATA POINT $x=3$ PREDICTS A WELL-DEFINED PEAK. WE CAN THEN PLOT THEIR PRODUCT, THE OVERALL $N=3$ LIKELIHOOD:

$\mathcal{L}(\lambda) = f_{\vec{X}}(\vec{x}|\{\lambda\})$



WHICH GIVES US THE MAXIMUM AMOUNT OF INFORMATION ABOUT "WHAT IS λ?" THAT CAN BE HAD FROM THESE THREE DATA POINTS. (JUST GIVING $\hat{\lambda}$ IS TOO LITTLE INFO!).

MODEL COMPARISON:

[FINISH WRITING UP MACKAY LECTURE 10]

2022/03/05

Kolmogorov has an extensive discussion of independent experiments, i.e., partitions of M.

Independent Random Variables

Two random variables X and Y are said to be **independent** if, for any two Borel sets $A, B \subseteq \mathbb{R}$:

$$\text{“ } P(X \in A \cap Y \in B) = P(X \in A) \cdot P(Y \in B) \text{ ”}$$

In other words, two random variables are independent if, for all A and B, the events

$$E_A = \{e \in \mathcal{E} \mid X(e) \in A\}$$

$$E_B = \{f \in \mathcal{E} \mid Y(f) \in B\}$$

are independent.

- If X and Y are independent random variables w/ probability densities $f_X(x)$ and $f_Y(y)$, then:

$$f_{XY}(x,y) = f_X(x) \cdot f_Y(y)$$

- Random variables that are not independent are called **dependent**.