STA257 (Probability and Statistics I) Lecture Notes, Fall 2023

by Jeffrey S. Rosenthal, University of Toronto, www.probability.ca

(Last updated: December 4, 2023)

Note: I will update these notes regularly, by posting them on the course web page each evening after lectures. However, they are just rough, point-form notes, with no guarantee of completeness or accuracy. They should in no way be regarded as a substitute for attending and learning from all the lectures, studying the course textbook, or doing the suggested homework exercises.

Introduction

- Course Information: See the course web page at: probability.ca/sta257

- Who here is doing a specialist or major program involving: Statistics / Data Science? Mathematics? Actuarial Science? Computer Science? Economics/Commerce? Physics/Chemistry/Biology? Education? Psychology/Sociology? Engineering? Other?

- Who here has seen probabilities in elementary school? high school? STA130?

  → Don't worry, we will start from scratch. (Just need math.)

- Life is full of randomness and uncertainty: lotteries, card games, computer games, gambling, weather, TTC, airplanes, friends, jobs, classes, science, finance, elections, diseases, safety/risk, demographics, internet routing, legal cases, ... whenever we're not sure of the outcome or what will happen next.

- Lots of interesting probability questions to solve! Such as ...

  → What's the probability you'll win the Lotto Max jackpot, i.e. that you will choose the correct 7 distinct numbers between 1 and 50?

  → If 200 students each flip a fair coin, then how many Heads is most likely? What's the probability of more than 150 Heads?

  → If you repeatedly roll a fair 6-sided die [show], then how many rolls will there be on average before the first 5?

  → At a party of 40 people, what is the probability that some pair of them have the same birthday?

  → If a disease affects one person in a thousand, and a test for the disease has 99% accuracy, and you test positive, then what is the probability you have the disease?

  → If you pick a number uniformly at random between 0 and 1, then what is the probability that you pick exactly the number 3/4?

  → Three-Card Challenge. [demonstration] What are the probabilities of the initial (front) colour? Then, what are the probabilities of the back colour?

- History of Mathematical Probability Theory (in brief):

  → Mathematics is very precise and certain. For thousands of years, it simply ignored the uncertainty of probabilities.

  → Then, in 1654, the French writer Antoine Gombaud (the "Chevalier de Méré") asked the mathematician Pierre de Fermat some gambling questions:

→ Which is more likely (or are they the same) (and are they more than 50%):
(a) Get at least one six when rolling a fair six-sided die 4 times; or
(b) Get at least one pair of sixes when rolling two fair six-sided dice 24 times?

→ He thought (a) was $4 \times (1/6) = 2/3$, and (b) was $24 \times (1/36) = 2/3$. Correct?

→ Also: (c) Suppose a gambler is playing a best-of-seven match, where whoever wins 4 (fair) games first in the winner, and so far they have won 3 times and lost 1, but then the match gets interrupted. What is the probability that they would have won the match, if it had been allowed to continue?

→ Fermat then corresponded with the mathematician Blaise Pascal to solve these questions (later!), and mathematical probability theory was born!

- So, can probabilities be studied mathematically?

→ Can we use certain mathematics to study the uncertainty of probabilities?

→ Yes! That's why we're here! To be certain about our uncertainty!

→ But we have to define our terms carefully ...

Sample Space

- The first part of any probability model is the sample space, written S, which is the set of all possible outcomes.

→ e.g. flip a coin: $S = \{\text{Heads, Tails}\}$, or $S = \{H, T\}$.

→ e.g. flip a coin three times in a row:
$S = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}$.

→ Or, if we only care about the number of Heads: $S = \{0, 1, 2, 3\}$.

→ e.g. tonight's dinner: $S = \{\text{Beef, Chicken, Fish}\}$.

→ e.g. the number of bees I will see on my walk home: $S = \{0, 1, 2, 3, \ldots\}$.

→ e.g. the price of IBM stock next month: $S = [0, \infty)$.

→ e.g. the height (in cm) of the next student I meet: $S = (0, \infty)$. (Or ...)

→ e.g. your grade in this class: $S = \{0, 1, 2, 3, \ldots, 100\}$.

→ e.g. roll one six-sided die: $S = \{1, 2, 3, 4, 5, 6\}$.

→ e.g. roll two six-sided dice: $S = \{1, 2, 3, 4, 5, 6\}^2$, i.e.
$S = \{11, 12, 13, 14, 15, 16,\ 21, 22, 23, 24, 25, 26,$
$31, 32, 33, 34, 35, 36,\ 41, 42, 43, 44, 45, 46,$
$51, 52, 53, 54, 55, 56,\ 61, 62, 63, 64, 65, 66\}$.

→ Or, if we only care about the sum, instead maybe take $S = \{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$.

→ e.g. "Pick any integer between 1 and 10": $S = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$.

→ e.g. "Pick any number between 0 and 1": $S = [0, 1]$. (important case!)

- Summary: The sample space S can be any non-empty set which contains all of the possible outcomes. Simple!

- But it gets more interesting when we also have ...

Probabilities and Events

- An event A is "any" subset $A \subseteq S$.
- For any event A, we can define the probability $\mathrm{P}(A)$ that it will occur.

→ e.g. flip a "fair" coin: $\mathrm{P}(H) = \mathrm{P}(T) = 1/2$.

→ (Note: We often use e.g. "$\mathrm{P}(H)$" as shorthand for "$\mathrm{P}(\{H\})$", etc.)

→ e.g. roll a fair six-sided die: $\mathrm{P}(1) = \mathrm{P}(2) = \mathrm{P}(3) = \mathrm{P}(4) = \mathrm{P}(5) = \mathrm{P}(6) = 1/6$.

→ e.g. tonight's dinner: maybe P(Beef)=0.40, P(Chicken)=0.15, and P(Fish)=0.45.

→ (Note: We could also write P(Fish) = 45%, etc. Usually percentages are good for intuition, but pure probabilities (not percentages) are better for calculation.)

→ e.g. flip three fair coins: $\mathrm{P}(HHH) = \mathrm{P}(HHT) = \ldots = \mathrm{P}(TTT) = 1/8$.

→ e.g. roll two fair dice: $\mathrm{P}(11) = \mathrm{P}(12) = \ldots = \mathrm{P}(65) = \mathrm{P}(66) = 1/36$.

→ e.g. Pick any integer between 1 and 10. [Try it!]
Could be "uniform", i.e. $\mathrm{P}(1) = \mathrm{P}(2) = \ldots = \mathrm{P}(10) = 1/10$. Or instead, maybe ... P(3)=P(6)=P(7)=0.2, and P(5)=0.1, and P(1)=P(2)=P(4)=P(8)=P(9)=P(10)=0.05.

→ e.g. Pick any number between 0 and 1, "uniformly":
$\mathrm{P}([0, 1/2]) = 1/2$, $\mathrm{P}([1/2, 1]) = 1/2$, $\mathrm{P}([0, 1/3]) = 1/3$, $\mathrm{P}([1/3, 2/3]) = 1/3$,
and in general $\mathrm{P}([a, b]) = b - a$ whenever $0 \leq a \leq b \leq 1$. [Diagram]

Basic Properties of Probabilities

- Let's begin with a specific example (and then we will generalise):
- e.g. tonight's dinner, with P(Beef)=0.40, P(Chicken)=0.15, and P(Fish)=0.45.

→ Probability of Beef or Chicken = $\mathrm{P}(\{\text{Beef, Chicken}\}) = \mathrm{P}(\{\text{Beef}\}) + \mathrm{P}(\{\text{Chicken}\})$ $= 0.40 + 0.15 = 0.55$.

→ Probability of any dinner = Probability of Beef or Chicken or Fish = $\mathrm{P}(\{\text{Beef, Chicken, Fish}\}) = \mathrm{P}(\{\text{Beef}\}) + \mathrm{P}(\{\text{Chicken}\}) + \mathrm{P}(\{\text{Fish}\}) = 0.40 + 0.15 + 0.45 = 1$.

→ Probability of No dinner = $\mathrm{P}(\emptyset) = 0$.

- In general, certain properties must hold for any probability model ("axioms"):
- If A is an event, then $0 \leq \mathrm{P}(A) \leq 1$.
- If $A = S$ is the event corresponding to all outcomes, then $\mathrm{P}(A) = \mathrm{P}(S) = 1$.
- Or, if $A = \emptyset$ is the event corresponding to no outcomes, then $\mathrm{P}(A) = \mathrm{P}(\emptyset) = 0$.
- Additivity: If A and B are disjoint events (i.e. $A \cap B = \emptyset$), e.g. $A = \{\text{Beef}\}$ and $B = \{\text{Chicken}\}$, then $\mathrm{P}(A \cup B) = \mathrm{P}(A) + \mathrm{P}(B)$.
- More generally, if $A_1, A_2, A_3, \ldots$ are any sequence (finite or infinite) of disjoint events (i.e. $A_i \cap A_j = \emptyset$ whenever $i \neq j$), then $\mathrm{P}\left(\bigcup_i A_i\right) = \sum_i \mathrm{P}(A_i)$.

→ So, in particular, since $\mathrm{P}(S) = 1$, all of the probabilities have to add up to 1.

→ e.g. P(Heads) + P(Tails) = 0.5 + 0.5 = 1.

→ e.g. P(Beef) + P(Chicken) + P(Fish) = 0.40 + 0.15 + 0.45 = 1.

Suggested Homework: 1.2.1, 1.2.2, 1.2.3, 1.2.4, 1.2.8, 1.2.9, 1.2.10, 1.2.11, 1.2.12.

Derived Properties of Probabilities

- Once we know the above properties, then we can use them to prove others too:

- Fact: If A^C is the complement of A, i.e. the set of all outcomes which are not in A, then $\mathrm{P}(A^C) = 1 - \mathrm{P}(A)$. (Important! Remember this! Use this!)

  → Proof: Note that A and A^C are disjoint, so $\mathrm{P}(A \cup A^C) = \mathrm{P}(A) + \mathrm{P}(A^C)$. But $\mathrm{P}(A \cup A^C) = \mathrm{P}(S) = 1$, so $1 = \mathrm{P}(A) + \mathrm{P}(A^C)$, i.e. $\mathrm{P}(A^C) = 1 - \mathrm{P}(A)$. ∎

  → e.g. P(Fish) = P(not Beef or Chicken) $= 1 - \mathrm{P}$(Beef or Chicken) $= 1 - 0.55 = 0.45$.

- Fact: For any events A and B [Diagram], $\mathrm{P}(A) = \mathrm{P}(A \cap B) + \mathrm{P}(A \cap B^C)$. (∗)

  → Proof: The events $A \cap B$ and $A \cap B^C$ are disjoint, and $(A \cap B) \cup (A \cap B^C) = A$ [Diagram], so by additivity, $\mathrm{P}(A \cap B) + \mathrm{P}(A \cap B^C) = \mathrm{P}(A)$. ∎

  → e.g. integer between 1 and 10: P(even) = P(even and ≤ 4) + P(even and ≥ 5) $= \mathrm{P}(\{2, 4\}) + \mathrm{P}(\{6, 8, 10\})$.

END MONDAY #1

- Re-arranging (∗) also gives that: $\mathrm{P}(A \cap B^C) = \mathrm{P}(A) - \mathrm{P}(A \cap B)$. (∗∗)

- Fact: If $A \supseteq B$, then $\mathrm{P}(A) = \mathrm{P}(B) + \mathrm{P}(A \cap B^C)$. (∗ ∗ ∗)

  → Proof: This follows from (∗), since if $A \supseteq B$, then $A \cap B = B$. ∎

  → e.g. integer between 1 and 10: P(≤ 7) = P(≤ 4) + P(≤ 7 but ≥ 5).

- Monotonicity: If $A \supseteq B$, then $\mathrm{P}(A) \geq \mathrm{P}(B)$. (Remember this!)

  → Proof: We must have $\mathrm{P}(A \cap B^C) \geq 0$, so from (∗∗∗), $\mathrm{P}(A) = \mathrm{P}(B) + \mathrm{P}(A \cap B^C) \geq \mathrm{P}(B) + 0 = \mathrm{P}(B)$. ∎

  → e.g. P({Beef, Chicken}) $= 0.55 \geq 0.40 =$ P({Beef}).

- Law of Total Probability – Unconditioned Version: Suppose $A_1, A_2, \ldots$ are a sequence (finite or infinite) of events which form a partition of S, i.e. they are disjoint ($A_i \cap A_j = \emptyset$ for all $i \neq j$) and their union equals the entire sample space ($\bigcup_i A_i = S$), and let B be any event. Diagram:

Then $\mathrm{P}(B) = \sum_i \mathrm{P}(A_i \cap B)$. That is: $\mathrm{P}(B) = \mathrm{P}(A_1 \cap B) + \mathrm{P}(A_2 \cap B) + \ldots$.

→ Proof: Since the $\{A_i\}$ are disjoint, and $A_i \cap B \subseteq A_i$, therefore the $\{A_i \cap B\}$ are also disjoint. Furthermore, since $\bigcup_i A_i = S$, therefore $\bigcup_i (A_i \cap B) = S \cap B = B$. Hence, $\mathrm{P}(B) = \mathrm{P}\Big(\bigcup_i (A_i \cap B)\Big) = \sum_i \mathrm{P}(A_i \cap B)$. ∎

→ e.g. integer between 1 and 10: Suppose $A_1 = \{\leq 4\} = \{1, 2, 3, 4\}$, and $A_2 = \{\geq 5\} = \{5, 6, 7, 8, 9, 10\}$, and $B = \{\text{even}\} = \{2, 4, 6, 8, 10\}$. Then P(even) = P(even and ≤ 4) + P(even and ≥ 5), i.e. $\mathrm{P}(\{2, 4, 6, 8, 10\}) = \mathrm{P}(\{2, 4\}) + \mathrm{P}(\{6, 8, 10\})$.

• Principle of Inclusion-Exclusion: $\mathrm{P}(A \cup B) = \mathrm{P}(A) + \mathrm{P}(B) - \mathrm{P}(A \cap B)$.

→ (Of course, if they're disjoint ($A \cap B = \emptyset$), then $\mathrm{P}(A \cup B) = \mathrm{P}(A) + \mathrm{P}(B)$.)

→ Intuition: $\mathrm{P}(A) + \mathrm{P}(B)$ counts each element of $A \cap B$ twice, so we have to subtract one of them off.

→ Proof: The events $A \cap B$, and $A \cap B^C$, and $A^C \cap B$, are all disjoint, and their union is $A \cup B$. [Diagram.] Hence, $\mathrm{P}(A \cup B) = \mathrm{P}(A \cap B) + \mathrm{P}(A \cap B^C) + \mathrm{P}(A^C \cap B)$.
Then, from (**), $\mathrm{P}(A \cap B^C) = \mathrm{P}(A) - \mathrm{P}(A \cap B)$ and $\mathrm{P}(A^C \cap B) = \mathrm{P}(B) - \mathrm{P}(A \cap B)$.
Hence, $\mathrm{P}(A \cup B) = \mathrm{P}(A \cap B) + \big[\mathrm{P}(A) - \mathrm{P}(A \cap B)\big] + \big[\mathrm{P}(B) - \mathrm{P}(A \cap B)\big]$
$= \mathrm{P}(A) + \mathrm{P}(B) - \mathrm{P}(A \cap B)$. ∎

→ e.g. integer between 1 and 10: P(even or ≤ 4) = P(even) + P(≤ 4) − P(even and ≤ 4) = $\mathrm{P}(\{2,4,6,8,10\}) + \mathrm{P}(\{1,2,3,4\}) - \mathrm{P}(\{2,4\})$.

→ Or, P(even or perfect square) = P(even) + P(perfect square) − P(even and perfect square) = $\mathrm{P}(\{2,4,6,8,10\}) + \mathrm{P}(\{1,4,9\}) - \mathrm{P}(\{4\})$.

• $\mathrm{P}(A \cap B) \geq 0$, so $\mathrm{P}(A \cup B) = \mathrm{P}(A) + \mathrm{P}(B) - \mathrm{P}(A \cap B) \leq \mathrm{P}(A) + \mathrm{P}(B)$. Indeed:

• Subadditivity: For any sequence of events $A_1, A_2, \ldots$, not necessarily disjoint, we still always have $\mathrm{P}(A_1 \cup A_2 \cup \ldots) \leq \mathrm{P}(A_1) + \mathrm{P}(A_2) + \ldots$.

→ (Of course, it would be equal if they are disjoint.)

→ Proof: Let $B_1 = A_1$, and $B_2 = A_2 \cap (A_1)^C$, and $B_3 = A_3 \cap (A_1 \cup A_2)^C$, and $B_4 = A_4 \cap (A_1 \cup A_2 \cup A_3)^C$, and so on. (That is, each new B_n is the part of A_n which is not already part of $A_1, \ldots, A_{n-1}$.) Diagram:

The $\{B_i\}$ are disjoint, and $\bigcup_i B_i = \bigcup_i A_i$. Also $B_i \subseteq A_i$ so $\mathrm{P}(B_i) \leq \mathrm{P}(A_i)$. Hence, $\mathrm{P}(A_1 \cup A_2 \cup \ldots) = \mathrm{P}(B_1 \cup B_2 \cup \ldots) = \mathrm{P}(B_1) + \mathrm{P}(B_2) + \ldots \leq \mathrm{P}(A_1) + \mathrm{P}(A_2) + \ldots$. ∎

→ Alternative proof (for a finite number of events): Use induction! For $n = 2$ events, this follows from Inclusion-Exclusion. Then for $n \geq 3$ events, $\mathrm{P}(A_1 \cup \ldots \cup A_n) = \mathrm{P}\big((A_1 \cup \ldots \cup A_{n-1}) \cup A_n\big)$, which by Inclusion-Exclusion is $\leq \mathrm{P}\big(A_1 \cup \ldots \cup A_{n-1}\big) + \mathrm{P}(A_n)$, which by induction is $\leq \big(\mathrm{P}(A_1) + \ldots + \mathrm{P}(A_{n-1})\big) + \mathrm{P}(A_n)$. ∎

→ e.g. integer between 1 and 10: P(even or ≤ 4) $\leq$ P(even) + P(≤ 4), i.e. $\mathrm{P}(\{1,2,3,4,6,8,10\}) \leq \mathrm{P}(\{2,4,6,8,10\}) + \mathrm{P}(\{1,2,3,4\})$.

Suggested Homework: 1.2.13, 1.2.14, 1.2.15. (more theoretical)

Suggested Homework: 1.3.1, 1.3.2, 1.3.3, 1.3.4, 1.3.5, 1.3.7, 1.3.8, 1.3.9.

Optional: A more general Inclusion-Exclusion formula is in Challenge 1.3.10.

Uniform Probabilities on Finite Spaces

• Suppose $S = \{s_1, s_2, \ldots, s_n\}$ is some finite sample space, of finite size $|S| = n$, and each element is equally likely.

→ Then $\mathrm{P}(s_1) = \mathrm{P}(s_2) = \ldots = \mathrm{P}(s_n) = 1/n$. ("discrete uniform distribution")

→ And for any event $A = \{a_1, a_2, \ldots, a_k\}$, by additivity we have

$$\mathrm{P}(A) \;=\; \mathrm{P}(a_1) + \mathrm{P}(a_2) + \ldots + \mathrm{P}(a_k) \;=\; \frac{1}{n} + \frac{1}{n} + \ldots + \frac{1}{n} \;=\; \frac{k}{n} \;=\; \frac{|A|}{|S|}\,.$$

→ So, in this case, we just need to count the number of elements in A, and divide that by the number of elements in S. Easy!?! Sometimes!

- e.g. Roll a fair six-sided die. What is $\mathrm{P}(\geq 5)$?

→ Here $S = \{1, 2, 3, 4, 5, 6\}$ so $|S| = 6$. All equally likely.

→ Also $A = \{5, 6\}$ so $|A| = 2$.

→ So, $\mathrm{P}(\geq 5) = \mathrm{P}(A) = |A| \,/\, |S| = 2/6 = 1/3$. Easy!

- e.g. Roll one fair six-sided die, and flip two fair coins.

What is P(# Heads = Number Showing On The Die)?

→ Here $S = \{1HH, 1HT, 1TH, 1TT, 2HH, \ldots, 6TT\}$. All equally likely.

→ But what is $|S|$?

→ Multiplication Principle: If S is made up by choosing one element of each of the subsets $S_1, S_2, \ldots, S_k$, i.e. if $S = S_1 \times S_2 \times \ldots \times S_k$, then what is $|S|$? Well, ... $|S| = |S_1|\,|S_2|\,\ldots\,|S_k|$.

→ In our example, $S_1 = \{1, 2, 3, 4, 5, 6\}$, and $S_2 = \{H, T\}$, and $S_3 = \{H, T\}$, so $|S| = |S_1|\,|S_2|\,|S_3| = 6 \cdot 2 \cdot 2 = 24$.

→ And what about A? Well, think about the possibilities ...

$A = \{1HT, 1TH, 2HH\}$. (No other combination works. Why?) So, $|A| = 3$.

→ Hence, P(# Heads = Number Showing On The Die) $= |A| \,/\, |S| = 3/24 = 1/8$.

- e.g. Roll three fair six-sided dice. What is $\mathrm{P}(\text{sum} \geq 17)$?

→ Here $S = \{1, 2, 3, 4, 5, 6\}^3$ so $|S| = 6^3 = 216$. All equally likely.

→ But what is A? Think about it ...

Here $A = \{666, 566, 656, 665\}$ (why?), so $|A| = 4$.

→ So, $\mathrm{P}(\text{sum} \geq 17) = \mathrm{P}(A) = |A| \,/\, |S| = 4/216 = 1/54$.

→ Exercise: What about $\mathrm{P}(\text{sum} \geq 16)$? $\mathrm{P}(\text{sum} \geq 15)$?

- Chevalier's questions:

- (a) What is P(get at least one six when rolling a fair six-sided die 4 times)?

→ Here $S = \{1, 2, 3, 4, 5, 6\}^4$, so $|S| = 6^4 = 1296$. All equally likely.

→ And what is $|A|$? Tricky. Easier to consider ...

→ $A^C = \{\text{no sixes in four rolls}\} = \{1, 2, 3, 4, 5\}^4$, so $|A^C| = 5^4 = 625$.

→ So, $\mathrm{P}(A^C) = |A^C| \,/\, |S| = 5^4 \,/\, 6^4 = 625 \,/\, 1296 \doteq 0.482$.

→ So, $\mathrm{P}(A) = 1 - \mathrm{P}(A^C) \doteq 1 - 0.482 = 0.518$. More than 50%.

→ (Alternatively: By "independence" [later], $\mathrm{P}(A) = 1 - (5/6)^4 \doteq 0.518$.)

- (b) What is P(get at least one pair of sixes when rolling a pair of fair six-sided dice 24 times)?

→ Here $S = \left(\{1,2,3,4,5,6\}^2\right)^{24}$, so $|S| = (6^2)^{24} = 6^{48}$ ($>10^{37}$). All equally likely.

→ And what is $|A|$? Tricky. Again, easier to consider ...

→ A^C = {no pair of sixes in 24 rolls} = $\{11, 12, 13, \ldots, 64, 65\}^{24}$, so $|A^C| = 35^{24}$.

→ So, $\mathrm{P}(A^C) = |A^C| \,/\, |S| = 35^{24}/6^{48} \doteq 0.509$.

→ So, $\mathrm{P}(A) = 1 - \mathrm{P}(A^C) \doteq 1 - 0.509 = 0.491$. Less than 50%.

→ (Again, alternatively by independence [later], $\mathrm{P}(A) = 1 - (35/36)^{24} \doteq 0.491$.)

• (c) In a best-of-seven match with fair (50%) games, if a player has won 3 games and lost 1, then what is the probability they will win the match?

→ Various paths to victory: win right away, lose then win, etc. Tricky.

→ One solution: Pretend 3 more games will always be played. (Result same.)

→ Then $S = \{\text{Win, Lose}\}^3$, so $|S| = 2^3 = 8$, all equally likely.

→ What about A? Well, here $A^C = \{\text{Lose, Lose, Lose}\}$, so $|A^C| = 1$.

→ Hence, $\mathrm{P}(A^C) = |A^C|/|S| = 1/8$, and so $\mathrm{P}(A) = 1 - \mathrm{P}(A^C) = 7/8$.

→ Exercise: What if the player has won just 2 games and lost 1? (Trickier.)

Suggested Homework: 1.4.1, 1.4.2, 1.4.3, 1.4.9, 1.4.10, 1.4.11, 1.4.12, 1.4.13.

Warning about Non-Uniform Probabilities

• e.g. Roll two fair dice. What is P(sum is ≤ 3)?

→ POSSIBLE SOLUTION: The sum is in $S = \{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$. So, $|S| = 11$. And, the event "≤ 3" corresponds to $A = \{2, 3\}$, so $|A| = 2$. Hence, P(sum is ≤ 3) $= |A|/|S| = 2/11$. Right?

→ WRONG! These sums are not all equally likely, i.e. it is not uniform! So, $\mathrm{P}(A) \neq |A|/|S|$. That formula is only when all outcomes are equally likely. Important!

→ INSTEAD: Let S = {all ordered pairs of two dice}, i.e. $S = \{11, 12, 13, \ldots, 65, 66\}$. Then $|S| = 36$. Now each outcome in S is equally likely. And, now $A = \{11, 12, 21\}$. So, $\mathrm{P}(A) = |A|/|S| = 3/36 = 1/12$. Correct!

• And sometimes the sample space S is a discrete infinite set:

→ e.g. $S = \mathbf{N} := \{1, 2, 3, \ldots\}$, with $\mathrm{P}(i) = 2^{-i}$ for each $i \in S$.

→ Valid? Yes, since $2^{-i} \geq 0$, and $\sum_{i=1}^{\infty} 2^{-i} = \frac{2^{-1}}{1-2^{-1}} = 1$. (Geometric series.)

→ Then e.g. $\mathrm{P}(\text{Even Number}) = \sum_{i=2,4,6,\ldots} 2^{-i} = \frac{1}{4} + \frac{1}{16} + \frac{1}{64} + \ldots = \frac{1/4}{1-(1/4)} = 1/3$.

→ And, $\mathrm{P}(\leq 10) = \sum_{i=1}^{10} 2^{-i} = \frac{2^{-1}-2^{-11}}{1-2^{-1}} = \frac{(1/2)-(1/2048)}{1-(1/2)} = 1023/1024$. Close to 1.

→ On a discrete infinite space, cannot have a uniform distribution!

• Summary: Don't assume it's uniform when it isn't!

END WEDNESDAY #1

More Finite Uniform Probabilities

• e.g. Suppose there are ten people at a party, and you randomly pick three of the people, in order (1-2-3). What is the probability that your choices will also be the three richest people at the party (in the same order)?

→ S is the set of all ways of picking three people, in order. All equally likely.

→ But what is $|S|$?

→ The first person can be picked in 10 different ways.

→ Then, the second person can be picked in 9 different ways.

→ Then, the third person can be picked in 8 different ways.

→ So, $|S| = 10 \cdot 9 \cdot 8 = 720$.

→ Also, $|A| = 1$ since there is only one matching choice.

→ So, P(you picked the three richest, in order) $= |A|/|S| = 1/720$.

• More generally, the number of ways of picking k distinct items, in order, out of n items total, is equal to $n(n-1)(n-2)\dots(n-k+1) = n!/(n-k)!$. ("permutations")

→ In particular, if $k = n$, then the number of ways of picking all n items in order is equal to $n(n-1)(n-2)\dots(1) = n!$. ("n factorial")

• "The Birthday Problem": Suppose 40 (say) people at a party are each equally likely to be born on any one of 365 days of the year. Then what is the probability that at least one pair of them have the same birthday? (Any guesses?)

→ Here, S is the set of all 40-tuples of possible birthdays. All equally likely.

→ (Count them in order, since they might not all be distinct.)

→ So, by the Multiplication Principle, $|S| = 365^{40}$.

→ What about $|A|$? Not easy ...

→ Instead, consider A^C. (Then can use that $\mathrm{P}(A) = 1 - \mathrm{P}(A^C)$.)

→ A^C is the set of all ways of picking 40 distinct birthdays, in order.

→ So, $|A^C| = 365 \cdot 364 \cdot 363 \cdot \ldots \cdot 326 = 365! \,/\, 325!$.

→ So, $\mathrm{P}(A^C) = \big(365!/325!\big) \,/\, 365^{40} \doteq 0.109$.

→ So, $\mathrm{P}(A) = 1 - \mathrm{P}(A^C) \doteq 0.891$. Over 89%. Very likely! (Make a bet?)

→ (Generalisation to "C" people in the textbook's Challenge 1.4.21 ...)

• But suppose instead that we don't care about the order. Then, we have to divide by $k! = k(k-1)(k-2)\dots(2)(1)$, the number of different orderings of k items.

→ So, the number of ways of picking k distinct items out of n items total, ignoring order, is equal to $n(n-1)(n-2)\dots(n-k+1) \,/\, k! = n!/(n-k)!\,k!$. ("combinations"; "choose formula", or "binomial coefficient") Also written as: $\binom{n}{k}$.

• e.g. Suppose there are ten people at a party, and you randomly pick a collection of three of the people, but ignoring order. What is the probability that your choices will also be the three richest people at the party (in any order)?

→ S is all ways of picking three people (ignoring order). All equally likely.

→ But what is $|S|$?

→ Here $|S| = \binom{10}{3} = 10!\ /\ (7!\ 3!) = 120$.

→ And, again $|A| = 1$ since there is only one matching choice.

→ So, P(you picked the three richest, ignoring order) $= |A|/|S| = 1/120$.

→ Six times as large as before! Makes sense since $3! = 6$.

- e.g. Lotto Max jackpot:

→ Here $S = \{$all choices of 7 distinct numbers between 1 and 50$\}$.

→ All equally likely. And, we do not care about the order.

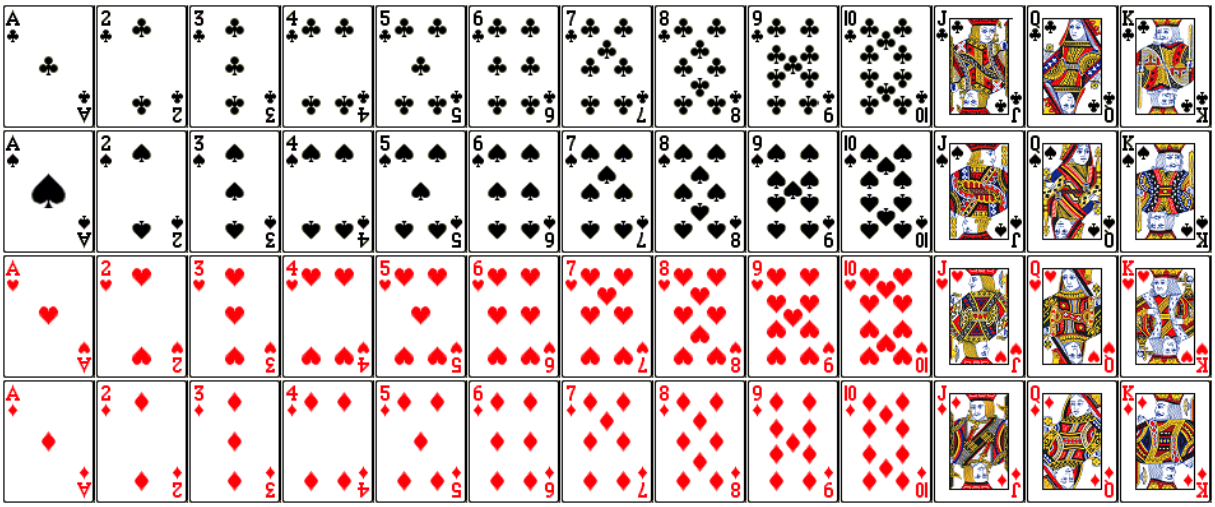
→ So, $|S| = 50!\ /\ (43!\ 7!) = 99,884,400 \doteq 100$ million.

→ Also, A is the one correct choice. So, $|A| = 1$.

→ So, P(jackpot) = P(choose the correct 7 distinct numbers between 1 and 50) $= |A|\ /\ |S| = 1/99,884,400 \doteq 1/100,000,000 = 0.000001\%$. Very small!

→ (For \$5, you get three choices of 7 numbers, which increases P(jackpot) to 3 / 99,884,400 = 1 /33,294,800 ... still very small ...)

- Recall that a standard deck of playing cards has four suits (Clubs, Spades, Hearts, Diamonds), and each suit has 13 ranks (A,2,3,4,5,6,7,8,9,10,J,Q,K), so 52 cards total:



- A card's value is its number, counting A as 1, J as 11, Q as 12, and K as 13.

- Suppose we pick one playing card from a standard deck, uniformly at random.

→ So S is the set of all cards in the deck, with $|S| = 52$, all equally likely.

→ Then what is P(Club or 7)? Can solve this directly, or ...

→ Here P(Club) = 13/52 = 1/4, and P(7) = 4/52 = 1/13.

→ Also, P(Club and 7) = P(7-of-Clubs) = 1/52.

→ So, by Inclusion-Exclusion, P(Club or 7) = P(Club) + P(7) − P(Club and 7) $= 1/4 + 1/13 - 1/52 = 16/52 = 4/13$.

- Or, suppose we draw a pair of distinct cards uniformly from a standard deck.

→ What is P(both are Face Cards), i.e. P(both are J/Q/K)?

→ Here $S = \{$all distinct pairs of cards, ignoring order$\}$.

→ So, $|S| = \binom{52}{2} = 52 \cdot 51/2 = 1326$.

→ And $A = \{$all distinct pairs of Face Cards$\}$, so $|A| = \binom{12}{2} = 12 \cdot 11/2 = 66$.

→ So, $\mathrm{P}(A) = |A|/|S| = 66/1326 \doteq 0.0498 \doteq 1/20$.

→ Alternatively, could let $S = \{$all distinct pairs of cards in order$\}$. Then $|S| = 52 \cdot 51 = 2652$, and $|A| = 12 \cdot 11 = 132$. So, $\mathrm{P}(A) = |A|/|S| = 132/2652$, which gives the same answer as before.

→ (Or, conditional probability [next]: $\mathrm{P}(A) = (12/52) \cdot (11/51) = 132/2652$.)

- e.g. Flip 4 fair coins. What is P(exactly 2 Heads)?

→ Here $S =$ all 4-tuples of H and T (in order). $|S| = 2^4 = 16$. All equally likely.

→ And $A =$ all 4-tuples with two H and two T. What is $|A|$?

→ Can write them all out [let's do it now]:

→ So $|A| = 6$, and $P(A) = |A|/|S| = 6/16 = 3/8$. Simpler way?

→ Each element of A can be specified by choosing which 2 of the 4 coins were H (without caring about the order).

→ So, $|A|$ = number of choices of 2 coins out of 4 $= \binom{4}{2} = 4!/((4-2)!\ 2!) = 24/(2 \cdot 2) = 6$, and $P(A) = |A|/|S| = 6/16$.

→ Same answer as before, but more systematic, and easier to use when we have lots of coins. Clear?

- e.g. Suppose we flip ten fair coins. What is P(exactly six Heads)?

→ S is the set of all "10-tuples" of H and T, i.e. length-10 sequences (in order) of H and T.

→ All equally likely. But what is $|S|$? Well, by the Multiplication Principle, $|S| = 2 \cdot 2 \cdot \ldots \cdot 2 = 2^{10} = 1024$.

→ What about $|A|$? Well, $A = \{HHHHHHTTTT, HHHHHTHTTT, \ldots, TTTTHHHHHH\}$. But how many elements does it include?

→ Well, an element of A is specified by "choosing" which 6 of the 10 coins are Heads. So, the size of A is equal to the corresponding binomial coefficient:

$$|A| = \binom{10}{6} = \frac{10!}{6!\ (10-6)!} = \frac{10!}{6!\ 4!} = \frac{10 \cdot 9 \cdot 8 \cdot 7}{4 \cdot 3 \cdot 2 \cdot 1} = \frac{5040}{24} = 210\,.$$

→ So, P(exactly six Heads) $= |A|\ /\ |S| = 210/1024 = 105/512 \doteq 0.205 = 20.5\%$.

- In general, if flip n fair coins, then P(exactly k Heads) $= \binom{n}{k}/2^n$, for $0 \le k \le n$.

→ (Special case of the "Binomial Distribution" – more later.)

Suggested Homework: 1.3.6, 1.4.4, 1.4.6, 1.4.7, 1.4.8, 1.4.15, 1.4.16, 1.4.17, 1.4.19, 1.4.21.

END MONDAY #2

Simulating Using the Computer Software "R"

- There is lots of computer software available for statistical computation. (Even spreadsheets etc.) One package used by most statisticians (and STA courses) is "R".
  - → Free and easy to install on any computer, e.g. on your laptop!
  - → For some basic info and links, see: http://probability.ca/Rinfo.html
  - → Also discussed in Appendix B of the textbook.
  - → In this course, you do not need to learn it.
  - → But I will use it for occasional demonstrations.
  - → It is interesting, and insightful, and used in other courses. [Try it!]
- For now, just a few simulation commands to get us started:
  - → sample(c("H","T"), 1) [one random sample from $\{H, T\}$]
  - → sample(1:6, 1) [one random sample from $\{1, 2, 3, 4, 5, 6\}$]
  - → sample(1:6, 3, replace=TRUE) [three samples, with replacement]
  - → sample(c("Beef","Chicken","Fish"), 1, prob=c(0.40,0.15,0.45)) [with probs]
  - → rgeom(1, 1/2) + 1 [sample where $P(i) = 2^{-i}$]

Conditional Probability

- e.g. Flip three fair coins.
  - → Then $S = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}$.
  - → All equally likely. So, P(first coin Heads) $= 4/8 = 1/2$.
  - → Suppose we are told that exactly 2 coins were Heads.
  - → Now what is the probability that the first coin was Heads?
  - → Well, the outcome must be in $\{HHT, HTH, THH\}$. Still all equally likely.
  - → And, two of these three outcomes have the first coin Heads.
  - → So, now the probability that the first coin was Heads is equal to $2/3$.
  - → That is: The probability that the first coin was Heads, given that 2 coins were Heads, is equal to $2/3$.
  - → In symbols: P(first coin Heads | 2 coins were Heads) $= 2/3$.
- In general, if A and B are two events, then the conditional probability of A given B is written as $\mathrm{P}(A \,|\, B)$, and represents the fraction of the times when B occurs, in which A also occurs. [Diagram.] So, it is equal to:

$$\mathrm{P}(A \,|\, B) \;=\; \frac{\mathrm{P}(A \cap B)}{\mathrm{P}(B)}\,.$$

- Note: If $\mathrm{P}(B) = 0$, then $\mathrm{P}(A \,|\, B)$ is ... undefined! It only makes sense if $\mathrm{P}(B) > 0$.
  - → (Reasonable since if $\mathrm{P}(B) = 0$, then B will "never" happen.)
- In the above example, $A = \{\text{first coin Heads}\}$, and $B = \{2 \text{ coins Heads}\}$.

→ Then, $B = \{HHT, HTH, THH\}$, so $\mathrm{P}(B) = |B| \,/\, |S| = 3/8$.

→ Also, $A \cap B = \{HHT, HTH\}$, so $\mathrm{P}(A \cap B) = |A \cap B| \,/\, |S| = 2/8$.

→ Hence, $\mathrm{P}(A \,|\, B) = \mathrm{P}(A \cap B) \,/\, \mathrm{P}(B) = (2/8) \,/\, (3/8) = 2/3$, same as before.

• e.g. Roll three fair six-sided dice. What is P(first die is 3 | at least one 3)?

→ Here $S = \{111, 112, \ldots, 665, 666\}$. So, $|S| = 6 \cdot 6 \cdot 6 = 6^3 = 216$.

→ Here $A = \{\text{first die is 3}\}$, and $B = \{\text{at least one 3}\}$. What is $\mathrm{P}(B)$?

→ Well, $B^C = \{\text{no 3}\}$, i.e. each die in $\{1, 2, 4, 5, 6\}$. (So, 5 choices.)

→ So, $|B^C| = 5^3$, and $\mathrm{P}(B^C) = |B^C|/|S| = 5^3/6^3 = 125/216$.

→ Then, $\mathrm{P}(B) = 1 - \mathrm{P}(B^C) = 1 - 125/216 = 91/216$. What about $\mathrm{P}(A)$?

→ Well, $A = \{311, 312, \ldots, 366\}$, so $|A| = 6^2 = 36$, and $\mathrm{P}(A) = 36/216 = 1/6$. (Of course – "independence" – coming soon.) But what we really need is ...

→ $\mathrm{P}(A \cap B)$. But $A \subseteq B$, so $A \cap B = A$, so $\mathrm{P}(A \cap B) = \mathrm{P}(A) = 36/216 = 1/6$.

→ Hence, $\mathrm{P}(A \,|\, B) = \mathrm{P}(A \cap B)/\mathrm{P}(B) = (1/6)/(91/216) = (36/216)/(91/216) = 36/91 \doteq 0.396$. Much more than $1/6 \doteq 0.167$. Surprising?

• e.g. Roll three fair six-sided dice. What is P(at least one 3 | sum is ≤ 5)?

→ Here $S = \{111, 112, \ldots, 665, 666\}$. So, $|S| = 6 \cdot 6 \cdot 6 = 216$.

→ Here $A = \{\text{at least one 3}\}$, and $B = \{\text{sum is} \leq 5\}$. What is $|B|$?

→ Well, $B = \{111, 112, 113, 121, 122, 131, 211, 212, 221, 311\}$.

→ So, $|B| = 10$, and $\mathrm{P}(B) = |B| \,/\, |S| = 10/216$.

→ What about A? Well, $A = \{311, 312, 313, \ldots\}$. Tricky? Use A^C!

→ Here $|A^C| = 5^3 = 125$, so $\mathrm{P}(A^C) = 125/216 \doteq 0.579$, so $\mathrm{P}(A) \doteq 0.421$.

→ But wait, here we don't need to know A, we only need $A \cap B$!

→ By looking at B, we see that $A \cap B = \{113, 131, 311\}$.

→ So, $|A \cap B| = 3$, and $\mathrm{P}(A \cap B) = |A \cap B| \,/\, |S| = 3/216$.

→ Then $\mathrm{P}(A \,|\, B) = \mathrm{P}(A \cap B) \,/\, \mathrm{P}(B) = (3/216) \,/\, (10/216) = 3/10 = 30\%$.

• Conditional Multiplication Formula: Since $\mathrm{P}(A \,|\, B) = \mathrm{P}(A \cap B)/\mathrm{P}(B)$, therefore $\mathrm{P}(A \cap B) = \mathrm{P}(B)\,\mathrm{P}(A \,|\, B)$. Similarly, $\mathrm{P}(A \cap B) = \mathrm{P}(A)\,\mathrm{P}(B \,|\, A)$. Useful!

• e.g. Suppose we are dealt two cards, in order, from a standard deck.

→ What is P(both are Face Cards)? Can instead use conditional prob ...

→ Let $A = \{\text{first card is Face Card}\}$, and $B = \{\text{second card is Face Card}\}$.

→ Then $P(A) = 12/52$. What about $\mathrm{P}(B \,|\, A)$?

→ Well, once we know that the first card is a Face Card, then there are 11 Face Cards remaining, out of 51 total remaining cards. So, $\mathrm{P}(B \,|\, A) = 11/51$.

→ Then $\mathrm{P}(A \cap B) = \mathrm{P}(A)\,\mathrm{P}(B \,|\, A) = (12/52)\,(11/51)$. Same as before. Easier?

• Combining this Conditional Multiplication Formula with our previous Law of Total Probability gives a new version:

• Law of Total Probability – Conditioned Version: Suppose $A_1, A_2, \ldots$ are a sequence (finite or infinite) of events which form a partition of S, i.e. they are dis-

joint ($A_i \cap A_j = \emptyset$ for all $i \neq j$) and their union equals the entire sample space ($\bigcup_i A_i = S$), and let B be any event. Then $\mathrm{P}(B) = \sum_i \mathrm{P}(A_i)\ \mathrm{P}(B \,|\, A_i)$, or equivalently $\mathrm{P}(B) = \mathrm{P}(A_1)\ \mathrm{P}(B \,|\, A_1) + \mathrm{P}(A_2)\ \mathrm{P}(B \,|\, A_2) + \ldots$.

• e.g. Flip one fair coin. If Heads, roll one die; if Tails, roll two dice. What is P(get at least one 5)?

→ Here $B = \{\text{at least one 5}\}$, and $A_1 = \{\text{Heads}\}$, and $A_2 = \{\text{Tails}\}$.

→ Then A_1, A_2 form a partition. And $\mathrm{P}(A_1) = \mathrm{P}(A_2) = 1/2$. Need $\mathrm{P}(B \,|\, A_i)$.

→ Well, $\mathrm{P}(B \,|\, A_1) = \mathrm{P}(\text{get at least one 5 when you roll one die}) = 1/6$.

→ Also, $\mathrm{P}(B \,|\, A_2) = \mathrm{P}(\text{get at least one 5 when you roll two dice}) = ??$

→ Well, its complement is P(get no 5 when you roll two dice) $= 5^2/6^2 = 25/36$.

→ So, $\mathrm{P}(B \,|\, A_2) = 1 - (25/36) = 11/36$.

→ Then, from the above Law of Total Probability,

$$\mathrm{P}(B) = \sum_i \mathrm{P}(A_i)\ \mathrm{P}(B \,|\, A_i) = \mathrm{P}(A_1)\ \mathrm{P}(B \,|\, A_1) + \mathrm{P}(A_2)\ \mathrm{P}(B \,|\, A_2)$$

$$= (1/2)(1/6) + (1/2)(11/36) = 17/72 \doteq 0.236\,.$$

• Three-Card Challenge: Have three cards: C1=Blue-Blue, C2=Yellow-Yellow, C3=Blue-Yellow. Pick a card uniformly at random. Then pick one side of that card, uniformly at random. What is P(the card is C2 | the side is Yellow)?

→ Let $B = \{\text{the side is Yellow}\}$. First of all, what is $\mathrm{P}(B)$?

→ Use Law of Total Probability! Since we pick one of the three cards, the three cards C1,C2,C3 form a partition.

→ So, $\mathrm{P}(B) = \mathrm{P}(\text{C1})\,\mathrm{P}(B \,|\, \text{C1}) + \mathrm{P}(\text{C2})\,\mathrm{P}(B \,|\, \text{C2}) + \mathrm{P}(\text{C3})\,\mathrm{P}(B \,|\, \text{C3})$
$= (1/3)(0) + (1/3)(1) + (1/3)(1/2) = 1/3 + 1/6 = 1/2$. (Of course.)

→ Now, let $A = \{\text{the card is C2}\}$. Then what is $\mathrm{P}(A \cap B)$?

→ Well, $A \cap B = \{\text{choose C2, then Yellow}\} = \{\text{choose C2, then either side}\}$.

→ So, $\mathrm{P}(A \cap B) = \mathrm{P}(A)\,\mathrm{P}(B \,|\, A) = \mathrm{P}(\text{C2})\,\mathrm{P}(\text{Yellow Side} \,|\, \text{C2}) = (1/3)\,(1) = 1/3$.

→ Hence, P(the card is C2 | the side is Yellow) $= \mathrm{P}(A \,|\, B) = \mathrm{P}(A \cap B)/\mathrm{P}(B) = (1/3)/(1/2) = 2/3$. Surprising? (Try it!)

→ Intuition: We picked one of the three Yellow sides, of which two are on C2.

• e.g. Suppose a disease affects one person in a thousand, and a test for the disease has 99% accuracy. Someone is selected at random, and tested for the disease.

→ (a) What is P(they test positive)?

→ Use the Law of Total Probability! Here $B = \{\text{test positive}\}$. And, partition is $A_1 = \{\text{have disease}\}$ and $A_2 = \{\text{do not have disease}\}$.

→ So, $\mathrm{P}(B) = \mathrm{P}(A_1)\,\mathrm{P}(B \,|\, A_1) + \mathrm{P}(A_2)\,\mathrm{P}(B \,|\, A_2)$
$= (1/1000)(0.99) + (999/1000)(0.01) = 0.01098$.

→ (b) Given that they tested positive (i.e., conditional on them testing positive), what is the conditional probability that they have the disease?

→ This is $\mathrm{P}(A_1 \,|\, B) = \mathrm{P}(A_1 \cap B)/\mathrm{P}(B)$. But how do we compute $\mathrm{P}(A_1 \cap B)$?

→ Use the Conditional Multiplication Formula! Here $\mathrm{P}(A_1 \cap B) = \mathrm{P}(A_1)\,\mathrm{P}(B \mid A_1) = (1/1000)(0.99) = 0.00099$.

→ So, $\mathrm{P}(A_1 \mid B) = \mathrm{P}(A_1 \cap B)/\mathrm{P}(B) = (0.00099)/(0.01098) = 0.0901639 \doteq 9\% \doteq 1/11$. Small! Why?

→ Intuition: So many more people do not have the disease, that even their false positives (1%) are more than the number of people who have the disease (0.1%).

- In the above example, we knew $\mathrm{P}(B \mid A_1)$ (it was 99%), but we wanted $\mathrm{P}(A_1 \mid B)$.

→ What is the connection between them?

- In general, $\mathrm{P}(B \mid A) = \mathrm{P}(A \cap B) \,/\, \mathrm{P}(A)$, and $\mathrm{P}(A \mid B) = \mathrm{P}(A \cap B) \,/\, \mathrm{P}(B)$.

→ So ... $\mathrm{P}(A \mid B) = \frac{\mathrm{P}(A)}{\mathrm{P}(B)}\,\mathrm{P}(B \mid A)$. ("Bayes Theorem", or "Bayes Rule")

→ (Aside: This formula is the inspiration for "Bayesian Statistics" ...)

Suggested Homework: 1.5.1, 1.5.2, 1.5.3, 1.5.4, 1.5.6, 1.5.7, 1.5.8, 1.5.10, 1.5.11, 1.5.12, 1.5.13, 1.5.16, 1.5.17.

Independence

- Recall: If we roll three fair six-sided dice, then P(first die shows 5) = ... 1/6. Of course! Why? Because the first die doesn't "care" about the other dice!

→ And, P(first die shows 5 | second die shows 4) = 1/6, too. Doesn't care!

→ More formally, we say the first die is "independent" of the other dice.

- If A and B are any two events, then saying they are independent means that they do not affect each others' probabilities, i.e. that $\mathrm{P}(A \mid B) = \mathrm{P}(A)$, and $\mathrm{P}(B \mid A) = \mathrm{P}(B)$.

→ But $\mathrm{P}(A \mid B) = \mathrm{P}(A \cap B) \,/\, \mathrm{P}(B)$, so $\mathrm{P}(A \mid B) = \mathrm{P}(A)$ if and only if ... $\mathrm{P}(A \cap B) = \mathrm{P}(A)\,\mathrm{P}(B)$. This is the official definition of independence. (Better, since it is symmetric in A and B, and it is valid even if $\mathrm{P}(A) = 0$ or $\mathrm{P}(B) = 0$.)

→ If A and B are independent, and $\mathrm{P}(B) > 0$, then $\mathrm{P}(A \mid B) = \mathrm{P}(A)$.

END WEDNESDAY #2

- If two parts of an experiment are physically completely unrelated, like two different coins, or a coin and a die, or multiple dice, then they must be independent.

→ We already implicitly used this fact, e.g. if you flip two coins, then P(both Heads) = P(first is Heads) P(second is Heads) = (1/2)(1/2) = 1/4, and so on.

→ But now we know why it was okay to multiply!

- e.g. Flip two fair coins. So, $S = \{HH, HT, TH, TT\}$, $|S| = 4$, all equally likely.

→ Let $A = \{$first coin Heads$\}$, $B = \{$second coin Heads$\}$, and $C = \{$both coins are the same$\}$.

→ Are A and B independent? Yes, of course! (physically unrelated)

→ Check: $\mathrm{P}(A) = |\{HH, HT\}| \,/\, 4 = 2/4 = 1/2$, and $\mathrm{P}(B) = |\{HH, TH\}| \,/\, 4 = 2/4 = 1/2$, and $\mathrm{P}(A \cap B) = |\{HH\}| \,/\, 4 = 1/4 = (1/2)(1/2) = \mathrm{P}(A)\,\mathrm{P}(B)$.

→ What about A and C? Well, $\mathrm{P}(C) = |\{HH, TT\}| \,/\, 4 = 2/4 = 1/2$, and $\mathrm{P}(A \cap C) = |\{HH\}| \,/\, 4 = 1/4$, So, $\mathrm{P}(A \cap C) = 1/4 = (1/2)(1/2) = \mathrm{P}(A)\,\mathrm{P}(C)$.

→ So, A and C are independent! And similarly, B and C are independent.

→ So, A and B and C are all pairwise independent.

→ Hence, $\mathrm{P}(A \,|\, C) = \mathrm{P}(A) = 1/2$, and $\mathrm{P}(C \,|\, A) = \mathrm{P}(C) = 1/2$, etc. Surprising?

→ But are they all truly independent? Well, suppose we know A and also know B. Then we would know that C is true, too!

→ That is, $\mathrm{P}(C \,|\, A \cap B) = 1 \neq 1/2 = \mathrm{P}(C)$.

→ Why? Since $\mathrm{P}(A \cap B \cap C) = |\{HH\}| \,/\, 4 = 1/4 \neq (1/2)(1/2)(1/2)$.

→ For A and B and C to be truly independent, we also need $\mathrm{P}(A \cap B \cap C) = \mathrm{P}(A)\,\mathrm{P}(B)\,\mathrm{P}(C)$. That would guarantee that e.g. $\mathrm{P}(C \,|\, A \cap B) = \mathrm{P}(C)$, etc.

• In general, a collection $A_1, A_2, A_3, \ldots$ of events are called independent if $\mathrm{P}(A_{i_1} \cap A_{i_2} \cap \ldots \cap A_{i_k}) = \mathrm{P}(A_{i_1})\ \mathrm{P}(A_{i_2})\ \ldots\ \mathrm{P}(A_{i_k})$ for any subcollection of the events.

→ If truly independent, then we can always multiply the probabilities.

→ e.g. Flip 5 fair coins: P(all Heads) $= (1/2)(1/2)(1/2)(1/2)(1/2) = 1/32$.

• Does it matter? Ask Sally Clark! Solicitor in Cheshire, England. Had two sons; each suffocated and died in infancy.

→ Sudden Infant Death Syndrome (SIDS)? Or murder!?!

→ 1999 testimony by paediatrician Sir Roy Meadow: "the odds against two [SIDS] in the same family are 73 million to one".

→ Sally Clark was arrested, jailed, and vilified, and her third son was temporarily taken away. Was this justified?

→ How did Meadow compute that "73 million to one"?

→ He said the probability of one child dying of SIDS was one in 8,543, so for two children dying, we multiply:
$(1/8,543) \times (1/8,543) = 1/72,982,849 \approx 1/73,000,000$. Was this valid?

→ No! We can't just multiply, since SIDS tends to run in families, i.e. not independent. Given one SIDS death, a second one is about 10 times more likely!

→ (Also, even the figure "one in 8,543" was misleading, since he included factors which lower the SIDS probability, but neglected other factors which raise it.)

→ (Separate point: Even if two SIDS deaths are quite unlikely, two murders are also unlikely! So, how to compare and evaluate? Even unlikely things will happen sometime to someone. Statistical inference! Interesting, but not part of this course.)

→ So what happened? Convicted! Jailed for three years! Then overturned.

→ More info in my article: probability.ca/justice

Suggested Homework: 1.5.9, 1.5.14, 1.5.15, 1.5.20.

Continuity of Probabilities

• Recall: For a function $f : \mathbf{R} \to \mathbf{R}$, "continuity" means if $\lim_{n\to\infty} x_n = x$, then $\lim_{n\to\infty} f(x_n) = f(x)$. Is there something similar for probabilities $\mathrm{P}(A_n)$? Sort of ...

• e.g. $S = \mathbf{N} := \{1, 2, 3, \ldots\}$, with $\mathrm{P}(i) = 2^{-i}$ for each $i \in S$.

• Question: In this example, what is $\lim_{n\to\infty} \mathrm{P}(\leq n)$, i.e. $\lim_{n\to\infty} \mathrm{P}\{1, 2, \ldots, n\}$?

→ It must be 1, of course. Not just in this example, but in general:

• Definition: Write that $\{A_n\} \nearrow A$ if $\bigcup_n A_n = A$, and they are "nested increasing", i.e. $A_n \subseteq A_{n+1}$ for all n, i.e. $A_1 \subseteq A_2 \subseteq A_3 \subseteq \ldots$. Like $\lim_{n\to\infty} A_n = A$. Diagram:

→ e.g. if $A_n = \{1, 2, \ldots, n\}$, then $\{A_n\} \nearrow \mathbf{N}$. [Check!] And therefore?

• Continuity Of Probabilities Theorem: If $\{A_n\} \nearrow A$, then $\lim_{n\to\infty} \mathrm{P}(A_n) = \mathrm{P}(A)$.

→ Proof: Let $B_1 = A_1$, and $B_n = A_n \cap A_{n-1}^C$ for $n \geq 2$.

→ Then A is the disjoint union of all of the B_n. [Diagram.]

→ Hence, by additivity, $\mathrm{P}(A) = \sum_{i=1}^{\infty} \mathrm{P}(B_i) := \lim_{n\to\infty} \sum_{i=1}^{n} \mathrm{P}(B_i)$.

→ But also, A_n is the disjoint union of just $B_1, B_2, \ldots, B_n$.

→ So, by additivity, $\mathrm{P}(A_n) = \sum_{i=1}^{n} \mathrm{P}(B_i)$.

→ Combining these two, $\mathrm{P}(A) = \lim_{n\to\infty} \sum_{i=1}^{n} \mathrm{P}(B_i) = \lim_{n\to\infty} \mathrm{P}(A_n)$. ∎

• Similarly, write that $\{A_n\} \searrow A$ if $\bigcap_n A_n = A$, and they are nested decreasing, i.e. $A_n \supseteq A_{n+1}$ for all n, i.e. $A_1 \supseteq A_2 \supseteq A_3 \supseteq \ldots$. Diagram:

→ It follows that $\{A_n\} \searrow A$ if and only if $\{A_n^C\} \nearrow A^C$. [Exercise!]

→ Hence, if $\{A_n\} \searrow A$, then $\{A_n^C\} \nearrow A^C$, so $\lim_{n\to\infty} \mathrm{P}(A_n^C) = \mathrm{P}(A^C)$, i.e. $\lim_{n\to\infty}[1 - \mathrm{P}(A_n)] = 1 - \mathrm{P}(A)$, so $\lim_{n\to\infty} \mathrm{P}(A_n) = \mathrm{P}(A)$, just like before.

• e.g. Suppose we have any probablities P defined on $S = \mathbf{N} = \{1, 2, 3, \ldots\}$.

→ Does there necessarily exist some finite number $n \in \mathbf{N}$ with $\mathrm{P}\{1, 2, \ldots, n\} = 1$?

→ No! e.g. in above example with $\mathrm{P}(i) = 2^{-i}$, we have $\mathrm{P}\{1, 2, \ldots, n\} = \sum_{i=1}^{n} 2^{-i} = \frac{2^{-1} - 2^{-n-1}}{1 - 2^{-1}} = 1 - 2^{-n}$, which is always < 1.

→ Is it necessarily true that $\lim_{n\to\infty} \mathrm{P}\{1, 2, \ldots, n\} = 1$?

→ Yes! Since $\{1, 2, \ldots, n\} \nearrow \mathbf{N} = S$, by Continuity Of Probabilities, we must have $\lim_{n\to\infty} \mathrm{P}\{1, 2, \ldots, n\} = \mathrm{P}(S) = 1$.

→ Does there necessarily exist some finite $n \in \mathbf{N}$ with $\mathrm{P}\{1, 2, \ldots, n\} > 0.99$?

→ Yes! Since $\lim_{n\to\infty} \mathrm{P}\{1, 2, \ldots, n\} = 1$, therefore $\mathrm{P}\{1, 2, \ldots, n\} > 0.99$ for all sufficiently large n.

• Suppose we flip an infinite number of fair coins. (!)

→ What is P(all the coins are Heads)? How to even think about that?

→ Let $A = \{$all the coins are Heads$\}$, and $A_n = \{$the first n coins are Heads$\}$.

→ Then $A_n \supseteq A_{n+1}$. Also $\bigcap_{n=1}^{\infty} A_n = A$. So, $\{A_n\} \searrow A$.

→ Hence, $\mathrm{P}(\text{all coins Heads}) = \lim_{n\to\infty} \mathrm{P}(A_n) = \lim_{n\to\infty}(1/2)^n = 0$.

→ (So, {all coins Heads} is "possible", but has probability 0; will never happen.)

END MONDAY #3

- e.g. Suppose we pick a number between 0 and 1. Diagram:

→ Suppose $\mathrm{P}\big([a,b]\big) = b - a$ whenever $0 \le a < b \le 1$, e.g. $\mathrm{P}\big([\frac{1}{2}, \frac{2}{3}]\big) = \frac{2}{3} - \frac{1}{2} = \frac{1}{6}$.

→ What about the open interval $\mathrm{P}\big((\frac{1}{2}, \frac{2}{3})\big)$? Is it necessarily the same?

→ Use Continuity Of Probabilities!

→ Let $A = (\frac{1}{2}, \frac{2}{3})$, and $A_n = \big[\frac{1}{2} + \frac{1}{n},\ \frac{2}{3} - \frac{1}{n}\big]$. Diagram:

→ Then $A_{n+1} \supseteq A_n$, and $\bigcup_{n=1}^{\infty} A_n = A$, so $\{A_n\} \nearrow A$.

→ Also, we know that $\mathrm{P}\big(\big[\frac{1}{2} + \frac{1}{n},\ \frac{2}{3} - \frac{1}{n}\big]\big) = [\frac{2}{3} - \frac{1}{n}] - [\frac{1}{2} + \frac{1}{n}] = \frac{1}{6} - \frac{2}{n}$.

→ Hence, by Continuity Of Probabilities, $\mathrm{P}(A) = \lim_{n\to\infty} \mathrm{P}(A_n)$,
i.e. $\mathrm{P}\big((\frac{1}{2}, \frac{2}{3})\big) = \lim_{n\to\infty}[\frac{1}{6} - \frac{2}{n}] = \frac{1}{6}$.

→ Similarly, must have $\mathrm{P}\big((a,b)\big) = b - a$ whenever $0 \le a < b \le 1$.

→ What about $\mathrm{P}(\{a\})$, for $a \in \mathbf{R}$? Zero? Again by Continuity of Probabilities,
$\mathrm{P}(\{a\}) = \lim_{n\to\infty} \mathrm{P}([a - \frac{1}{n}, a + \frac{1}{n}]) = \lim_{n\to\infty}((a + \frac{1}{n}) - (a - \frac{1}{n})) = \lim_{n\to\infty} \frac{2}{n} = 0$.

Suggested Homework: 1.6.1, 1.6.2, 1.6.3, 1.6.4, 1.6.5, 1.6.6, 1.6.7, 1.6.8, 1.6.9, 1.6.10. Optional: 1.6.11.

[END OF TEXTBOOK CHAPTER #1]

Random Variables

- A random variable is "any" function from S to $\mathbf{R}$.

→ Intuitively, it represents some random quantity in an experiment.

- e.g. Roll 3 dice: X = number showing on the first die.

→ X could be 1,2,3,4,5,6, depending on result: $X(265) = 2$, $X(513) = 5$, etc.

→ Or, Y = sum of the three numbers showing, so $Y(265) = 13$, $Y(513) = 9$, etc.

→ Or, Z = first number divided by third number: $Z(265) = 2/5$, $Z(513) = 5/3$.

• Or: Roll three fair dice, $X(s)$ = number of 5's, $Y(s)$ = number of 3's, $Z = X - Y$.

→ Then $X(335) = 1$, $Y(335) = 2$, $Z(335) = -1$, etc. Values can be negative, too!

• e.g. Flip 10 coins: X = # of Heads, or $Y = (\#\text{ of Heads})^2$, or $Z = 1$ if first coin Heads otherwise $Z = 0$, etc.

→ So $X(HHHTTTHTTT) = 4$, $X(TTHHHHHHHT) = 7$, etc.

→ In this example, can also write $Y = X^2$ (function of another random variable).

• e.g. $X(s) = 5$ for all $s \in S$: "constant random variable". (Or any constant.)

• Special case: $I_A(s) = 1$ if $s \in A$ otherwise $I_A(s) = 0$. "indicator function"

• e.g. $S = \mathbf{N} := \{1, 2, 3, \ldots\}$, with $\mathrm{P}(i) = 2^{-i}$ for each $i \in S$.

→ Maybe $X(s) = s$, and $Y(s) = s^2$. What are their largest possible values?

→ None! They can be arbitrarily large. "unbounded random variables"

→ Also, for all $s \in S$ we have $s \le s^2$, i.e. $X(s) \le Y(s)$ for all $s \in S$, so "$X \le Y$".

Suggested Homework: 2.1.1, 2.1.2, 2.1.4, 2.1.5, 2.1.6, 2.1.10, 2.1.11, 2.1.12, 2.1.15.

Distributions of Random Variables

• The distribution of a random variable is the collection of all of the probabilities of the variable being in every possible subset of $\mathbf{R}$.

• e.g. tonight's dinner, with $S = \{\text{Beef, Chicken, Fish}\}$, and P(Beef)=0.40, P(Chicken)=0.15, and P(Fish)=0.45.

→ Let $X(\text{Beef}) = 1$, $X(\text{Chicken}) = 2$, $X(\text{Fish}) = 5$. Probabilities for X?

→ Here $\mathrm{P}(X = 1) = \mathrm{P}\{\text{Beef}\} = 0.40$, and $\mathrm{P}(X = 2) = \mathrm{P}\{\text{Chicken}\} = 0.15$, and $\mathrm{P}(X = 5) = \mathrm{P}\{\text{Fish}\} = 0.45$. What about $\mathrm{P}(X \le 3)$?

→ Well, $\mathrm{P}(X \le 3) = \mathrm{P}\{\text{Beef, Chicken}\} = 0.40 + 0.15 = 0.55$. And $\mathrm{P}(X = 7) = 0$.

→ And $\mathrm{P}(X < 20) = \mathrm{P}\{\text{Beef, Chicken, Fish}\} = 0.40 + 0.15 + 0.45 = 1$.

→ And $\mathrm{P}(1 < X < 6) = \mathrm{P}\{\text{Chicken, Fish}\} = 0.15 + 0.45 = 0.60$. And so on.

→ Can also write that for "any" subset $B \subseteq \mathbf{R}$, we have $\mathrm{P}(X \in B) = 0.40\, I_B(1) + 0.15\, I_B(2) + 0.45\, I_B(5)$.

→ e.g. If B is the event "≤ 3", then $I_B(1) = 1$, $I_B(2) = 1$, and $I_B(5) = 0$, so $\mathrm{P}(X \in B) = 0.40\,(1) + 0.15\,(1) + 0.45\,(0) = 0.55$, like before.

• In general, "$\mathrm{P}(X \in B)$" means $\mathrm{P}(X^{-1}(B)) := \mathrm{P}\{s \in S : X(s) \in B\}$.

→ e.g. If B is the event "≤ 3", then $B = \{x \in \mathbf{R} : x \le 3\}$, so $\mathrm{P}(X \in B) = \mathrm{P}(X \le 3) = \mathrm{P}\big(X \in (-\infty, 3]\big) = \mathrm{P}\big(X^{-1}(-\infty, 3]\big)$.

Suggested Homework: 2.2.1, 2.2.2, 2.2.3, 2.2.4, 2.2.5, 2.2.6, 2.2.8, 2.2.9, 2.2.10.

Discrete Random Variables

• A random variable is called discrete if $\sum_{x \in \mathbf{R}} \mathrm{P}(X = x) = 1$.

→ i.e., all of its probability is on individual values.

→ Not always true! e.g. if we "pick a number uniformly between 0 and 1", then we know that $\mathrm{P}(X = x) = 0$ for all values of x, so $\sum_{x \in \mathbf{R}} \mathrm{P}(X = x) = 0 < 1$.

• If it's true, there's a distinct sequence $x_1, x_2, x_3, \ldots \in \mathbf{R}$, and corresponding probabilities $p_1, p_2, p_3, \ldots \geq 0$, with $\sum_i p_i = 1$, such that $\mathrm{P}(X = x_i) = p_i$ for each i.

→ In above example, $x_1 = 1$, $x_2 = 2$, $x_3 = 5$, with $p_1 = 0.40$, $p_2 = 0.15$, $p_3 = 0.45$.

• Can also define the "probability function" as $p_X(x) := \mathrm{P}(X = x)$.

→ So, $p_X(x_i) = p_i$ for all i, with $p_X(x) = 0$ for all $x \notin \{x_1, x_2, \ldots\}$.

→ In above example, $p_X(1)$=0.40, $p_X(2)$=0.15, $p_X(3)$=0.45, otherwise $p_X(x)$=0.

• e.g. Flip one fair coin, and let $X = \#$ Heads.

→ Then $\mathrm{P}(X = 0) = 1/2$, and $\mathrm{P}(X = 1) = 1/2$.

→ So, here $x_1 = 0$, and $x_2 = 1$, and $p_1 = p_2 = 1/2$.

→ Also, $p_X(0) = 1/2$ and $p_X(1) = 1/2$, with $p_X(x) = 0$ for all $x \neq 0, 1$.

• e.g. Flip <u>two</u> fair coins, and let $X = \#$ Heads. Then $\mathrm{P}(X = 0) = \binom{2}{0}/2^2 = 1/4$, and $\mathrm{P}(X = 1) = \binom{2}{1}/2^2 = 2/4 = 1/2$, and $\mathrm{P}(X = 2) = \binom{2}{2}/2^2 = 1/4$.

→ So $x_1 = 0$, and $x_2 = 1$, and $x_3 = 2$, and $p_1 = 1/4$, and $p_2 = 1/2$, and $p_3 = 1/4$.

→ Also, $p_X(0) = 1/4$ and $p_X(1) = 1/2$ and $p_X(2) = 1/4$, otherwise $p_X(x) = 0$.

Suggested Homework: 2.3.1, 2.3.2, 2.3.3, 2.3.4, 2.3.5.

Some Important Discrete Distributions

• e.g. Shoot one "free throw" in basketball, with probability "θ" of scoring (for some value of θ with $0 < \theta < 1$, e.g. $\theta = 0.5$, or $\theta = 1/3$, or $\ldots$).

→ Let $X = 1$ if you score, or $X = 0$ if you miss. Probabilities for X?

→ Here $\mathrm{P}(X = 1) = \mathrm{P}\{\text{score}\} = \theta$, and $\mathrm{P}(X = 0) = \mathrm{P}\{\text{miss}\} = 1 - \theta$.

→ This is the "Bernoulli(θ) distribution".

→ Can also write $X \sim \text{Bernoulli}(\theta)$.

→ Then $p_X(0) = 1 - \theta$ and $p_X(1) = \theta$, with $p_X(x) = 0$ for all $x \neq 0, 1$.

→ e.g. Bernoulli(0.5), or Bernoulli(1/3), or $\ldots$

→ (Of course, it doesn't have to be free throws! This distribution applies to <u>any</u> situation involving any "attempt" or "trial" having probability θ of "success" and probability $1 - \theta$ of "failure". And similarly for the below, too.)

• e.g. Shoot <u>2</u> free throws, each <u>independent</u> with probability θ of scoring (for some value of θ with $0 < \theta < 1$ like 0.5 or 1/3).

→ Let $X = \#$ Successes. Probabilities for X?

→ Here $\mathrm{P}(X = 0) = \mathrm{P}\{\text{miss-miss}\} = (1 - \theta)(1 - \theta) = (1 - \theta)^2$.
(We can <u>multiply</u> because they are <u>independent</u>.)

→ And, $\mathrm{P}(X=2) = \mathrm{P}\{\text{score-score}\} = (\theta)(\theta) = \theta^2$.

→ And, $\mathrm{P}(X=1) = \mathrm{P}\{\text{score-miss, miss-score}\} = (\theta)(1-\theta)+(1-\theta)(\theta) = 2\theta(1-\theta)$.

→ This is the "Binomial$(2,\theta)$ distribution".

→ Then $p_X(0) = (1-\theta)^2$, $p_X(1) = 2\theta(1-\theta)$, $p_X(2) = \theta^2$, otherwise $p_X(x) = 0$.

• e.g. Shoot "n" free throws, each independent with probability θ of scoring (for some value of θ with $0 < \theta < 1$, and some value of $n \in \mathbf{N}$ like 2 or 10 or 286).

→ Let $X = \#$ Successes. Probabilities for X?

→ Here $\mathrm{P}(X=0) = \mathrm{P}\{\text{miss-miss-...-miss}\} = (1-\theta)^n$.

→ And, $\mathrm{P}(X=n) = \mathrm{P}\{\text{score-score-...-score}\} = \theta^n$.

→ And, $\mathrm{P}(X=1) = \mathrm{P}\{\text{score-miss-...-miss, miss-score-miss-...-miss, ...}\} = ??$

→ Well, each such sequence has probability $\theta(1-\theta)\dots(1-\theta) = \theta(1-\theta)^{n-1}$.

→ And, there are n such sequences (one for each shot which could score).

→ So, $\mathrm{P}(X=1) = n\theta(1-\theta)^{n-1}$.

→ What about $\mathrm{P}(X=k)$ for any integer $k \in \{0,1,2,\dots,n\}$?

→ Well, $\mathrm{P}(X=k) = \mathrm{P}\{\text{all sequences of } k \text{ scores and } n-k \text{ misses}\}$.

→ Each such sequence has probability $\theta^k(1-\theta)^{n-k}$.

→ And, the number of such sequences is $\binom{n}{k}$. ("Choose" which k shots scored.)

→ So, $p_X(k) := \mathrm{P}(X=k) = \binom{n}{k}\theta^k(1-\theta)^{n-k}$, for any $k \in \{0,1,2,\dots,n\}$.

→ This is the "Binomial(n,θ) distribution". Can write $X \sim \text{Binomial}(n,\theta)$.

→ Check: $k=0$: $\mathrm{P}(X=0) = \binom{n}{0}\theta^0(1-\theta)^{n-0} = (1-\theta)^n$. Yep!

→ Check: $k=n$: $\mathrm{P}(X=n) = \binom{n}{n}\theta^n(1-\theta)^{n-n} = \theta^n$. Yep!

→ Check: $k=1$: $\mathrm{P}(X=1) = \binom{n}{1}\theta^1(1-\theta)^{n-1} = n\theta(1-\theta)^{n-1}$. Yep!

→ Check: $\mathrm{P}(X=k) \geq 0$. Yep!

→ Check: $\sum_{k=0}^{n} \mathrm{P}(X=k) = \sum_{k=0}^{n} \binom{n}{k}\theta^k(1-\theta)^{n-k} = ??$
$= [\theta + (1-\theta)]^n = 1^n = 1$ (by using the "Binomial Theorem"). Yep!

• Special case: Binomial$(1,\theta)$ is the same as Bernoulli(θ).

• Suppose $X_1, X_2, \dots, X_n \sim \text{Bernoulli}(\theta)$, for independent trials.

→ Let $Y = X_1 + X_2 + \dots + X_n$. What is the distribution of Y?

→ Here Y represents the number of successes in n independent attempts, each with probability θ of success, so $Y \sim \text{Binomial}(n,\theta)$.

→ Special case: if $\theta = 1/2$, then the Binomial$(n,1/2)$ distribution has $\mathrm{P}(X=k) = \binom{n}{k}(1/2)^k(1-(1/2))^{n-k} = \binom{n}{k}(1/2)^n = \binom{n}{k} \,/\, 2^n$, same as before.

• e.g. Suppose 1/4 of students have long hair. You pick four students at random, with replacement. What is P(exactly 2 of them have long hair)?

→ Let $Y = \#$ students with long hair. Then $Y \sim \text{Binomial}(4,\ 1/4)$. So, $\mathrm{P}(Y=2) = \binom{4}{2}(1/4)^2\Big(1-(1/4)\Big)^{4-2} = 6(1/4)^2(3/4)^2 = 54/256 = 27/128 \doteq 0.21$.

• e.g. Repeatedly shoot free throws, each independent with probability θ of scoring. Let $Z = \#$ misses before the first score. Probabilities for Z?

→ Here $\mathrm{P}(Z = 0) = \mathrm{P}(\text{score first time}) = \theta$.

→ And, $\mathrm{P}(Z = 1) = \mathrm{P}(\text{miss-score}) = (1-\theta)\theta$.

→ And, $\mathrm{P}(Z = 2) = \mathrm{P}(\text{miss-miss-score}) = (1-\theta)^2\theta$.

→ In general, $\mathrm{P}(Z = k) = \mathrm{P}(\text{miss-miss-...-miss-score}) = (1-\theta)^k\theta$, valid for all $k = 0, 1, 2, 3, \ldots$.

→ This is the "Geometric(θ) distribution". Can write $Z \sim \text{Geometric}(\theta)$.

→ Check: $\mathrm{P}(Z = k) \geq 0$ for all k. Yep!

→ Check: $\sum_{k=0}^{\infty}(1-\theta)^k\theta = \theta[1 + (1-\theta) + (1-\theta)^2 + (1-\theta)^3 + \ldots]$
$= \theta[\frac{1}{1-(1-\theta)}] = \theta[\frac{1}{\theta}] = 1$. (Geometric series.) Yep!

• [Some books count # attempts up to and including first success: one more.]

• e.g. Suppose 1/4 of students have long hair. You repeatedly pick students at random, with replacement. What is P(the sixth student is the first with long hair)?

→ Let $X = \#$ students before first one with long hair. Then we want to find $\mathrm{P}(X = 5)$. And, here $X \sim \text{Geometric}(1/4)$.

→ So, $\mathrm{P}(X = 5) = (1/4)\big(1 - (1/4)\big)^5 = (1/4)(3/4)^5 = 243/4096 \doteq 0.059$.

• Suppose again that $X \sim \text{Geometric}(1/4)$. What is $\mathrm{P}(X = \infty)$?

→ Well, $\mathrm{P}(X \leq m) = \sum_{k=0}^{m}\mathrm{P}(X = k) = \sum_{k=0}^{m}(1/4)(3/4)^k = (1/4)[1 + (3/4) + (3/4)^2 + \ldots + (3/4)^k] = (1/4)\frac{1-(3/4)^{m+1}}{1-(3/4)} = 1 - (3/4)^{m+1}$. This is < 1.

→ So, $\mathrm{P}(X > m) = 1 - \mathrm{P}(X \leq m) = 1 - [1 - (3/4)^{m+1}] = (3/4)^{m+1}$.

→ Hence, $\mathrm{P}(X > m) > 0$ for any $m \in \mathbf{N}$. ("unbounded random variable")

→ But also, $\{X > m\} \searrow \{X = \infty\}$. [check!]

→ Hence, by Continuity of Probabilities,
$\mathrm{P}(X = \infty) = \lim_{m\to\infty}\mathrm{P}(X > m) = \lim_{m\to\infty}(3/4)^{m+1} = 0$. Phew!

Suggested Homework: 2.3.6, 2.3.7, 2.3.10, 2.3.11, 2.3.14, 2.3.15, 2.3.16(a,b), 2.3.23, 2.3.24, 2.3.27.

END WEDNESDAY #3

• Now, if X is a discrete variable which always equals one of the values $x_1, x_2, \ldots$, then the events $\{X = x_i\}$ form a partition. So, we get that ...

• [Law of Total Probability – Discrete Random Variable Version]
If X is a discrete random variable, with possible values $x_1, x_2, \ldots$, and corresponding probabilities $p_1, p_2, \ldots$, and B is any event, then
$\mathrm{P}(B) = \sum_i \mathrm{P}(X = x_i)\ \mathrm{P}(B \,|\, X = x_i) = \sum_i p_i\ \mathrm{P}(B \,|\, X = x_i)$.

→ In fact, since $\mathrm{P}(X = x) = 0$ for all other x, we can also write this as:
$\mathrm{P}(B) = \sum_{x\in\mathbf{R}}\mathrm{P}(X = x)\ \mathrm{P}(B \,|\, X = x)$.

• e.g. Suppose we roll one fair six-sided die, and then flip a number of coins equal to the number showing on the die. Let $X = \#$ Heads. Compute $\mathrm{P}(X = 3)$.

→ Let Y = number on die. Then Y is discrete, with possible values $\{1, 2, 3, 4, 5, 6\}$.

→ Use the values of Y as a partition! Then ...

$\mathrm{P}(X = 3) = \sum_{y \in \mathbf{R}} \mathrm{P}(Y = y)\ \mathrm{P}(X = 3 \,|\, Y = y) = \sum_{y=1}^{6} \mathrm{P}(Y = y)\ \mathrm{P}(X = 3 \,|\, Y = y) = \sum_{y=3}^{6} (1/6)\ [\binom{y}{3} \ / \ 2^y] = \frac{1}{6}\left(\frac{1}{8} + \frac{4}{16} + \frac{10}{32} + \frac{20}{64}\right) = 1/6.$ (Just like before.)

→ And, $\mathrm{P}(X = 4) = \sum_{y \in \mathbf{R}} \mathrm{P}(Y = y)\ \mathrm{P}(X = 4 \,|\, Y = y) = \sum_{y=1}^{6} \mathrm{P}(Y = y)\ \mathrm{P}(X = 4 \,|\, Y = y) = \sum_{y=4}^{6} (1/6)\ [\binom{y}{4} \ / \ 2^y] = \frac{1}{6}\left(\frac{1}{16} + \frac{5}{32} + \frac{15}{64}\right) = 29/384 \doteq 0.0755.$

• e.g. Suppose we roll one fair six-sided die, and then attempt a number of free throws equal to the number showing on the die. Assume we have independent probability 1/3 of scoring on each free throw. Let X = # Scores. Compute $\mathrm{P}(X = 3)$.

→ Let Y = number on die. Then by the Law of Total Probability,

$\mathrm{P}(X = 3) = \sum_{y \in \mathbf{R}} \mathrm{P}(Y = y)\ \mathrm{P}(X = 3 \,|\, Y = y) = \sum_{y=1}^{6} \mathrm{P}(Y = y)\ \mathrm{P}(X = 3 \,|\, Y = y) = \sum_{y=3}^{6} (1/6)\ \left[\binom{y}{3}(1/3)^3(2/3)^{y-3}\right] = (1/6)\Big[(1)(1/3)^3(2/3)^0 + (4)(1/3)^3(2/3)^1 + (10)(1/3)^3(2/3)^2 + (20)(1/3)^3(2/3)^3\Big] = \ldots = (1/6)\,[379/729] \doteq 0.087.$

Poisson Distribution

• e.g. Suppose Toronto has an average of $\lambda = 5$ house fires per day.

→ Intuitively, this is caused by a very large number n of buildings, each of which has a very small probability θ of having a fire.

→ Let $\lambda = n\theta$, i.e. $\theta = \lambda/n$. (Then λ is the "average" number of fires – later.)

→ Then the number of fires has the distribution Binomial$(n, \lambda/n)$, so

$$\begin{aligned} \mathrm{P}(\#\text{fires} = k) &= \binom{n}{k}\theta^k(1-\theta)^{n-k} \\ &= \frac{n(n-1)(n-2)\ldots(n-k+1)}{k!}\,(\lambda/n)^k\,[1-(\lambda/n)]^{n-k}. \end{aligned}$$

→ Now, what happens as $n \to \infty$, for a fixed value $k \in \{0, 1, 2, \ldots\}$?

→ Well, since $k \ll n$, we have $\frac{n}{n} = 1$, $\frac{n-1}{n} \to 1$, $\frac{n-2}{n} \to 1$, $\ldots$ $\frac{n-k+1}{n} \to 1$.

→ Hence, $\frac{n(n-1)(n-2)\ldots(n-k+1)}{n^k} \to 1$.

→ Also, from calculus, $e^x = 1 + x + \frac{x^2}{2!} + \ldots$, so for small $x \in \mathbf{R}$, $e^x \approx 1 + x$.

→ So, $[1 - (\lambda/n)]^{n-k} \approx [1 - (\lambda/n)]^n \approx [e^{-\lambda/n}]^n = e^{-\lambda}$.

→ Hence, as $n \to \infty$, we have $\mathrm{P}(\#\text{fires} = k) \ \to \ \frac{1}{k!}\lambda^k e^{-\lambda} \ = \ e^{-\lambda}\frac{\lambda^k}{k!}$.

→ This is the Poisson(λ) distribution: $\mathrm{P}(k) = e^{-\lambda}\,\frac{\lambda^k}{k!}$, for $k = 0, 1, 2, 3, \ldots$.

• Check: $\sum_{k=0}^{\infty} e^{-\lambda}\frac{\lambda^k}{k!} = e^{-\lambda}\,[1 + \lambda + \frac{\lambda^2}{2!} + \frac{\lambda^3}{3!} + \ldots] = e^{-\lambda}\,[e^{\lambda}] = 1$. Yep!

• In general, if n is very large, and θ is very small, then Binomial(n, θ) is well approximated by Poisson(λ) where $\lambda = n\theta$. "Poisson approximation"

• e.g. Suppose $Y \sim$ Poisson(3). What is $\mathrm{P}(Y = 4)$?

→ Well, $\mathrm{P}(Y = 4) = e^{-\lambda}\ \frac{\lambda^k}{k!} = e^{-3}\ \frac{3^4}{4!} = e^{-3}\ \frac{81}{24} \doteq 0.168$.

• e.g. Suppose $Y \sim$ Binomial$(20000, 0.0001)$.

→ Then $\mathrm{P}(Y=4) = \binom{20000}{4}(0.0001)^4(0.9999)^{20000-4} \doteq 0.09022352216$.

→ Poisson Approximation: Here $\lambda = n\theta = 20000 \cdot 0.0001 = 2$.

→ So, $\mathrm{P}(Y=4) \approx e^{-2}\frac{(2)^4}{4!} \doteq 0.09022352178$.

• Or, if $Y \sim \text{Binomial}(200, 0.01)$, then still $\lambda = 200 \cdot 0.01 = 2$, so Poisson approximation is the same, but now $\mathrm{P}(Y=4) = \binom{200}{4}(0.01)^4(0.99)^{200-4} \doteq 0.0902197$.

→ Still pretty close, but not as close.

Suggested Homework: 2.3.8, 2.3.12, 2.3.19, 2.3.27. Optional: 2.3.18, 2.3.30.

Understanding Distributions Using the Computer Software "R"

• Recall – basic info and links at: http://probability.ca/Rinfo.html

→ Also discussed in Appendix B of the textbook.

• Can use "R" to simulate from probability distributions!

→ e.g. "rbinom(1,10,1/2)", "rgeom(1,0.2)", "rpois(1,5)".

• Can also plot probabilities, e.g. "plot(dbinom(0:10,10,1/2))", "plot(dgeom(0:10,0.2))"

→ [Also: other parameter values, and different options like "type='b'", etc.]

Some Additional Discrete Distributions

• e.g. Repeatedly attempt free throws, with independent probability θ each time. Let $r \in \mathbf{N}$, and Y be the number of misses before the r^{th} score. What is $\mathrm{P}(Y=k)$?

→ Well, if $Y=k$, then the first $r-1+k$ shots must have included $r-1$ scores and k misses. Binomial Distribution! This probability is $\binom{r-1+k}{r-1}\,\theta^{r-1}\,(1-\theta)^k$.

→ Then we had to score on the final attempt, which has probability θ.

→ So, $\mathrm{P}(Y=k) = \binom{r-1+k}{r-1}\,\theta^{r-1}\,(1-\theta)^k\,\theta = \binom{r-1+k}{k}\,\theta^{r}\,(1-\theta)^k$, for $k = 0, 1, 2, \ldots$.

→ This is the Negative-Binomial(r, θ) Distribution.

→ Special case: If $r=1$, then $\mathrm{P}(Y=k) = \binom{1-1+k}{k}\,\theta^{1}\,(1-\theta)^k = \theta\,(1-\theta)^k$. This is the same as the Geometric(θ) Distribution (of course!).

• e.g. Suppose an urn contains N balls, of which M are Red and $N-M$ are Blue.

→ We draw n balls from the urn without replacement, so each collection of n balls has the same probability $1/\binom{N}{n}$.

→ Let X be the number of Red balls drawn. Probabilities?

→ Clearly $X \le n$, and $X \le M$, so $X \le \min(n, M)$. And $X \ge 0$.

→ Also, at most $N-M$ balls could be Blue, so $X \ge n-(N-M) = n+M-N$.

→ So, we want to find $\mathrm{P}(X=k)$, where $\max(0, n+M-N) \le k \le \min(n, M)$.

→ Well, $X=k$ if we chose k Red and $n-k$ Blue.

→ The number of such choices is $\binom{M}{k}\,\binom{N-M}{n-k}$.

→ Hence, $\mathrm{P}(X=k) = \binom{M}{k}\,\binom{N-M}{n-k} \,/\, \binom{N}{n}$.

→ This is the Hypergeometric(N, M, n) Distribution.

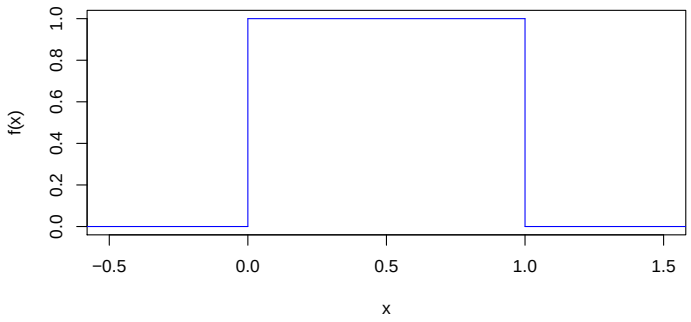
Suggested Homework: 2.3.9, 2.3.13, 2.3.16(c).

END MONDAY #4

Continuous Random Variables

- A random variable X is continuous if $\mathrm{P}(X = x) = 0$ for all x.
  → Then $\sum_{x \in \mathbf{R}} \mathrm{P}(X = x) = \sum_{x \in \mathbf{R}} 0 = 0$. The "opposite" of discrete!
- e.g. The Uniform[0,1] distribution (already mentioned):
  → $X \sim \text{Uniform}[0,1]$ if $\mathrm{P}(a \le X \le b) = b - a$ whenever $0 \le a \le b \le 1$.
  → Then e.g. $\mathrm{P}(X \in [0,1]) = \mathrm{P}(0 \le X \le 1) = 1 - 0 = 1$,
$\mathrm{P}(1/3 \le X \le 3/4) = (3/4) - (1/3) = 5/12$,
$\mathrm{P}(X \ge 2/3) = \mathrm{P}(2/3 \le X \le 1) = 1 - (2/3) = 1/3$, etc.
  → Also, $\mathrm{P}(X > 1) = 0$, and $\mathrm{P}(X < 0) = 0$, so e.g. $\mathrm{P}(1/3 \le X \le 5) = \mathrm{P}(1/3 \le X \le 1) = 1 - (1/3) = 2/3$, etc.
  → And, we previously showed (using Continuity Of Probabilities) that we can always replace "$\le$" with "$<$", or "$>$" by "$\ge$", etc. (Also true since $\mathrm{P}(X = x) = 0$.)
- Alternative representation: Let

$$f(x) = \begin{cases} 0, & x < 0 \\ 1, & 0 \le x \le 1 \\ 0, & x > 1 \end{cases}$$



  → Then for <u>any</u> $a \le b$,

$$\mathrm{P}(a \le X \le b) = \int_a^b f(x)\,dx\,.$$

  → And as a check, $f(x) \ge 0$, and $\int_{-\infty}^{\infty} f(x)\,dx = 1$. More complicated, but ...
- A density function is "any" $f : \mathbf{R} \to \mathbf{R}$ with $f(x) \ge 0$ and $\int_{-\infty}^{\infty} f(x)\,dx = 1$.
  → Given <u>any</u> density function, can define $\mathrm{P}(a \le X \le b) = \int_a^b f(x)\,dx$ for $a \le b$.
  → This defines a new distribution! Very general! ("absolutely continuous")
- Follows that $\mathrm{P}(X = a) = \mathrm{P}(a \le X \le a) = \int_a^a f(x)\,dx = 0$, i.e. X is continuous.
- If $f(x)$ is the density function for a random variable X, write it as $f_X(x)$.
- e.g. the Uniform[5,12] distribution has density: $f_X(x) = \begin{cases} 0, & x < 5 \\ 1/7, & 5 \le x \le 12 \\ 0, & x > 12 \end{cases}$
  → Then $f_X(x) \ge 0$, and $\int_{-\infty}^{\infty} f_X(x)\,dx = \int_{-\infty}^{5} (0)\,dx + \int_5^{12} (1/7)\,dx + \int_{12}^{\infty} (0)\,dx = 0 + (1/7)(7) + 0 = 1$. Good.
  → And then for $5 \le a \le b \le 12$, we have $P(a \le X \le b) = \frac{1}{7}(b - a)$.

- For any $L < R$, the Uniform[L,R] density is: $f_X(x) = \begin{cases} 0, & x < L \\ 1/(R-L), & L \le x \le R \\ 0, & x > R \end{cases}$

  → Then $f_X(x) \ge 0$, and $\int_{-\infty}^{\infty} f_X(x)\,dx = \int_{-\infty}^{L}(0)\,dx + \int_{L}^{R} \frac{1}{R-L}\,dx + \int_{R}^{\infty}(0)\,dx = 0 + \frac{1}{R-L}(R-L) + 0 = 1$. Good.

  → And then whenever $L \le a \le b \le R$, then $P(a \le X \le b) = \frac{b-a}{R-L}$.

- e.g. Let $f(x) = e^{-x}$ for $x \ge 0$, otherwise $f(x) = 0$.

  → Then $f(x) \ge 0$, and $\int_{-\infty}^{\infty} f(x)\,dx = \int_{-\infty}^{0}(0)\,dx + \int_{0}^{\infty} e^{-x}\,dx = (0) + (-e^{-x})\Big|_{x=0}^{x=\infty} = (-0) - (-1) = 1$.

  → If X has this density f, for $0 \le a \le b$, $\mathrm{P}(a \le X \le b) = \int_a^b e^{-x}\,dx = e^{-a} - e^{-b}$.

  → Also $\mathrm{P}(X \ge a) = e^{-a}$. This is the Exponential(1) distribution.

- More generally, for any $\lambda > 0$, let $f(x) = \lambda\, e^{-\lambda x}$ for $x \ge 0$, otherwise $f(x) = 0$.

  → Then $f(x) \ge 0$, and $\int_{-\infty}^{\infty} f(x)\,dx = \int_{-\infty}^{0}(0)\,dx + \int_{0}^{\infty}(\lambda\, e^{-\lambda x})\,dx = -e^{-\lambda x}\Big|_{x=0}^{x=\infty} = (-0) - (-1) = 1$.

  → If X has this density f, for $0 \le a \le b$, $\mathrm{P}(a \le X \le b) = e^{-\lambda a} - e^{-\lambda b}$.

  → Also $\mathrm{P}(X \ge a) = e^{-\lambda a}$. This is the Exponential(λ) distribution.

  → Many useful properties. Good model of e.g. how long a lightbulb will last.

Suggested Homework: 2.4.1, 2.4.2, 2.4.3, 2.4.4, 2.4.5, 2.4.6, 2.4.7, 2.4.8, 2.4.9, 2.4.10, 2.4.11, 2.4.12, 2.4.14.

The Normal Distribution

- Let $\phi(x) = \frac{1}{\sqrt{2\pi}}\, e^{-x^2/2}$ for $x \in \mathbf{R}$.

  → "Standard normal density"

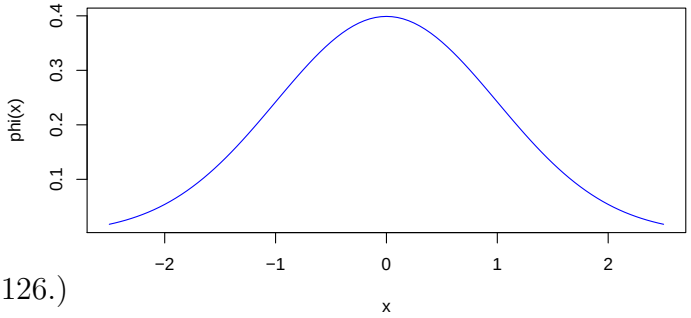
  → "bell curve", "Gaussian"

  → Clearly $\phi(x) \ge 0$.

  → Fact: $\int_{-\infty}^{\infty} \phi(x)\,dx = 1$.

  → (Proof uses polar coordinates: p. 126.)

  → So, it's a density. Important! Amazing!



- If X has density ϕ, then we say that X has the Normal(0,1) or N(0,1) distribution.

  → Then $\mathrm{P}(a \le X \le b) = \int_a^b \phi(x)\,dx = \int_a^b \frac{1}{\sqrt{2\pi}}\, e^{-x^2/2}\,dx$ for all $a \le b$.

  → Cannot be computed analytically. (No exact anti-derivative function.)

  → But can be computed using software, or tables like Appendix D.2.

- More generally, for any $\mu \in \mathbf{R}$ and $\sigma > 0$, let $f(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-(x-\mu)^2/2\sigma^2}$.

  → Then $f(x) \ge 0$. By change-of-variable theorem, $\int_{-\infty}^{\infty} f(x)\,dx = \int_{-\infty}^{\infty} \phi(x)\,dx = 1$.

  → This is the density of the Normal(μ, σ^2) or N(μ, σ^2) distribution.

  → Previous case was: $\mu = 0$, $\sigma = 1$. ("Standard normal distribution")

  → Curve is centered at μ, so changing μ "shifts" it.

→ Increasing σ makes it "fatter"; decreasing σ makes it "thinner".

→ [Plot in R: e.g. "plot(\(x) dnorm(x,2,3), xlim=c(-4,4), ylim=c(0,1))"]

• In fact, if $Z \sim \text{Normal}(0,1)$, and $W = \mu + \sigma Z$, then by the change-of-variable formula (coming soon), $W \sim \text{Normal}(\mu, \sigma^2)$.

• So, there is a normal density for every "location" μ and "scale" σ.

→ Good model for e.g. human heights, weights of eggs, etc.

→ See e.g. https://www.statology.org/example-of-normal-distribution/

→ The key distribution for the Central Limit Theorem and more! (Later.)

Suggested Homework: 2.4.13, 2.4.26.

Cumulative Distribution Functions (cdf)

• For any random variable X, the cumulative distribution function (cdf) is the function F_X defined by $F_X(x) = \text{P}(X \leq x)$ for all $x \in \mathbf{R}$.

→ If X is discrete, then $F_X(x) = \sum_{u \leq x} \text{P}(X = u)$.

→ Or, if X is absolutely continuous, then $F_X(x) = \int_{-\infty}^{x} f_X(u)\, du$.

• Then for any $a < b$, $\text{P}(a < X \leq b) = \text{P}(X \leq b) - \text{P}(X \leq a) = F_X(b) - F_X(a)$.

→ Also, by Continuity Of Probabilities, $\text{P}(a \leq X \leq b) = \text{P}(X \leq b) - \text{P}(X < a) = \text{P}(X \leq b) - \lim_{n\to\infty} \text{P}(X \leq a - \frac{1}{n}) = F_X(b) - \lim_{n\to\infty} F_X(a - \frac{1}{n})$.

→ Special case: $\text{P}(X = a) = \text{P}(a \leq X \leq a) = F_X(a) - \lim_{n\to\infty} F_X(a - \frac{1}{n})$.

→ (In particular, if F_X is continuous, then $\text{P}(a \leq X \leq b) = F_X(b) - F_X(a)$.)

→ So, all probabilites for X can be found from F_X. ("distribution function")

• Basic properties of any cumulative distribution function F_X:

→ $0 \leq F_X(x) \leq 1$ for all $x \in \mathbf{R}$.

→ If $x \leq y$, then $F_X(x) \leq F_X(y)$, i.e. F_X is an increasing function.

→ What about $\lim_{x\to-\infty} F_X(x)$ and $\lim_{x\to\infty} F_X(x)$?

→ By Continuity of Probabilities, $\lim_{x\to-\infty} F_X(x) = 0$ and $\lim_{x\to\infty} F_X(x) = 1$.

[Reminder: No lecture nor tutorial next Monday Oct 9 (Thanksgiving).]

[Reminder: Extra TA and Prof office hours added on web page.]

[Reminder: Midterm #1 next Wednesday Oct 11 in EX200.]

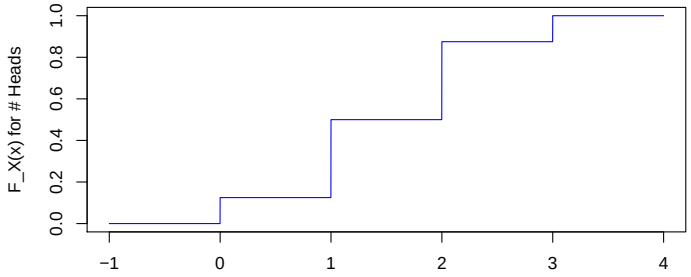
END WEDNESDAY #4

(Thanksgiving holiday.)

END MONDAY #5

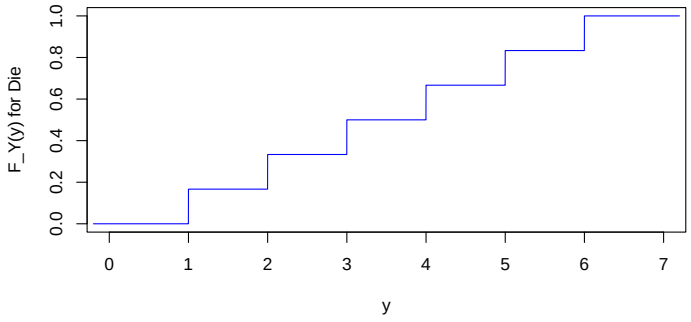
END WEDNESDAY #5

- Are cumulative distribution functions (cdfs) continuous?
- If $A = (-\infty, x]$ and $A_n = (-\infty, x + \frac{1}{n}]$, then:
$\{A_n\} \searrow A$, so $\mathrm{P}(A_n) \to \mathrm{P}(A)$, i.e. $F_X(x + \frac{1}{n}) \to F_X(x)$. "right-continuous"
- If $A = (-\infty, x]$ and $A_n = (-\infty, x - \frac{1}{n}]$, does $\{A_n\} \nearrow A$?
  - → No! $\{A_n\} \nearrow (-\infty, x)$. [Since $x \notin A_n$ for any n.]
  - → So, $\mathrm{P}(A_n) \to \mathrm{P}((-\infty, x)) = \mathrm{P}(X < x)$. [Not $\mathrm{P}(X \le x)$.]
  - → i.e. $F_X(x - \frac{1}{n}) \to \mathrm{P}(X < x) = \mathrm{P}(X \le x) - \mathrm{P}(X = x) = F_X(x) - \mathrm{P}(X = x)$.
- If $\mathrm{P}(X{=}x) = 0$, e.g. X continuous, then $F_X(x - \frac{1}{n}) \to F_X(x)$. "left-continuous"
  - → And, if it's right-continuous and left-continuous, then it is continuous!
- But if $\mathrm{P}(X = x) > 0$, then $F_X(x)$ is discontinuous at x.
  - → Furthermore, the jump-size at x is equal to $\mathrm{P}(X = x)$.
- e.g. Flip 3 coins, $X = \#$ Heads.
  - → Know $\mathrm{P}(X = 0) = 1/8$, $\mathrm{P}(X = 1) = 3/8$, $\mathrm{P}(X = 2) = 3/8$, $\mathrm{P}(X = 3) = 1/8$.
  - → So, for $x < 0$, $F_X(x) = \mathrm{P}(X \le x) = 0$.
  - → And, for $0 \le x < 1$, $F_X(x) = \mathrm{P}(X \le x) = \mathrm{P}(X = 0) = 1/8$.
  - → And, for $1 \le x < 2$, $F_X(x) = \mathrm{P}(X \le x) = \mathrm{P}(X = 0) + \mathrm{P}(X = 1) = 1/8 + 3/8 = 4/8 = 1/2$.
  - → And, for $2 \le x < 3$, $F_X(x) = \mathrm{P}(X \le x) = \mathrm{P}(X = 0) + \mathrm{P}(X = 1) + \mathrm{P}(X = 2) = 1/8 + 3/8 + 3/8 = 7/8$.



  - → And, for $x \ge 3$, $F_X(x) = \mathrm{P}(X \le x) = \mathrm{P}(X = 0) + \mathrm{P}(X = 1) + \mathrm{P}(X = 2) + \mathrm{P}(X = 3) = 1/8 + 3/8 + 3/8 + 1/8 = 1$.
  - → [Graph.] All properties satisfied!
- All discrete distributions have somewhat similar cdfs. (piecewise-constant)
- e.g. $Y =$ roll of one fair six-sided die.

$$F_Y(y) = \begin{cases} 0, & y < 1 \\ 1/6, & 1 \le y < 2 \\ 2/6, & 2 \le y < 3 \\ 3/6, & 3 \le y < 4 \\ 4/6, & 4 \le y < 5 \\ 5/6, & 5 \le y < 6 \\ 1, & y \ge 6 \end{cases}$$

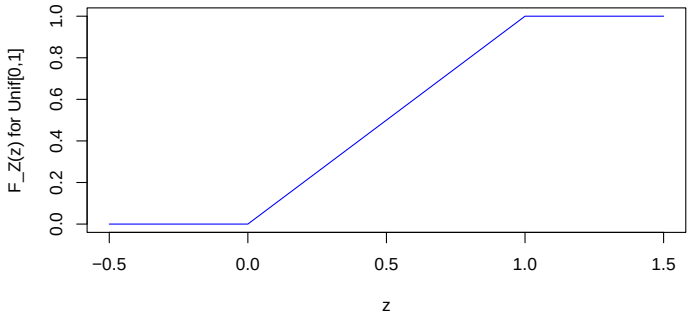


- Continuous? e.g. $X \sim \mathrm{Uniform}[0, 1]$.
  - → Then $\mathrm{P}(X \le x) = 0$ for $x < 0$.

→ And, $\mathrm{P}(X \le x) = 1$ for $x > 1$.

→ For $0 \le x \le 1$, $\mathrm{P}(X \le x) = \mathrm{P}(0 \le X \le x) = x - 0 = x$.

→ Hence, $F_X(x) = \begin{cases} 0, & x < 0 \\ x, & 0 \le x < 1 \\ 1, & x \ge 1 \end{cases}$



- e.g. $Z \sim \text{Uniform}[L, R]$ for some $L < R$. Then, similarly, $F_Z(z) = \begin{cases} 0, & z < L \\ \frac{z-L}{R-L}, & L \le z < R \\ 1, & z \ge R \end{cases}$

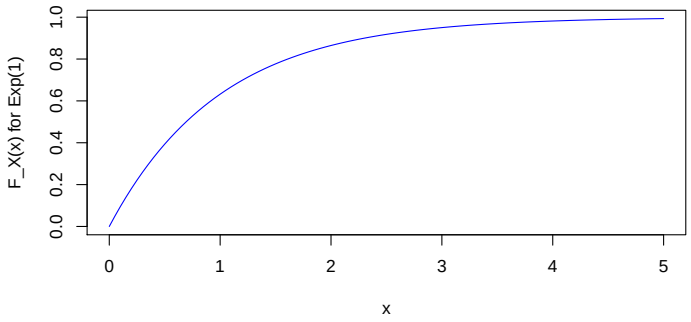
→ So e.g. if $L = 2$ and $R = 5$, then $F_Z(z) = \frac{z-2}{3}$ for $2 \le z \le 5$.

- e.g. $X \sim \text{Exponential}(1)$.

→ Then $\mathrm{P}(X < 0) = 0$.

→ So, for $x < 0$, $F_X(x) = 0$.

→ For $x \ge 0$, $F_X(x) = \mathrm{P}(X \le x) = \int_{-\infty}^{x} f_X(u)\,du = \int_0^x e^{-u}\,du = 1 - e^{-x}$. [Graph.] All properties satisfied!



- e.g. $Y \sim \text{Exponential}(5)$.

→ Then $\mathrm{P}(Y < 0) = 0$. So, for $y < 0$, $F_Y(y) = 0$.

→ For $y \ge 0$, $F_Y(y) = \mathrm{P}(Y \le y) = \int_{-\infty}^{y} f_Y(u)\,du = \int_0^y 5\,e^{-5u}\,du = 1 - e^{-5y}$.

- In general, if $W \sim \text{Exponential}(\lambda)$ for some $\lambda > 0$, then $F_W(w) = 0$ for $w < 0$, otherwise $F_W(w) = 1 - e^{-\lambda w}$.

- e.g. Suppose $X \sim \text{Exponential}(3)$. What is $\mathrm{P}(X \ge 2.6)$?

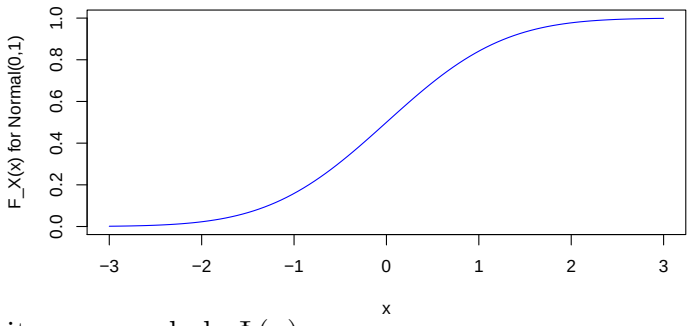
→ Here F_X is continuous, so $\mathrm{P}(X \ge 2.6) = 1 - \mathrm{P}(X < 2.6) = 1 - \mathrm{P}(X \le 2.6) = 1 - F_X(2.6) = 1 - [1 - e^{-3(2.6)}] = e^{-3(2.6)} = e^{-7.8} \doteq 0.00041$.

Suggested Homework: 2.5.2, 2.5.3, 2.5.7, 2.5.8, 2.5.9, 2.5.12.

- e.g. $Z \sim \text{Normal}(0, 1)$.

→ Then $F_Z(x) = \mathrm{P}(Z \le x) = \int_{-\infty}^{x} \phi(u)\,du = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-u^2/2}\,du$.

→ [Graph.] All properties satisfied!



→ Formula for this $F_Z(x)$?

→ There isn't one!

→ But it is so important that it has its own symbol: $\Phi(x)$.

→ It can be computed using software (R: "pnorm"), or tables like Appendix D.2.

→ By symmetry, $\mathrm{P}(Z \le x) = \mathrm{P}(Z \ge -x)$. So, $\Phi(x) = 1 - \Phi(-x)$ for all $x \in \mathbf{R}$, i.e. $\Phi(x) + \Phi(-x) = 1$. Also, $\Phi(0) = 1/2$.

- e.g. Suppose $Z \sim \text{Normal}(0, 1)$. What is $\mathrm{P}(Z \le 1.43)$?

→ Well, $\mathrm{P}(Z \leq 1.43) = \Phi(1.43) = 1 - \Phi(-1.43)$.

→ From the table in Appendix D.2, this is $\doteq 1 - (0.0764) = 0.9236$.

- e.g. Suppose $W \sim \text{Normal}(5, 4^2)$. What is $\mathrm{P}(6 \leq W \leq 8)$?

→ Well, here $W = 5 + 4\,Z$ where $Z \sim \text{Normal}(0, 1)$.

→ So, $\mathrm{P}(6 \leq W \leq 8) = \mathrm{P}(6 \leq 5 + 4\,Z \leq 8) = \mathrm{P}(1/4 \leq Z \leq 3/4)$.

→ By definition of Φ, this is $\mathrm{P}(Z \leq 3/4) - \mathrm{P}(Z \leq 1/4) = \Phi(3/4) - \Phi(1/4)$.

→ Then, this equals $[1 - \Phi(-3/4)] - [1 - \Phi(-1/4)] = \Phi(-1/4) - \Phi(-3/4) = \Phi(-0.25) - \Phi(-0.75)$.

→ From the Appendix D.2 table, this is $\doteq 0.4013 - 0.2266 = 0.1747$.

→ So, here $\mathrm{P}(6 \leq W \leq 8) \doteq 0.1747$.

Suggested Homework: 2.5.4, 2.5.5.

END MONDAY #6

- Suppose that X is absolutely continuous, with density function $f_X(x)$, and cumulative distribution function $F_X(x)$. What is the relationship between f_X and F_X?

→ Well, we know that $F_X(x) := \mathrm{P}(X \leq x) = \int_{-\infty}^{x} f_X(u)\,du$.

→ So, by the Fundamental Theorem of Calculus, the derivative $F_X'(x) := \frac{d}{dx} F_X(x)$ equals $f_X(x)$, at least if f_X is continuous at x.

→ That is, the derivative of the cdf is the density!

- e.g. Suppose $X \sim \text{Exponential}(1)$. Then we know $F_X(x) = 1 - e^{-x}$ for $x \geq 0$.

→ Then for $x > 0$, $F_X'(x) = \frac{d}{dx}[1 - e^{-x}] = -(-e^{-x}) = e^{-x} = f_X(x)$. Yep!

- e.g. Similarly, for any $\lambda > 0$, if $Y \sim \text{Exponential}(\lambda)$, then for $y > 0$, $F_Y(y) = 1 - e^{-\lambda y}$, and $F_Y'(y) = \frac{d}{dy}[1 - e^{-\lambda y}] = (-\lambda)(-e^{-\lambda y}) = \lambda e^{-\lambda y} = f_Y(y)$. Yep!

- If $Z \sim \text{Normal}(0, 1)$, then we know $\Phi'(z) = \phi(z) = \frac{1}{\sqrt{2\pi}}\, e^{-z^2/2}$.

→ Even though we don't really know exactly what $\Phi(z)$ is!

- e.g. Suppose a r.v. X has cdf $F_X(x) = \begin{cases} 0, & x < 5 \\ (x-5)^4, & 5 \leq x < 6 \\ 1, & x \geq 6 \end{cases}$

→ Valid cdf? (Yes! Increases from 0 to 1, right-continuous . . .)

→ Then e.g. $\mathrm{P}(3 < X \leq 5.5) = F_X(5.5) - F_X(3) = (5.5 - 5)^4 - 0 = 0.0625$.

→ e.g. Also, X has density function $f_X(x) = F_X'(x) = \begin{cases} 0, & x < 5 \\ 4(x-5)^3, & 5 < x < 6 \\ 0, & x > 6 \end{cases}$

- Mixture Distributions: e.g. Consider the following random variables:

→ Y is the result of rolling one fair six-sided die, with cdf $F_Y(y)$ as above.

→ $Z \sim \text{Uniform}[2, 5]$, with cdf $F_Z(z) = \frac{z-2}{3}$ for $2 \leq z \leq 5$ as above.

→ $W \sim \text{Bernoulli}(1/3)$ (indep.), so $\mathrm{P}(W = 1) = 1/3$ and $\mathrm{P}(W = 0) = 2/3$.

→ Then, we let $X = \begin{cases} Y, & W = 1 \\ Z, & W = 0 \end{cases}$

→ Intuitively, X is equal either to the result of the die (with probability 1/3), or to a Uniform[2,5] variable (with probability 2/3).

→ Then what is, say, $F_X(4.4)$?

→ Well, by the Law of Total Probability, $F_X(4.4) := \mathrm{P}(X \leq 4.4)$
$= \mathrm{P}(X \leq 4.4,\ W = 1) + \mathrm{P}(X \leq 4.4,\ W = 0)$
$= \mathrm{P}(Y \leq 4.4,\ W = 1) + \mathrm{P}(Z \leq 4.4,\ W = 0)$
$= \mathrm{P}(Y \leq 4.4)\ \mathrm{P}(W = 1) + \mathrm{P}(Z \leq 4.4)\ \mathrm{P}(W = 0)$
$= F_Y(4.4)\,(1/3) + F_Z(4.4)\,(2/3) = (4/6)\,(1/3) + (2.4/3)\,(2/3).$

→ More generally, $F_X(x) = (1/3)\,F_Y(x) + (2/3)\,F_Z(x)$, for all $x \in \mathbf{R}$.

→ (Can then plug in $F_Y(x)$ and $F_Z(x)$ to compute $F_X(x)$.)

→ The distribution of X is a mixture of the distributions of Y and of Z.

- In this example, is X continuous?

→ No! By independence, we have that e.g. $\mathrm{P}(X = 2) = \mathrm{P}(W = 1,\ Y = 2) = \mathrm{P}(W = 1)\ \mathrm{P}(Y = 2) = (1/3)(1/6) = 1/18 > 0$. Not zero, like for the continuous case.

- Ah, so then is X discrete?

→ No! Here $\sum_{x \in \mathbf{R}} \mathrm{P}(X = x) = \sum_{x=1}^{6} \mathrm{P}(X = x) = \sum_{x=1}^{6} \mathrm{P}(W = 1,\ Y = x) = \sum_{x=1}^{6} \mathrm{P}(W = 1)\ \mathrm{P}(Y = x) = \sum_{x=1}^{6} (1/3)(1/6) = 1/3 < 1$. Not one, like for the discrete case.

- Here X is has a mixture distribution. Neither discrete nor continuous!

→ (In this course we'll usually stick with either discrete or absolutely continuous. But there are other kinds of random variables too. Even beyond mixtures!)

Suggested Homework: 2.5.6, 2.5.13, 2.5.14, 2.5.15, 2.5.17, 2.5.18.

Change of Variable Formula (one-dimensional)

- Suppose X is a random variable, and $h : \mathbf{R} \to \mathbf{R}$ is some function.

→ Then we can define $Y = h(X)$, i.e. $Y(s) = h(X(s))$ for all $s \in S$. (e.g. $Y = X^2$)

→ Then Y is another random variable. ("function of a random variable")

→ So, Y has its own distribution. What is it??

- Discrete Case: Suppose X discrete: $\mathrm{P}(X = x_i) = p_i$ where $p_i \geq 0$ and $\sum_i p_i = 1$.

→ Then, Y is discrete too, with $\mathrm{P}(Y = y) = \mathrm{P}(h(X) = y) = \sum\{p_i : h(x_i) = y\}$.

→ That is, $\mathrm{P}(Y = y) = \mathrm{P}(X \in \{x : h(x) = y\})$.

→ Or, in terms of probability functions, $p_Y(y) = \sum_{x\,:\,h(x)=y} p_X(x)$.

→ Discrete Change-of-Variable Theorem.

- e.g. X = roll of fair die, and $Y = (X-3)^2$. What is $\mathrm{P}(Y = 4)$?

→ Well, $\mathrm{P}(Y = 4) = \mathrm{P}(X \in \{x : (x-3)^2 = 4\}) = \mathrm{P}(X \in \{1, 5\}) = (1/6)+(1/6) = 2/6 = 1/3$.

→ Also, $\mathrm{P}(Y=1) = \mathrm{P}(X \in \{x : (x-3)^2 = 1\}) = \mathrm{P}(X \in \{2,4\}) = (1/6)+(1/6) = 2/6 = 1/3$.

→ And, $\mathrm{P}(Y=9) = \mathrm{P}(X \in \{x : (x-3)^2 = 9\}) = \mathrm{P}(X \in \{6\}) = (1/6)$. More?

→ Yes! Also $\mathrm{P}(Y=0) = \mathrm{P}(X \in \{x : (x-3)^2 = 0\}) = \mathrm{P}(X \in \{3\}) = (1/6)$.

→ That is, $p_Y(y) = 1/3$ for $y = 1, 4$; $p_Y(y) = 1/6$ for $y = 0, 9$; otherwise 0.

- Easy! But what if X is continuous? Trickier!
- Absolutely Continuous Case: Suppose X has density $f_X(x)$, and $Y = h(X)$.

→ Then what is the density function $f_Y(y)$ for Y?

→ Will Y necessarily even be absolutely continuous too??

→ No, not necessarily!

- e.g. $X \sim \text{Uniform}[0,1]$, and $h(x) = \begin{cases} 2, & x \leq 1/3 \\ 4, & x > 1/3 \end{cases}$

→ Then if $Y = h(X)$, then $\mathrm{P}(Y=2) = \mathrm{P}(X \leq 1/3) = 1/3$, and $\mathrm{P}(Y=4) = \mathrm{P}(X > 1/3) = 1 - (1/3) = 2/3$. That is, $p_Y(2) = 1/3$, and $p_Y(4) = 2/3$.

→ So, Y is discrete! Not continuous at all!

- But what if h is strictly increasing? (or decreasing?) Then what is $f_Y(y)$?
- Absolutely Continuous Change-of-Variable Theorem: Suppose X has density $f_X(x)$, and $Y = h(X)$, where $h : \mathbf{R} \to \mathbf{R}$ is differentiable and strictly increasing or decreasing (at least on $\{x : f_X(x) > 0\}$), with inverse function $h^{-1}(y)$. Then Y is also absolutely continuous, with density function $f_Y(y) = f_X\big(h^{-1}(y)\big) \,/\, \big|h'\big(h^{-1}(y)\big)\big|$.
- Proof: Suppose h is strictly increasing.

→ Then h has an inverse function, $h^{-1}(y)$. So, $X = h^{-1}(Y)$.

→ Also assume h has a derivative, $h'(x)$.

→ Then by the Inverse Function Theorem, $\frac{d}{dy}\, h^{-1}(y) := \big(h^{-1}\big)'(y) = 1 \,/\, h'(h^{-1}(y))$.

- Method #1:

→ Then $\mathrm{P}(a \leq Y \leq b) = \mathrm{P}(h^{-1}(a) \leq X \leq h^{-1}(b)) = \int_{h^{-1}(a)}^{h^{-1}(b)} f_X(x)\, dx$.

→ We now make the "substitution" $x = h^{-1}(y)$.

→ Then by "integration by subsitution" or the "chain rule" from calculus, we have $dx = d\big(h^{-1}(y)\big) = \big(h^{-1}\big)'(y)\, dy = [1/h'(h^{-1}(y))]\, dy$.

→ Hence, from above, $\mathrm{P}(a \leq Y \leq b) = \int_a^b \big[f_X\big(h^{-1}(y)\big)/h'(h^{-1}(y))\big]\, dy$, $\forall a \leq b$.

→ But this equals $\int_a^b f_Y(y)\, dy$, so we must have $f_Y(y) = f_X\big(h^{-1}(y)\big)/h'\big(h^{-1}(y)\big)$.

→ (The first part $f_X\big(h^{-1}(y)\big)$ is intuitive. The rest is from the chain rule.)

- Method #2:

→ Then $F_Y(y) = \mathrm{P}(Y \leq y) = \mathrm{P}(h(X) \leq y) = \mathrm{P}(X \leq h^{-1}(y)) = F_X\big(h^{-1}(y)\big)$.

→ So, $f_Y(y) = \frac{d}{dy} F_Y(y) = \frac{d}{dy} F_X\big(h^{-1}(y)\big) = f_X\big(h^{-1}(y)\big)\, \frac{d}{dy}\, h^{-1}(y)$
$= f_X\big(h^{-1}(y)\big)\, [1/\, h'\big(h^{-1}(y)\big)] = f_X\big(h^{-1}(y)\big) \,/\, h'\big(h^{-1}(y)\big)$.

- Note: We need h to be increasing only where $f_X(x) > 0$; other x don't matter.

• If instead h is strictly decreasing, then everything is still the same, except that h' and $(h^{-1})'$ are negative, so we need to put an absolute value sign on it.

→ Or, in Method #2, $\mathrm{P}(Y \le y) = \mathrm{P}(X \ge h^{-1}(y)) = 1 - \mathrm{P}(X \le h^{-1}(y)) = 1 - F_X(h^{-1}(y))$ which gives a negative. ∎

• e.g. Suppose $X \sim \text{Uniform}[0,1]$, and $Y = 5X + 4$.

→ Then $f_X(x) = 1$ for $0 \le x \le 1$, otherwise 0.

→ Also $h(x) = 5x + 4$, strictly increasing, $h'(x) = 5$.

→ And, if $y = 5x + 4$, then $x = (y-4)/5$, so $h^{-1}(y) = (y-4)/5$.

→ So, $f_X(h^{-1}(y)) = f_X\big((y-4)/5\big)$, which $= 1$ for $4 \le y \le 9$ otherwise 0.

→ And, $h'\big(h^{-1}(y)\big) = h'\big((y-4)/5\big) = 5$.

→ So, $f_Y(y) = f_X\big(h^{-1}(y)\big) \,/\, \big|h'\big(h^{-1}(y)\big)\big| = 1/5$ for $4 \le y \le 9$ otherwise 0.

→ That is, $Y \sim \text{Uniform}[4,9]$, a familiar distribution! (Makes sense.)

• Alternatively, use cdfs!

→ In above example, for $4 \le y \le 9$:

→ $F_Y(y) = \mathrm{P}(Y \le y) = \mathrm{P}(5X + 4 \le y) = \mathrm{P}(X \le (y-4)/5) = (y-4)/5$.

→ Hence, for $4 \le y \le 9$, $f_Y(y) = \frac{d}{dy} F_Y(y) = \frac{d}{dy}(y-4)/5 = 1/5$. Same as before!

• e.g. Suppose $X \sim \text{Uniform}[0,1]$, and $Y = X^2$.

→ Then $f_X(x) = 1$ for $0 \le x \le 1$, otherwise 0.

→ Also $h(x) = x^2$, strictly increasing for $x \ge 0$, and $h'(x) = 2x$.

→ And, $h^{-1}(y) = \sqrt{y}$ for $y \ge 0$, so $f_X(h^{-1}(y))$ is 1 for $0 < y \le 1$ otherwise 0.

→ Therefore, $h'(h^{-1}(y)) = 2\,h^{-1}(y) = 2\sqrt{y}$ for $y > 0$, otherwise 0.

→ So, $f_Y(y) = f_X\big(h^{-1}(y)\big) \,/\, \big|h'\big(h^{-1}(y)\big)\big| = 1/(2\sqrt{y})$ for $0 < y \le 1$ otherwise 0.

→ Is that really correct? Check: $\int_{-\infty}^{\infty} f_Y(y)\,dy = \int_0^1 [1/(2\sqrt{y})]\,dy = \frac{1}{2}\int_0^1 y^{-1/2}\,dy = \frac{1}{2}\big(2y^{1/2}\big)\Big|_{y=0}^{y=1} = \frac{1}{2}\big(2[1^{1/2} - 0^{1/2}]\big) = \frac{1}{2}\cdot 2 \cdot 1 = 1$. Phew! (And Y is not uniform.)

→ Alternatively: For $0 \le y \le 1$, $F_y(y) = \mathrm{P}(Y \le y) = \mathrm{P}(X^2 \le y) = \mathrm{P}(X \le \sqrt{y}) = \sqrt{y}$, so $f_Y(y) = \frac{d}{dy} F_Y(y) = \frac{d}{dy}\sqrt{y} = \frac{d}{dy} y^{1/2} = (1/2)y^{-1/2} = 1/(2\sqrt{y})$.

Suggested Homework: 2.6.1, 2.6.2, 2.6.3, 2.6.4, 2.6.5, 2.6.6, 2.6.7, 2.6.9, 2.6.10, 2.6.12, 2.6.14, 2.6.15.

END WEDNESDAY #6

• e.g. Suppose $X \sim \text{Exponential}(5)$, and $Y = X^2$.

→ Then for $y > 0$, $F_y(y) = \mathrm{P}(Y \le y) = \mathrm{P}(X^2 \le y) = \mathrm{P}(X \le \sqrt{y}) = 1 - e^{-5\sqrt{y}}$.

→ So, for $y > 0$, $f_Y(y) = \frac{d}{dy} F_Y(y) = \frac{d}{dy}[1 - e^{-5\sqrt{y}}] = -e^{-5\sqrt{y}}\,(-5y^{-1/2}/2)) = (5/2)e^{-5\sqrt{y}}/\sqrt{y}$. (Otherwise $f_Y(y) = 0$.) Crazy, but true! [Check: Integrates to 1.]

→ Or, use the usual Theorem: Again $h(x) = x^2$, strictly increasing for $x \ge 0$, $h'(x) = 2x$, $h^{-1}(y) = \sqrt{y}$ for $y \ge 0$, and here $f_X(x) = 5e^{-5x}$ for $x \ge 0$, so for $y \ge 0$, $f_Y(y) = f_X\big(h^{-1}(y)\big) \,/\, \big|h'\big(h^{-1}(y)\big)\big| = 5e^{-5\sqrt{y}}/2\sqrt{y}$. Same!

- e.g. Suppose $Z \sim \text{Normal}(0,1)$, and $Y = 6 + 3Z$.

  → Then $f_Z(z) = \phi(z) = \frac{1}{\sqrt{2\pi}}\, e^{-z^2/2}$.

  → Also $h(z) = 6+3z$, strictly increasing, with $h'(z) = 3$. And, $h^{-1}(y) = (y-6)/3$.

  → So, $f_Y(y) = f_Z\big(h^{-1}(y)\big) \;/\; \big|h'\big(h^{-1}(y)\big)\big| = \phi\big((y-6)/3\big) \;/\; 3$
$= \frac{1}{\sqrt{2\pi}}\, e^{-[(y-6)/3]^2/2} \;/\; 3 = \frac{1}{3\sqrt{2\pi}}\, e^{-(y-6)^2/(2\cdot 3^2)}$.

  → This is the same as $\frac{1}{\sigma\sqrt{2\pi}}\, e^{-(y-\mu)^2/(2\sigma^2)}$ where $\mu = 6$ and $\sigma = 3$.

  → Hence, $Y \sim \text{Normal}(6, 3^2)$, as we expected.

  → (Similarly for any μ besides 6, and σ besides 3.)

  → This demonstrates that if $Z \sim \text{Normal}(0,1)$, and $Y = \mu + \sigma Z$, then $Y \sim \text{Normal}(\mu, \sigma^2)$, as we claimed before. (Phew.)

Joint Distributions

- Suppose X and Y are two random variables.

  → Suppose we know the distribution of X and also know the distribution of Y.

  → Does that tell us the whole story? Maybe not!

- e.g. Suppose we flip two fair (independent) coins.

  → Let $X = I_{\text{first coin Heads}}$, i.e. $X = 1$ if first coin Heads, otherwise $X = 0$.

  → Then $X \sim \text{Bernoulli}(1/2)$, i.e. $\text{P}(X = 0) = \text{P}(X = 1) = 1/2$.

  → Let $Y_1 = X$, $Y_2 = 1 - X$, and $Y_3 = I_{\text{second coin Heads}}$. Distributions?

  → Here $Y_1 \sim \text{Bernoulli}(1/2)$, and $Y_2 \sim \text{Bernoulli}(1/2)$, and $Y_3 \sim \text{Bernoulli}(1/2)$.

  → But what about their relationships to X? e.g. $\text{P}(X = 1,\ Y_i = 1)$?

  → Here $\text{P}(X = 1,\ Y_1 = 1) = 1/2$, and $\text{P}(X = 1,\ Y_2 = 1) = 0$, and $\text{P}(X = 1,\ Y_3 = 1) = 1/4$. All different!

- To really understand multiple variables, we need their joint distribution.

  → How to keep track? Joint probability functions (discrete case), joint density functions (absolutely continuous case), joint cdfs (most general; we'll do them first).

Joint Cumulative Distribution Functions

- Given random variables X and Y, their joint cumulative distribution function or joint cdf is the function $F_{X,Y} : \mathbf{R}^2 \to [0,1]$ given by $F_{X,Y}(x,y) = \text{P}(X \le x,\ Y \le y) \equiv \text{P}(X \le x \text{ and } Y \le y)$. Can get tricky!

- Again let $X = I_{\text{first coin Heads}}$, $Y_1 = X$, $Y_2 = 1 - X$, and $Y_3 = I_{\text{second coin Heads}}$.

  → If $x < 0$ or $y < 0$ (or both), then $F_{X,Y_i}(x,y) = 0$ for each i (of course).

  → If $x \ge 1$ and $y \ge 1$, then $F_{X,Y_i}(x,y) = 1$ for each i (of course).

  → For $Y_1 = X$: If $0 \le \min[x,y] < 1$, then $F_{X,Y_1}(x,y) = \text{P}(X \le x,\ Y_1 \le y) = \text{P}(X \le x,\ X \le y) = \text{P}(X \le \min[x,y]) = \text{P}(X = 0) = 1/2$. Hence,

$$F_{X,Y_1}(x,y) \;=\; \begin{cases} 1, & x \ge 1 \text{ and } y \ge 1 \\ 1/2, & 0 \le \min[x,y] < 1 \\ 0, & x < 0 \text{ or } y < 0 \text{ or both} \end{cases}$$

→ Alternatively (easier?), compute $F_{X,Y_1}(x,y)$ systematically using a big table:

$F_{X,Y_1}(x,y)$	$x<0$	$0\le x<1$	$x\ge 1$
$y<0$	0	0	0
$0\le y<1$	0	1/2	1/2
$y\ge 1$	0	1/2	1

→ What about $Y_2 = 1 - X$? Well, if $0 \le x < 1$ and $y \ge 1$, then $F_{X,Y_2}(x,y) = \mathrm{P}(X \le x,\ Y_1 \le y) = \mathrm{P}(X \le x,\ 1 - X \le y) = \mathrm{P}(X = 0,\ 1 - X = 1) = \mathrm{P}(X = 0) = 1/2$. Also true if $0 \le y < 1$ and $x \ge 1$. But if $x < 1$ and $y < 1$, then cannot have both $X \le x$ and $1 - X \le y$, so $F_{X,Y_2}(x,y) = 0$. Hence,

$F_{X,Y_2}(x,y)$	$x<0$	$0\le x<1$	$x\ge 1$
$y<0$	0	0	0
$0\le y<1$	0	0	1/2
$y\ge 1$	0	1/2	1

→ What about $Y_3 = I_{\text{second coin Heads}}$? Well, if $0 \le x < 1$ and $y \ge 1$, then $F_{X,Y_3}(x,y) = \mathrm{P}(X \le x,\ Y_1 \le y) = \mathrm{P}(X = 0) = 1/2$. Also true if $0 \le y < 1$ and $x \ge 1$. But if $0 \le x < 1$ and $0 \le y < 1$, then $\mathrm{P}(X \le x,\ Y \le y) = \mathrm{P}(X = 0,\ Y = 0) = (1/2)(1/2) = 1/4$. Hence,

$F_{X,Y_3}(x,y)$	$x<0$	$0\le x<1$	$x\ge 1$
$y<0$	0	0	0
$0\le y<1$	0	1/4	1/2
$y\ge 1$	0	1/2	1

→ So, e.g. $F_{X,Y_1}(1/2,1/2) = 1/2$, $F_{X,Y_2}(1/2,1/2) = 0$, $F_{X,Y_3}(1/2,1/2) = 1/4$.

→ All different! Relationships matter! (But $F_{X,Y}(x,y)$ awkward to work with.)

- Some "limit" properties of $F_{X,Y}(x,y) := \mathrm{P}(X \le x,\ Y \le y)$:

→ $\lim_{x\to-\infty} F_{X,Y}(x,y) = 0$ for all y, and $\lim_{y\to-\infty} F_{X,Y}(x,y) = 0$ for all x.

→ $\lim_{x\to+\infty} F_{X,Y}(x,y) = F_Y(y)$ for all y, and $\lim_{y\to+\infty} F_{X,Y}(x,y) = F_X(x)$ for all x.

→ "Marginal cdfs": the joint cdf tells us all about the individual ones.

→ In above example, bottom row is $F_X(x)$, and right column is $F_Y(y)$.

- What about $\mathrm{P}(a < X \le b,\ c < Y \le d)$?

→ Well, $\mathrm{P}(a < X \le b,\ Y \le d) = \mathrm{P}(X \le b,\ Y \le d) - \mathrm{P}(X \le a,\ Y \le d) = F_{X,Y}(b,d) - F_{X,Y}(a,d)$.

→ Hence, $\mathrm{P}(a < X \le b,\ c < Y \le d) = \mathrm{P}(a < X \le b,\ Y \le d) - \mathrm{P}(a < X \le b,\ Y \le c) = [F_{X,Y}(b,d) - F_{X,Y}(a,d)] - [F_{X,Y}(b,c) - F_{X,Y}(a,c)]$,

→ So, $\mathrm{P}(a < X \le b,\ c < Y \le d) = F_{X,Y}(b,d) - F_{X,Y}(a,d) - F_{X,Y}(b,c) + F_{X,Y}(a,c)$.

→ Intuitive from Diagram:

Suggested Homework: 2.7.1, 2.7.2, 2.7.5, 2.7.6(a), 2.7.11.

END MONDAY #7

Joint Probability Functions

• If X and Y are discrete, then we can keep track of their relationship by the joint probability function $p_{X,Y}(x,y) := \mathrm{P}(X = x, Y = y)$.

• e.g. In above example, $p_{X,Y_1}(1,1) = 1/2$ and $p_{X,Y_1}(0,0) = 1/2$ (otherwise $p_{X,Y_1}(x,y) = 0$, e.g. $p_{X,Y_1}(1,0) = 0$). Also $p_{X,Y_2}(1,0) = 1/2$ and $p_{X,Y_2}(0,1) = 1/2$. Also $p_{X,Y_3}(1,1) = 1/4$ and $p_{X,Y_3}(1,0) = 1/4$ and $p_{X,Y_3}(0,1) = 1/4$ and $p_{X,Y_3}(0,0) = 1/4$.

• If we know $p_{X,Y}(x,y)$, can we find $p_X(x)$ and $p_Y(y)$?

→ Yes! From the Law of Total Probability (Unconditioned Version), $p_X(x) = \mathrm{P}(X = x) = \sum_y \mathrm{P}(X = x,\ Y = y) = \sum_y p_{X,Y}(x,y)$ for all x. Similarly $p_Y(y) = \sum_x p_{X,Y}(x,y)$ for all y. ("marginals") So, $p_{X,Y}(x,y)$ has all the information.

• e.g. In above example, $p_X(1) = p_{X,Y_3}(1,0) + p_{X,Y_3}(1,1) = 1/4 + 1/4 = 1/2$, etc.

→ Can also write e.g. $p_{X,Y_3}(x,y)$ in a table, with $p_X(x)$ and $p_{Y_3}(y)$ at the right and bottom margins, which is why they are called the "marginals":

	$Y_3 = 0$	$Y_3 = 1$	$p_X(x)$
$X = 0$	1/4	1/4	1/2
$X = 1$	1/4	1/4	1/2
$p_{Y_3}(y)$	1/2	1/2	

• Then e.g. $\mathrm{P}(a \le X \le b,\ c \le Y \le d) = \sum_{a \le x \le b} \sum_{c \le y \le d} p_{X,Y}(x,y)$, etc.

Suggested Homework: 2.7.3, 2.7.6.

Joint Density Functions

• Random variables X and Y are jointly absolutely continuous if there is a joint density function $f_{X,Y} : \mathbf{R}^2 \to \mathbf{R}$, which is ≥ 0, with $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dx\,dy = 1$, such that $\mathrm{P}(a \le X \le b,\ c \le Y \le d) = \int_c^d \int_a^b f_{X,Y}(x,y)\,dx\,dy$ for all $a \le b$ and $c \le d$.

• Two-dimensional ("iterated") integral! (e.g. Appendix A.6.) [MAT237 – later]

→ Compute the "inner" integral first, treating the outer variable as constant.

→ Then, integrate the resulting expression as the outer integral.

→ Trickiest part: specify the inner limits of integration correctly, to ensure that the point (x,y) is always within the correct region (see example below).

→ Can integrate in either order ("Fubini's Thm"), provided you do it correctly!

• Marginals? Similar to discrete case – "add up" the other variable.

→ $\mathrm{P}(a \le X \le b) = \mathrm{P}(a \le X \le b,\ -\infty < Y < \infty) = \int_{-\infty}^{\infty} \int_a^b f_{X,Y}(x,y)\,dx\,dy$.

→ But $\mathrm{P}(a \le X \le b) = \int_a^b f_X(x)\,dx$, for all $a \le b$.

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

→ Similarly, $f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dx$.

- Running example: $f_{X,Y}(x,y) = \frac{15}{32}xy^2$ for $0 \le y \le x \le 2$, otherwise 0. Diagram:

- Valid joint density function?

→ Here $f_{X,Y} \ge 0$, and $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dx\,dy = \int_0^2\int_y^2(\frac{15}{32}xy^2)\,dx\,dy = \int_0^2(\frac{15}{32}\frac{1}{2}x^2y^2)\Big|_{x=y}^{x=2}dy =$ $\int_0^2[\frac{15}{64}(2^2-y^2)y^2]\,dy = \frac{15}{64}[2^2\frac{1}{3}y^3 - \frac{1}{5}y^5]\Big|_{y=0}^{y=2} = \frac{15}{64}[\frac{4}{3}(2^3-0) - \frac{1}{5}(2^5-0)] = 1$. So, yes!

- What is $\mathrm{P}(0 \le X \le 1/2,\ 0 \le Y \le 1/4)$? We compute this as ...

→ $\int_0^{1/4}\int_y^{1/2}(\frac{15}{32}xy^2)\,dx\,dy = \int_0^{1/4}(\frac{15}{32}\frac{1}{2}x^2y^2)\Big|_{x=y}^{x=1/2}dy = \int_0^{1/4}[\frac{15}{64}((1/2)^2-y^2)y^2]\,dy =$ $\frac{15}{64}[(1/2)^2\frac{1}{3}y^3 - \frac{1}{5}y^5]\Big|_{y=0}^{y=1/4} = \frac{15}{64}[\frac{1}{12}((1/4)^3-0) - \frac{1}{5}((1/4)^5-0)] = 17/65536 \doteq 0.00026$.

→ Exercise: Compute $\mathrm{P}(7/4 \le X \le 2,\ 3/2 \le Y \le 2)$. Is it larger?

- What is $f_X(x)$, the density function of X?

→ For $0 \le x \le 2$, $f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dy = \int_0^x(\frac{15}{32}xy^2)\,dy = (\frac{15}{32}\frac{1}{3}xy^3)\Big|_{y=0}^{y=x} =$ $\frac{15}{32}\frac{1}{3}x(x^3-0^3) = (5/32)\,x^4$. (Otherwise $f_X(x) = 0$ if $x < 0$ or $x > 2$.)

→ Check: $\int_{-\infty}^{\infty} f_X(x)\,dx = \int_0^2(5/32)\,x^4\,dx = (5/32)\,\frac{1}{5}x^5\Big|_{x=0}^{x=2} = (5/32)\,\frac{1}{5}(2^5-0^5) =$ 1. Phew!

→ So e.g. $\mathrm{P}(X \le 1/3) = \int_0^{1/3} f_X(x)\,dx = \int_0^{1/3}(5/32)x^4\,dx = (5/32)\frac{1}{5}x^5\Big|_{x=0}^{x=1/3} =$ $(5/32)\frac{1}{5}((1/3)^5-0^5) = 1/7776 \doteq 0.00013$.

- What is $f_Y(y)$, the density function of Y?

→ For $0 \le y \le 2$, $f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dx = \int_y^2(\frac{15}{32}xy^2)\,dx = (\frac{15}{32}\frac{1}{2}x^2y^2)\Big|_{x=y}^{x=2} =$ $\frac{15}{32}\frac{1}{2}(2^2-y^2)y^2 = \frac{15}{64}(4y^2-y^4)$. (Otherwise $f_Y(y) = 0$ if $y < 0$ or $y > 2$.)

→ Check: $\int_{-\infty}^{\infty} f_Y(y)\,dy = \int_0^2\frac{15}{64}(4y^2-y^4)\,dy = \frac{15}{64}[4\frac{1}{3}y^3 - \frac{1}{5}y^5)]\Big|_{y=0}^{y=2} = \frac{15}{64}[4\frac{1}{3}(2^3 -$ $0^3) - \frac{1}{5}(2^5-0^5)] = 1$. Phew!

Suggested Homework: 2.7.4, 2.7.7, 2.7.8, 2.7.9, 2.7.14, 2.7.15, 2.7.16.

Conditioning and Independence for Discrete Random Variables

- Suppose X and Y are discrete with joint probability function $p_{X,Y}$ given (in tabular form) by:

	$Y=5$	$Y=6$	$p_X(x)$
$X=2$	0.0	0.1	0.1
$X=3$	0.1	0.2	0.3
$X=4$	0.2	0.4	0.6
$p_Y(y)$	0.3	0.7	

(Meaning that $p_{X,Y}(2,5) = 0.0$, $p_{X,Y}(3,5) = 0.1$, $p_{X,Y}(4,6) = 0.4$, etc.)

(Marginals $p_X(x)$ and $p_Y(y)$ are also shown, found by summing.)

→ Then we can compute e.g. $\mathrm{P}(Y=5\,|\,X=3) = \frac{\mathrm{P}(X=3,\ Y=5)}{\mathrm{P}(X=3)} = \frac{0.1}{0.3} = 1/3$.

→ Similarly $\mathrm{P}(Y=6\,|\,X=3) = \frac{\mathrm{P}(X=3,\ Y=6)}{\mathrm{P}(X=3)} = \frac{0.2}{0.3} = 2/3$.

→ Can write this as $p_{Y|X}(5\,|\,3) = 1/3$, $p_{Y|X}(6\,|\,3) = 2/3$, otherwise $p_{Y|X}(x\,|\,3) = 0$.

→ So, $p_{Y|X}(\cdot\,|\,3)$ is a proper probability function (≥ 0, and sums to 1): the conditional distribution of Y given that $X = 3$.

→ Also, $\mathrm{P}(X=2\,|\,Y=6) = \frac{\mathrm{P}(X=2,\ Y=6)}{\mathrm{P}(Y=6)} = \frac{0.1}{0.7} = 1/7$, and $\mathrm{P}(X=3\,|\,Y=6) = 2/7$, and $\mathrm{P}(X=4\,|\,Y=6) = 4/7$. So, $p_{X|Y}(2\,|\,6) = 1/7$, $p_{X|Y}(3\,|\,6) = 2/7$, $p_{X|Y}(4\,|\,6) = 4/7$, the conditional distribution of X given that $Y = 6$.

→ Exercise: Find $p_{X\,|\,Y}(x\,|\,5)$ for all $x \in \mathbf{R}$, i.e. the conditional distribution of X given that $Y = 5$.

- In general, $p_{X|Y}(x\,|\,y) = \frac{\mathrm{P}(X=x,\ Y=y)}{\mathrm{P}(Y=y)}$, and $p_{Y|X}(y\,|\,x) = \frac{\mathrm{P}(X=x,\ Y=y)}{\mathrm{P}(X=x)}$.

→ Then e.g. $\mathrm{P}(a \leq Y \leq b\,|\,X=x) = \sum_{a\leq y\leq b} \mathrm{P}(Y=y\,|\,X=x) = \sum_{a\leq y\leq b} p_{Y|X}(y|x) = \sum_{a\leq y\leq b} \frac{p_{X,Y}(x,y)}{p_X(x)} = \frac{\mathrm{P}(a\leq Y\leq b,\, X=x)}{\mathrm{P}(X=x)}$, as it should.

- Definition: Two random variables X and Y are independent if the events $\{X \in B\}$ and $\{Y \in C\}$ are independent for all subsets $B, C \subseteq \mathbf{R}$, i.e. if we always have $\mathrm{P}(X \in B,\ Y \in C) = \mathrm{P}(X \in B)\ \mathrm{P}(Y \in C)$.

→ For example, if we take $B = (-\infty, x]$ and $C = (-\infty, y]$, this means that $\mathrm{P}(X \leq x,\ Y \leq y) = \mathrm{P}(X \leq x)\ \mathrm{P}(Y \leq y)$, i.e. $F_{X,Y}(x,y) = F_X(x)\,F_Y(y)$ for all $x, y \in \mathbf{R}$. (Equivalent definition.)

→ For discrete random variables X and Y, it suffices that the events $\{X = x\}$ and $\{Y = y\}$ are independent, i.e. $\mathrm{P}(X = x,\ Y = y) = \mathrm{P}(X = x)\ \mathrm{P}(Y = y)$, i.e. $p_{X,Y}(x,y) = p_X(x)\,p_Y(y)$ for all $x, y \in \mathbf{R}$.

→ Then for any B and C, we have $\mathrm{P}(X \in B,\ Y \in C) = \sum_{x\in B}\sum_{y\in C} p_{X,Y}(x,y) = \sum_{x\in B}\sum_{y\in C} p_X(x)\,p_Y(y) = \left(\sum_{x\in B} p_X(x)\right)\left(\sum_{y\in C} p_Y(y)\right) = \mathrm{P}(X \in B)\ \mathrm{P}(Y \in C)$.

- If X and Y are discrete and independent, then $p_{X\,|\,Y}(x\,|\,y) = \frac{\mathrm{P}(X=x,\ Y=y)}{\mathrm{P}(Y=y)} = \frac{\mathrm{P}(X=x)\ \mathrm{P}(Y=y)}{\mathrm{P}(Y=y)} = \mathrm{P}(X=x)$, and similarly $p_{Y\,|\,X}(y\,|\,x) = \mathrm{P}(Y=y)$.

→ This means the values of Y do not affect the probabilities for X.

→ In above example, X and Y are not independent, since e.g. $p_{X,Y}(3,5) = 0.1$ but $p_X(3)\,p_Y(5) = (0.3)(0.3) = 0.09 \neq 0.1$.

Suggested Homework: 2.8.1, 2.8.2, 2.8.5, 2.8.9, 2.8.10, 2.8.12, 2.8.13, 2.8.20.

Conditioning and Independence for Continuous Random Variables

- Suppose X and Y have joint density function $f_{X,Y}(x,y)$. Conditionals?
- Does $\mathrm{P}(a \leq Y \leq b\,|\,X = x)$ even make sense?

→ No, since $\mathrm{P}(X = x) = 0$, so we can't divide by it.

→ Trick: Do it anyway!

• Intuitively, imagine replacing the event $\{X = x\}$ by the event $\{x \le X \le x+\epsilon\}$ for small $\epsilon > 0$, so that $\mathrm{P}(x \le X \le x+\epsilon) > 0$. In fact, $\mathrm{P}(x \le X \le x+\epsilon) = \int_x^{x+\epsilon} f_X(u)\,du$.

→ If f_X is continuous at x, and $\epsilon > 0$ is small, then $\mathrm{P}(x \le X \le x+\epsilon) \approx \epsilon\, f_X(x)$.

→ ["First-order approximation": formally, $\lim_{\epsilon \searrow 0} \frac{1}{\epsilon} \int_x^{x+\epsilon} f_X(u)\,du = f_X(x)$.]

→ But also, if $f_{X,Y}$ is continuous at (x, y) for $a \le y \le b$, then $\mathrm{P}(x \le X \le x+\epsilon,\ a \le Y \le b) = \int_a^b \int_x^{x+\epsilon} f_{X,Y}(u, y)\,du\,dy \approx \epsilon \int_a^b f_{X,Y}(x, y)\,dy$.

→ So, $\mathrm{P}(a \le Y \le b \,|\, x \le X \le x+\epsilon) \approx \frac{\epsilon \int_a^b f_{X,Y}(x,y)\,dy}{\epsilon\, f_X(x)} = \int_a^b \frac{f_{X,Y}(x,y)}{f_X(x)}\,dy$.

• Therefore, we define the conditional density of Y given that $X = x$, to be the density function $f_{Y|X}(y \,|\, x) = \frac{f_{X,Y}(x,y)}{f_X(x)}$, valid whenever $f_X(x) > 0$.

→ Then we say that $\mathrm{P}(a \le Y \le b \,|\, X = x) = \int_a^b f_{Y|X}(y \,|\, x)\,dy := \int_a^b \frac{f_{X,Y}(x,y)}{f_X(x)}\,dy$.

• Definition: X and Y are independent if $f_{X,Y}(x, y) = f_X(x)\, f_Y(y)$ for "all" $x, y \in \mathbf{R}$, or equivalently if $f_{Y|X}(y \,|\, x) = f_Y(y)$ whenever $f_X(x) > 0$.

→ Then for any B and C, we have $\mathrm{P}(X \in B,\ Y \in C) = \int_{y \in C} \int_{x \in B} f_{X,Y}(x, y)\,dx\,dy = \int_{y \in C} \int_{x \in B} f_X(x)\, f_Y(y)\,dx\,dy = \left(\int_{x \in B} f_X(x)\,dx\right)\left(\int_{y \in C} f_Y(y)\,dy\right) = \mathrm{P}(X \in B)\,\mathrm{P}(Y \in C)$.

• Previous running example: $f_{X,Y}(x, y) = \frac{15}{32} x y^2$ for $0 \le y \le x \le 2$, otherwise 0.

→ Found that $f_X(x) = (5/32)x^4$ for $0 \le x \le 2$, otherwise 0.

→ And that $f_Y(y) = \frac{15}{64}(4y^2 - y^4)$ for $0 \le y \le 2$, otherwise 0.

→ Hence, for $0 \le y \le x \le 2$, we have $f_{Y|X}(y \,|\, x) = \frac{f_{X,Y}(x,y)}{f_X(x)} = \frac{\frac{15}{32} x y^2}{(5/32)x^4} = 3x^{-3}y^2$.

→ So e.g. $\mathrm{P}(0 \le Y \le 1 \,|\, X = 3/2) = \int_0^1 f_{Y|X}(y \,|\, 3/2)\,dy = \int_0^1 (3(3/2)^{-3}y^2)\,dy = 3(3/2)^{-3}\frac{1}{3}(1^3 - 0^3) = (3/2)^{-3} = 8/27$.

→ Also $\mathrm{P}(0 \le Y \le 3/2 \,|\, X = 3/2) = \int_0^{3/2} f_{Y|X}(y \,|\, 3/2)\,dy = \int_0^{3/2} (3(3/2)^{-3}y^2)\,dy = 3(3/2)^{-3}\frac{1}{3}((3/2)^3 - 0^3) = (3/2)^{-3}(3/2)^3 = 1$. Makes sense since here $0 \le Y \le X$.

→ Here $f_{X,Y}(x, y) \ne f_X(x)\, f_Y(y)$, and $f_{Y|X}(y \,|\, x) \ne f_Y(y)$, so not independent.

• Summary: X and Y are independent if and only if any one of:

→ $\mathrm{P}(X \in B,\ Y \in C) = \mathrm{P}(X \in B)\ \mathrm{P}(Y \in C)$ for all $B, C \subseteq \mathbf{R}$. (general)

→ $F_{X,Y}(x, y) = F_X(x)\, F_Y(y)$ for all $x, y \in \mathbf{R}$. (general)

→ $p_{X,Y}(x, y) = p_X(x)\, p_Y(y)$ for all $x, y \in \mathbf{R}$. (discrete)

→ $p_{Y|X}(y \,|\, x) = p_Y(y)$ for "all" $x, y \in \mathbf{R}$, or vice-versa. (discrete)

→ $f_{X,Y}(x, y) = f_X(x)\, f_Y(y)$ for "all" $x, y \in \mathbf{R}$. (abs. continuous)

→ $f_{Y|X}(y \,|\, x) = f_Y(y)$ for "all" $x, y \in \mathbf{R}$, or vice-versa. (abs. continuous)

Suggested Homework: 2.8.3, 2.8.4, 2.8.7, 2.8.8, 2.8.14, 2.8.15, 2.8.17.

END WEDNESDAY #7

Multivariable Change-Of-Variable – Discrete

• Suppose X and Y are discrete, with joint probability function $p_{X,Y}(x, y)$.

• Suppose $(Z, W) = h(X, Y)$, for some function $h : \mathbf{R}^2 \to \mathbf{R}^2$.

- Then what is $p_{Z,W}(z,w) := \mathrm{P}(Z=z, W=w)$?
- By the Law of Total Probability,

$$p_{Z,W}(z,w) = \mathrm{P}(h(X,Y)=(z,w)) = \sum \{p_{X,Y}(x,y) : h(x,y)=(z,w)\}\,.$$

- Similar to one-variable case. Not difficult.

Suggested Homework: 2.9.6, 2.9.9.

Multivariable Change-Of-Variable – Continuous

- Recall one-variable case: If $Y = h(X)$, where $h : \mathbf{R} \to \mathbf{R}$ is differentiable and strictly increasing or decreasing, then $f_Y(y) = f_X\big(h^{-1}(y)\big) \,/\, \big|h'\big(h^{-1}(y)\big)\big|$.
- Two-variable version? Trickier!
  - → Now $(Z,W) = h(X,Y)$, where $h : \mathbf{R}^2 \to \mathbf{R}^2$.
  - → i.e., $Z = h_1(X,Y)$ and $W = h_2(X,Y)$.
  - → Need h to be (two-dimensional) differentiable, and one-to-one (invertible).
  - → Then $f_{Z,W}(z,w) = f_{X,Y}\big(h^{-1}(z,w)\big) \,/\, \big|J_h\big(h^{-1}(z,w)\big)\big|$.
  - → Here J_h is the Jacobian determinant: $J_h(x,y) = \det\begin{pmatrix} \frac{\partial h_1}{\partial x} & \frac{\partial h_1}{\partial y} \\ \frac{\partial h_2}{\partial x} & \frac{\partial h_2}{\partial y} \end{pmatrix}$.
  - → See e.g. Textbook's Example 2.9.2 and Example 2.9.3 (page 111).
- e.g. Let U and V be independent Uniform[0,1].
  - → Then let $Z = \sqrt{2\log(1/U)}\,\cos(2\pi V)$ and $W = \sqrt{2\log(1/U)}\,\sin(2\pi V)$.
  - → What are the distributions of Z and W?
  - → Fact (textbook pp. 111–112): Z and W are independent, and are both ... Normal(0,1)!! This is important! Best way to simulate normal random variables.

Suggested Homework: 2.9.2, 2.9.3, 2.9.4, 2.9.5, 2.9.11.

- Note: We are omitting a few related topics from the end of Chapter 2, e.g.:
  - → Order Statistics (when you sort the sample values, from smallest to largest).
  - → Simulating probability distributions on a computer: algorithms.
  - → All interesting! Check them out! Try the exercises! Ask me questions!

[END OF TEXTBOOK CHAPTER #2]

Expected Values: Discrete Case

- Intuitively, the expected or average or mean value of a random variable is what it equals "on average".
  - → e.g. If $\mathrm{P}(X=0) = \mathrm{P}(X=12) = 1/2$, then $\mathrm{E}(X) = 6$, the average value.
  - → e.g. If $\mathrm{P}(X=0) = 2/3$ and $\mathrm{P}(X=12) = 1/3$, then $\mathrm{E}(X) = 4$: weighted av.

• Definition: If X is a discrete random variable, then its expected value is given by $\mathrm{E}(X) = \sum_{x\in\mathbf{R}} x\ \mathrm{P}(X=x) = \sum_{x\in\mathbf{R}} x\ p_X(x)$. (Also sometimes written as μ_X.)

→ If $\mathrm{P}(X=x_i) = p_i$ where $p_i \geq 0$ and $\sum_i p_i = 1$, then $\mathrm{E}(X) = \sum_i x_i\, p_i$.

• e.g. If $\mathrm{P}(X=0) = \mathrm{P}(X=12) = 1/2$, $\mathrm{E}(X) = 0(1/2) + 12(1/2) = 6$.

→ Or, if $\mathrm{P}(X=0) = 2/3$ and $\mathrm{P}(X=12) = 1/3$, $\mathrm{E}(X) = 0(2/3) + 12(1/3) = 4$.

→ Or, if $X = c$ is constant, i.e. $\mathrm{P}(X=c) = 1$, then $\mathrm{E}(X) = c(1) = c$.

• e.g. If X is the number showing on a fair six-sided die, then $\mathrm{E}(X) = \sum_{x\in\mathbf{R}} x\,\mathrm{P}(X = x) = \sum_{k=1}^{6} k\,(1/6) = (1+2+3+4+5+6)/6 = 21/6 = 3.5$. (Not 3!)

• e.g. If $X \sim \text{Bernoulli}(\theta)$, then $\mathrm{E}(X) = 0(1-\theta) + 1(\theta) = \theta$.

• e.g. Suppose $Y \sim \text{Binomial}(n,\theta)$. What is $\mathrm{E}(Y)$?

→ Well, $\mathrm{E}(Y) = \sum_{y\in\mathbf{R}} y\,\mathrm{P}(Y=y) = \sum_{k=0}^{n} k\binom{n}{k}\theta^k(1-\theta)^{n-k} = \sum_{k=0}^{n} k\frac{n!}{(n-k)!\,k!}\theta^k(1-\theta)^{n-k} = \sum_{k=1}^{n} n\frac{(n-1)!}{(n-k)!\,(k-1)!}\theta^k(1-\theta)^{n-k} = n\theta\sum_{k=1}^{n}\binom{n-1}{k-1}\theta^{k-1}(1-\theta)^{n-k}$.

→ Now, set $j = k-1$, and use the Binomial Theorem again: $\mathrm{E}(Y) = n\theta\sum_{j=0}^{n-1}\binom{n-1}{j}\theta^j(1-\theta)^{n-1-j} = n\theta[\theta+(1-\theta)]^{n-1} = n\theta$. Easier way?

→ e.g. Shoot $n=10$ free throws, prob $\theta = 1/4$ on each: $\mathrm{E}(\#\text{ successes}) = 2.5$.

• e.g. If $Z \sim \text{Geometric}(\theta)$, then $\mathrm{E}(Z) = \sum_{z\in\mathbf{R}} z\,\mathrm{P}(Z=z) = \sum_{k=0}^{\infty} k\,(1-\theta)^k\,\theta$ =??

→ Trick: Here $(1-\theta)\,\mathrm{E}(Z) = \sum_{k=0}^{\infty} k\,(1-\theta)^{k+1}\,\theta = \sum_{\ell=0}^{\infty}\ell\,(1-\theta)^{\ell+1}\,\theta$.

→ Letting $k = \ell+1$, this equals $\sum_{k=1}^{\infty}(k-1)\,(1-\theta)^k\,\theta$.

→ Hence, $\mathrm{E}(Z) - (1-\theta)\,\mathrm{E}(Z) = \sum_{k=1}^{\infty}(1)\,(1-\theta)^k\,\theta = \frac{1-\theta}{1-(1-\theta)}\,\theta = 1-\theta$.

→ But $\mathrm{E}(Z) - (1-\theta)\,\mathrm{E}(Z) = \theta\,\mathrm{E}(Z)$. Hence, $\mathrm{E}(Z) = \frac{1-\theta}{\theta}$. Phew!

→ e.g. if $\theta = 1/2$ then $\mathrm{E}(Z) = 1$, but if $\theta = 1/5$ then $\mathrm{E}(Z) = 4$.

• e.g. If $X \sim \text{Poisson}(\lambda)$, then $\mathrm{E}(X) = \sum_{x\in\mathbf{R}} x\,\mathrm{P}(X=x) = \sum_{k=0}^{\infty} k\,e^{-\lambda}\lambda^k/k! = e^{-\lambda}\lambda\,\left[\sum_{k=1}^{\infty}\lambda^{k-1}/(k-1)!\right] = e^{-\lambda}\lambda\,\left[\sum_{\ell=0}^{\infty}\lambda^{\ell}/\ell!\right] = e^{-\lambda}\lambda\,[e^{\lambda}] = \lambda$.

• e.g. Suppose $\mathrm{P}(X=2) = 1/2$, $\mathrm{P}(X=4) = 1/4$, $\mathrm{P}(X=8) = 1/8$, and in general $\mathrm{P}(X=2^k) = 2^{-k}$ for $k = 1,2,3,\ldots$.

→ Then $\mathrm{E}(X) = \sum_{k=1}^{\infty}(2^k)(2^{-k}) = \sum_{k=1}^{\infty}(1) = \infty$.

→ So, $\mathrm{E}(X) = \infty$, even though $\mathrm{P}(X < \infty) = 1$. Infinite expectation!

• Can also sum to get expectations of functions of discrete random variables:

→ If $Z = g(X)$, then $\mathrm{E}(Z) = \mathrm{E}(g(X)) = \sum_{z\in\mathbf{R}} z\,\mathrm{P}(Z=z) = \sum_{x\in\mathbf{R}} g(x)\,\mathrm{P}(X=x)$.

→ Or, if $Z = h(X,Y)$, $\mathrm{E}(Z) = \sum_{z\in\mathbf{R}} z\,\mathrm{P}(Z=z) = \sum_{x,y\in\mathbf{R}} h(x,y)\,\mathrm{P}(X=x, Y=y)$.

→ (Here Z is also discrete; and get the same expected value either way.)

• e.g. if $X \sim \text{Binomial}(3, 1/4)$, then know $\mathrm{E}(X) = 3(1/4) = 3/4$, but also $\mathrm{E}(5X^2) = \sum_{x\in\mathbf{R}} 5x^2\,\mathrm{P}(X=x) = \sum_{k=0}^{3} 5k^2\binom{3}{k}(1/4)^k(3/4)^{3-k}$
$= 5(0)^2\binom{3}{0}(1/4)^0(3/4)^3 + 5(1)^2\binom{3}{1}(1/4)^1(3/4)^2 + 5(2)^2\binom{3}{2}(1/4)^2(3/4)^1 + 5(3)^2\binom{3}{3}(1/4)^3(3/4)^0$
$= 0 + 5\cdot 1\cdot 3\cdot 3^2/4^3 + 5\cdot 4\cdot 3\cdot 3/4^3 + 5\cdot 9\cdot 1\cdot 1/4^3 = 45/8 = 5.625$.

Suggested Homework: 3.1.1, 3.1.2, 3.1.3, 3.1.8, 3.1.9, 3.1.10, 3.1.14.

• If $Z = aX + bY$, where $a, b \in \mathbf{R}$, and X and Y are discrete random variables,
$\mathrm{E}(Z) = \sum_{z\in\mathbf{R}} z\,\mathrm{P}(Z=z) = \sum_{x,y\in\mathbf{R}} (ax+by)\,\mathrm{P}(X=x, Y=y)$
$= a\sum_{x,y\in\mathbf{R}} x\,\mathrm{P}(X=x, Y=y) + b\sum_{x,y\in\mathbf{R}} y\,\mathrm{P}(X=x, Y=y)$
$= a\sum_{x\in\mathbf{R}} x\sum_{y\in\mathbf{R}} \mathrm{P}(X=x, Y=y) + b\sum_{y\in\mathbf{R}} y\sum_{x\in\mathbf{R}} \mathrm{P}(X=x, Y=y)$
$= a\sum_{x\in\mathbf{R}} x\,\mathrm{P}(X=x) + b\sum_{y\in\mathbf{R}} y\,\mathrm{P}(Y=y) = a\,\mathrm{E}(X) + b\,\mathrm{E}(Y)$. Linear property.

• If $Y \sim \text{Binomial}(n, \theta)$, then we can think of Y as $Y = X_1 + X_2 + \ldots + X_n$ where each $X_i \sim \text{Bernoulli}(\theta)$. (e.g. $X_i = 1$ if you score on the i^{th} free throw, otherwise 0)

→ By linearity, $\mathrm{E}(Y) = \mathrm{E}(X_1) + \mathrm{E}(X_2) + \ldots + \mathrm{E}(X_n) = \theta + \theta + \ldots + \theta = n\theta$.

→ Same answer as before! Easier!

END MONDAY #8

• e.g. Suppose $X \sim \text{Binomial}(5, 1/4)$, and $Y \sim \text{Geometric}(1/3)$, and $Z = 2X - 6Y$.

→ Then from linearity and the above calculations, $\mathrm{E}(Z) = \mathrm{E}(2X-6Y) = 2\,\mathrm{E}(X) - 6\,\mathrm{E}(Y) = 2[(5)(1/4)] - 6[\frac{2/3}{1/3}] = -19/2 = -9.5$.

• Caution: This is only for linear functions! e.g. If $X \sim \text{Bernoulli}(1/2)$, then $\mathrm{E}(X^2) = \mathrm{E}(X) = 1/2$, which is not the same as $\big(\mathrm{E}(X)\big)^2 = (1/2)^2 = 1/4$.

• Suppose X and Y are discrete, and $X \le Y$, i.e. $X(s) \le Y(s)$ for all $s \in S$.

→ Or more generally, suppose that $\mathrm{P}(X \le Y) = 1$.

→ Let $Z = Y - X$. Then Z is discrete, and $\mathrm{P}(Z \ge 0) = 1$.

→ So, $\mathrm{P}(Z = z) = 0$ whenever $z < 0$.

→ Hence, $\mathrm{E}(Z) = \sum_{z\in\mathbf{R}} z\,\mathrm{P}(Z=z) = \sum_{z\in[0,\infty)} z\,\mathrm{P}(Z=z) \ge 0$.

→ But $\mathrm{E}(Z) = \mathrm{E}(Y-X) = \mathrm{E}(Y) - \mathrm{E}(X)$, so $\mathrm{E}(Y) - \mathrm{E}(X) \ge 0$, i.e. $\mathrm{E}(X) \le \mathrm{E}(Y)$.

→ This is the monotonicity property: If $\mathrm{P}(X \le Y) = 1$, then $\mathrm{E}(X) \le \mathrm{E}(Y)$.

Suggested Homework: 3.1.4, 3.1.5, 3.1.11(a), 3.1.15, 3.1.16.

• Also, expectation preserves products of independent random variables:

→ Suppose X and Y are discrete random variables which are independent.

→ Then $\mathrm{E}(XY) = \sum_{x,y\in\mathbf{R}} xy\,\mathrm{P}(X=x,\, Y=y) = \sum_{x,y\in\mathbf{R}} xy\,\mathrm{P}(X=x)\,\mathrm{P}(Y=y) = \Big(\sum_{x\in\mathbf{R}} x\,\mathrm{P}(X=x)\Big)\Big(\sum_{y\in\mathbf{R}} y\,\mathrm{P}(Y=y)\Big) = \mathrm{E}(X)\;\mathrm{E}(Y)$. Useful!

• e.g. Suppose $X \sim \text{Binomial}(5, 1/4)$, and $Y \sim \text{Geometric}(1/3)$, and X and Y are independent, and $Z = XY$.

→ Then $\mathrm{E}(Z) = \mathrm{E}(XY) = \mathrm{E}(X)\,\mathrm{E}(Y) = [(5)(1/4)]\;[\frac{2/3}{1/3}] = 10/4 = 2.5$.

• e.g. Suppose $X \sim \text{Bernoulli}(1/2)$ and $Y = X$, and let $Z = XY$.

→ Then $\mathrm{E}(X) = 1/2$, and $\mathrm{E}(Y) = 1/2$, and $\mathrm{E}(Z) = \mathrm{E}(XY) = \mathrm{E}(X^2) = 1/2$.

→ So $\mathrm{E}(XY) \ne \mathrm{E}(X)\,\mathrm{E}(Y)$. Why not? Because X and Y are not independent!

Suggested Homework: 3.1.11(b), 3.1.12, 3.1.17, 3.1.20.

Expected Values: Absolutely Continuous Case

- If X is continuous, then $\mathrm{P}(X = x) = 0$, so $\sum_{x \in \mathbf{R}} x\,\mathrm{P}(X = x) = 0$. Useless!
  - → Can we still "add up" the values times their probabilities?
  - → Yes, by integrating instead of summing!
- Definition: If X is an absolutely continuous random variable, then its expected value is given by the integral $\mathrm{E}(X) = \int_{-\infty}^{\infty} x\ f_X(x)\,dx$. (Sometimes written as μ_X.)
  - → Intuitively, we are adding up values times little "bits" of probability.
- e.g. If $X \sim \mathrm{Uniform}[0,1]$, then what is $\mathrm{E}(X)$? We compute that:
$\mathrm{E}(X) = \int_{-\infty}^{\infty} x\,f_X(x)\,dx = \int_0^1 x\,(1)\,dx = \frac{1}{2}x^2\Big|_{x=0}^{x=1} = \frac{1}{2}(1^2 - 0^2) = \frac{1}{2}$.
- e.g. If $X \sim \mathrm{Uniform}[L,R]$, then what is $\mathrm{E}(X)$? We compute that:
$\mathrm{E}(X) = \int_{-\infty}^{\infty} x\ f_X(x)\,dx = \int_L^R x\ (\frac{1}{R-L})\,dx = \frac{1}{2}x^2\ (\frac{1}{R-L})\Big|_{x=L}^{x=R} = \frac{1}{2}\ (\frac{1}{R-L})(R^2 - L^2) =$
$\frac{1}{2}\ (\frac{1}{R-L})(R-L)(R+L) = \frac{1}{2}(R+L)$.
  - → e.g. If $X \sim \mathrm{Uniform}[-8,2]$, then $\mathrm{E}(X) = \frac{1}{2}(-8+2) = -3$. Negative!
- If $Y \sim \mathrm{Exponential}(\lambda)$, then $\mathrm{E}(Y) = \int_{-\infty}^{\infty} y\ f_Y(y)\,dy = \int_0^{\infty} y\ \lambda e^{-\lambda y}\,dy = ??$
  - → Need to use "integration by parts"!
  - → Set $u(y) = y$ and $v(y) = -e^{-\lambda y}$, then $du = dy$ and $dv = \lambda e^{-\lambda y}\,dy$.
  - → Then $\mathrm{E}(Y) = \int_0^{\infty} u\,dv = u(y)v(y)\Big|_{y=0}^{y=\infty} - \int_0^{\infty} du\,v = -ye^{-\lambda y}\Big|_{y=0}^{y=\infty} - \int_0^{\infty} dy\,(-e^{-\lambda y}) =$
$-0 + 0 + \int_0^{\infty} e^{-\lambda y}\,dy = -\frac{1}{\lambda}e^{-\lambda y}\Big|_{y=0}^{y=\infty} = -\frac{1}{\lambda}(0-1) = \frac{1}{\lambda}$. (Not λ.)
- If $Z \sim \mathrm{Normal}(0,1)$, then $\mathrm{E}(Z) = \int_{-\infty}^{\infty} z\ \phi(z)\,dz = \int_{-\infty}^{\infty} z\ \frac{1}{\sqrt{2\pi}}\,e^{-z^2/2}\,dz = ??$
  - → The integrand is an "odd" function, so by symmetry, $\mathrm{E}(Z) = 0$.
- Now suppose $W \sim \mathrm{Normal}(\mu, \sigma^2)$. Then what is $\mathrm{E}(W)$?
  - → Well, this means that $W = \mu + \sigma Z$ where $Z \sim \mathrm{Normal}(0,1)$.
  - → So, maybe $\mathrm{E}(W) = \mathrm{E}(\mu + \sigma Z) = \mu + \sigma\,\mathrm{E}(Z) = \mu + 0 = \mu$? Yes, because ...
- Expectation still satisfies the same general properties as for discrete r.v.:
- Can still calculate expectations of functions of abs. cont. random variables:
  - → If $Z = g(X)$, then $\mathrm{E}(Z) = \mathrm{E}(g(X)) = \int_{-\infty}^{\infty} g(x)\,f_X(x)\,dx$.
  - → Or, if $Z = h(X,Y)$, then $\mathrm{E}(Z) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(x,y)\,f_{X,Y}(x,y)\,dx\,dy$.
  - → (If Z is abs. cont. or discrete, then get the same expected value either way.)
- Expectation is still linear! Let $Z = aX + bY$, where $a, b \in \mathbf{R}$, and X and Y are jointly absolutely continuous random variables. Then:
$\mathrm{E}(Z) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(ax+by)\,f_{X,Y}(x,y)\,dx\,dy$
$= a\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x\,f_{X,Y}(x,y)\,dx\,dy + b\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} y\,f_{X,Y}(x,y)\,dx\,dy$
$= a\int_{-\infty}^{\infty} x\,\left(\int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dy\right)\,dx + b\int_{-\infty}^{\infty} y\,\left(\int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dx\right)\,dy$
$= a\int_{-\infty}^{\infty} x\,f_X(x)\,dx + b\int_{-\infty}^{\infty} y\,f_Y(y)\,dy = a\,\mathrm{E}(X) + b\,\mathrm{E}(Y)$.
- And, still monotone: If $\mathrm{P}(X \le Y) = 1$, and $Z = Y - X$, then $f_Z(z) = 0$ whenever $z < 0$, so $\mathrm{E}(Z) = \int_0^{\infty} z\,f_Z(z)\,dz \ge 0$, so $\mathrm{E}(Y - X) \ge 0$, so $\mathrm{E}(X) \le \mathrm{E}(Y)$.

- And, still preserves products of independent random variables:

→ Assume X and Y are jointly absolutely continuous, and independent.

→ Then $\mathrm{E}(XY) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x\,y\,f_{X,Y}(x,y)\,dx\,dy = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x\,y\,f_X(x)\,f_Y(y)\,dx\,dy = \left(\int_{-\infty}^{\infty} x\,f_X(x)\,dx\right)\left(\int_{-\infty}^{\infty} y\,f_Y(y)\,dy\right) = \mathrm{E}(X)\ \mathrm{E}(Y)$.

Suggested Homework: 3.2.1, 3.2.2, 3.2.3, 3.2.4, 3.2.5, 3.2.6, 3.2.7, 3.2.9, 3.2.10, 3.2.12, 3.2.14, 3.2.15.

Variance and Standard Deviation

- Suppose X has expected value $\mathrm{E}(X)$, or μ_X. Does that tell us everything?
- e.g. $X_1 \sim \mathrm{Uniform}[4.9, 5.1]$, $X_2 \sim \mathrm{Uniform}[4, 6]$, $X_3 \sim \mathrm{Uniform}[0, 10]$.

→ Then $\mathrm{E}(X_1) = 5$, and $\mathrm{E}(X_2) = 5$, and $\mathrm{E}(X_3) = 5$. All the same.

→ But X_1 is always very close to 5, while X_3 can be quite far away. (X_2 medium.)

- The variance of any random variable X is $\mathrm{Var}(X) := \mathrm{E}[(X - \mu_X)^2]$.

→ A measure of how far X usually is from μ_X.

→ Why not $\mathrm{E}(X - \mu_X)$? Always zero! Useless!

→ Why not $\mathrm{E}(|X - \mu_X|)$? That turns out to be less convenient ...

- So, we'll stick with $\mathrm{Var}(X) := \mathrm{E}[(X - \mu_X)^2]$.

→ But $\mathrm{Var}(X)$ has "squared units" (e.g. if X in meters (m), then $\mathrm{Var}(X)$ is in meters-squared (m^2)). This can be awkward.

→ So, often use the standard deviation, $\mathrm{Sd}(X) := \sqrt{\mathrm{Var}(X)} = \sqrt{\mathrm{E}[(X - \mu_X)^2]}$.

- e.g. $X \sim \mathrm{Bernoulli}(\theta)$. Then $\mu_X = \theta$, so $\mathrm{Var}(X) = \mathrm{E}[(X - \theta)^2] = (0 - \theta)^2(1 - \theta) + (1 - \theta)^2(\theta) = -\theta^2 + \theta^3 + \theta - \theta^3 = -\theta^2 + \theta = \theta(1 - \theta)$.
- By linearity, we always have $\mathrm{Var}(X) := \mathrm{E}[(X - \mu_X)^2] = \mathrm{E}[X^2 - 2X\mu_X + (\mu_X)^2] = \mathrm{E}[X^2] - 2\mathrm{E}[X]\mu_X + (\mu_X)^2 = \mathrm{E}[X^2] - (\mu_X)^2$.

→ So, if $X \sim \mathrm{Bernoulli}(\theta)$, then could instead compute $\mathrm{Var}(X)$ by: $\mathrm{Var}(X) = \mathrm{E}[X^2] - (\mu_X)^2 = 0^2(1 - \theta) + 1^2(\theta) - (\theta)^2 = \theta - \theta^2 = \theta(1 - \theta)$. Easier?

- Suppose $Y \sim \mathrm{Uniform}[0, 1]$. Know $\mu_Y = 1/2$.

→ And, $\mathrm{E}(Y^2) = \int_{-\infty}^{\infty} y^2\,f_Y(y)\,dy = \int_0^1 y^2\,(1)\,dy = \frac{1}{3}y^3\Big|_{y=0}^{y=1} = \frac{1}{3}(1^3 - 0^3) = \frac{1}{3}$.

→ Hence, $\mathrm{Var}(Y) = \mathrm{E}(Y^2) - (\mu_Y)^2 = (1/3) - (1/2)^2 = (1/3) - (1/4) = 1/12$.

→ So then $\mathrm{Sd}(Y) = \sqrt{\mathrm{Var}(Y)} = \sqrt{1/12} = 1/\sqrt{12} = 1/(2\sqrt{3})$.

- Suppose $Z \sim \mathrm{Uniform}[L, R]$ (where $L < R$). Know that $\mu_Z = (L + R)/2$.

→ And, $\mathrm{E}(Z^2) = \int_{-\infty}^{\infty} z^2\,f_Z(z)\,dz = \int_L^R z^2\,\frac{1}{R-L}\,dz = \frac{1}{3(R-L)}z^3\Big|_{z=L}^{z=R} = \frac{1}{3(R-L)}(R^3 - L^3) = \frac{1}{3(R-L)}(R - L)(R^2 + RL + L^2) = \frac{1}{3}(R^2 + RL + L^2)$.

→ Hence, $\mathrm{Var}(Z) = \mathrm{E}(Z^2) - (\mu_Z)^2 = \frac{1}{3}(R^2 + RL + L^2) - (\frac{L+R}{2})^2$.

→ After a bit of algebra (exercise!), this works out to ... $(R - L)^2/12$.

→ So then $\mathrm{Sd}(Z) = \sqrt{\mathrm{Var}(Z)} = (R - L)/\sqrt{12}$.

• e.g. if $X_1 \sim \text{Uniform}[4.9, 5.1]$, $X_2 \sim \text{Uniform}[4, 6]$, and $X_3 \sim \text{Uniform}[0, 10]$, then: $\text{Var}(X_1) = (0.2)^2/12 \doteq 0.0033$, $\text{Var}(X_2) = (1)^2/12 = 1/12 \doteq 0.083$, and $\text{Var}(X_3) = (10)^2/12 = 100/12 \doteq 8.33$. So $\text{Var}(X_3) \gg \text{Var}(X_2) \gg \text{Var}(X_1)$, which makes sense.

• In general, $(X - \mu_X)^2 \geq 0$, so always have $\text{Var}(X) := \text{E}[(X - \mu_X)^2] \geq 0$.

→ But $\text{Var}(X) = \text{E}[X^2] - (\mu_X)^2$, so $\text{E}[X^2] - (\mu_X)^2 \geq 0$, i.e. $\text{E}[X^2] \geq (\mu_X)^2$.

→ And, since $(\mu_X)^2 \geq 0$, always have $\text{Var}(X) = \text{E}[X^2] - (\mu_X)^2 \leq \text{E}[X^2]$, too.

• If $a, b \in \mathbf{R}$, then $\text{Var}(aX + b) = \text{E}[(aX + b - \mu_{aX+b})^2] = \text{E}[(aX + b - a\mu_X - b)^2] = \text{E}[(a(X - \mu_X))^2] = a^2\,\text{E}[(X - \mu_X)^2] = a^2\,\text{Var}(X)$. (Note: a^2, not a. And b irrelevant.)

→ Hence, $\text{Sd}(aX + b) = \sqrt{\text{Var}(aX + b)} = \sqrt{a^2\,\text{Var}(X)} = |a|\,\text{Sd}(X)$.

→ What about $\text{Var}(X + Y)$ or $\text{Var}(aX + bY)$? Later!

• e.g. $W \sim \text{Exponential}(\lambda)$. Know $\mu_W := \text{E}(W) = 1/\lambda$. $\text{Var}(W) = ??$

→ Well, $\text{E}(W^2) = \int_{-\infty}^{\infty} w^2\, f_W(w)\, dw = \int_0^{\infty} w^2\, \lambda e^{-\lambda w}\, dw$.

→ Integration by parts (check!): this $= 0 - 0 + \int_0^{\infty} 2w\, e^{-\lambda w}\, dw$.

→ Integration by parts again: this $= 0 - 0 + \int_0^{\infty} 2\,\frac{1}{\lambda}\, e^{-\lambda w}\, dw$.

→ But $\int_0^{\infty} e^{-\lambda w}\, dw = -\frac{1}{\lambda} e^{-\lambda w}\Big|_{w=0}^{w=\infty} = -\frac{1}{\lambda}(0 - 1) = \frac{1}{\lambda}$.

→ So, $\text{E}(W^2) = 2\,\frac{1}{\lambda}\,\frac{1}{\lambda} = 2/\lambda^2$.

→ Then $\text{Var}(W) = \text{E}(W^2) - (\mu_W)^2 = (2/\lambda^2) - (1/\lambda)^2 = 1/\lambda^2$. Phew!

→ Hence, $\text{Sd}(W) = 1/\lambda$.

• e.g. $Z \sim \text{Normal}(0, 1)$. We know $\mu_Z := \text{E}(Z) = 0$.

→ Also $\text{E}(Z^2) = \int_{-\infty}^{\infty} z^2\, \frac{1}{\sqrt{2\pi}}\, e^{-z^2/2}\, dz$.

→ Then, integration by parts with $u = z$ and $v = -e^{-z^2/2}$ and $dv = z\, e^{-z^2/2}\, dz$ gives $\text{E}(Z^2) = 0 - 0 + \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}}\, e^{-z^2/2}\, dz = \int_{-\infty}^{\infty} \phi(z)\, dz = 1$ since ϕ is a density.

→ Hence, $\text{Var}(Z) = 1 - (\mu_z)^2 = 1 - 0^2 = 1$. (As expected.) Also $\text{Sd}(Z) = \sqrt{1} = 1$.

• Now suppose $W \sim \text{Normal}(\mu, \sigma^2)$, where $\sigma > 0$. What is $\text{Var}(W)$?

→ Well, this means that $W = \mu + \sigma Z$ where $Z \sim \text{Normal}(0, 1)$.

→ So, $\text{Var}(W) = \text{Var}(\mu + \sigma Z) = \sigma^2\,\text{Var}(Z) = \sigma^2$. Also $\text{Sd}(W) = \sqrt{\sigma^2} = \sigma$.

• Suppose $X \sim \text{Poisson}(\lambda)$. Know $\text{E}(X) = \lambda$. What is $\text{Var}(X)$?

→ We compute that: $\text{E}(X^2) = \sum_{k=0}^{\infty} k^2 e^{-\lambda} \frac{\lambda^k}{k!} = \lambda e^{-\lambda} \sum_{k=1}^{\infty} \Big((k-1) + 1\Big) \frac{\lambda^{k-1}}{(k-1)!} = \lambda e^{-\lambda} \left(\lambda \sum_{k=2}^{\infty} \frac{\lambda^{k-2}}{(k-2)!} + \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!}\right) = \lambda e^{-\lambda} \left(\lambda e^{\lambda} + e^{\lambda}\right) = \lambda^2 + \lambda$.

→ Then $\text{Var}(X) = \text{E}(X^2) - (\text{E}(X))^2 = (\lambda^2 + \lambda) - (\lambda)^2 = \lambda$. Phew! Simple!

Suggested Homework: 3.3.1(b), 3.3.2(a,c), 3.3.4(first four), 3.3.10(first four), 3.3.11(first three).

END WEDNESDAY #8

Covariance and Correlation

- We know that $\mathrm{E}(X+Y) = \mathrm{E}(X) + \mathrm{E}(Y)$. What about $\mathrm{Var}(X+Y)$?

- Well, $\mathrm{Var}(X+Y) = \mathrm{E}[(X+Y-\mu_{X+Y})^2] = \mathrm{E}[(X+Y-\mu_X-\mu_Y)^2] = \mathrm{E}[((X-\mu_X)+(Y-\mu_Y))^2] = \mathrm{E}[(X-\mu_X)^2 + (Y-\mu_Y)^2 + 2(X-\mu_X)(Y-\mu_Y)]$.

- This equals $\mathrm{Var}(X) + \mathrm{Var}(Y) + 2\,\mathrm{Cov}(X,Y)$, where $\mathrm{Cov}(X,Y) := \mathrm{E}[(X-\mu_X)(Y-\mu_Y)]$ is the covariance of X and Y.

  → We always have $\mathrm{Cov}(X,Y) = \mathrm{Cov}(Y,X)$.

  → If $\mathrm{Cov}(X,Y) > 0$, then X and Y tend to increase together.

  → If $\mathrm{Cov}(X,Y) < 0$, then X and Y tend to increase oppositely.

- Special case: If $Y = X$, then $\mathrm{Cov}(X,Y) = \mathrm{Cov}(X,X) = \mathrm{E}[(X-\mu_X)(X-\mu_X)] = \mathrm{E}[(X-\mu_X)^2] = \mathrm{Var}(X)$. In particular, $\mathrm{Cov}(X,X) \geq 0$.

  → Or, if $Y = -X$, then $\mathrm{Cov}(X,Y) = \mathrm{Cov}(X,-X) = \mathrm{E}[(X-\mu_X)(-X-\mu_{-X})] = \mathrm{E}[-(X-\mu_X)^2] = -\mathrm{Var}(X)$. In particular, $\mathrm{Cov}(X,-X) \leq 0$.

- If X and Y are independent, then $\mathrm{Cov}(X,Y) = \mathrm{E}[(X-\mu_X)(Y-\mu_Y)]$ $= \mathrm{E}[X-\mu_X]\,\mathrm{E}[Y-\mu_Y] = [\mu_X-\mu_X]\,[\mu_Y-\mu_Y] = 0 \cdot 0 = 0$.

  → Then $\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\,\mathrm{Cov}(X,Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$.

  → That is: variances add for sums of independent random variables.

  → Since $\mathrm{Sd}(X) = \sqrt{\mathrm{Var}(X)}$, can also write $\mathrm{Sd}(X+Y) = \sqrt{\mathrm{Sd}(X)^2 + \mathrm{Sd}(Y)^2}$.

  → ("propagation of uncertainty" for independent sums; e.g. quantum mechanics?)

- e.g. If $Y \sim \mathrm{Binomial}(n,\theta)$, then we can think of Y as $Y = X_1 + X_2 + \ldots + X_n$ where each $X_i \sim \mathrm{Bernoulli}(\theta)$ and they are independent.

  → By independence, $\mathrm{Cov}(X_i, X_j) = 0$ for all $i \neq j$.

  → Hence, $\mathrm{Var}(Y) = \mathrm{Var}(X_1) + \mathrm{Var}(X_2) + \ldots + \mathrm{Var}(X_n) = \theta(1-\theta) + \theta(1-\theta) + \ldots + \theta(1-\theta) = n\theta(1-\theta)$. This gives the variance of the $\mathrm{Binomial}(n,\theta)$ distribution!

- In general, by multiplying out, we have $\mathrm{Cov}(X,Y) = \mathrm{E}[(X-\mu_X)(Y-\mu_Y)] = \mathrm{E}[XY - \mu_X Y - X\mu_Y + \mu_X\mu_Y] = \mathrm{E}[XY] - \mu_X\mu_Y - \mu_X\mu_Y + \mu_X\mu_Y = \mathrm{E}[XY] - \mu_X\mu_Y$.

  → (Just like how $\mathrm{Var}(X) = \mathrm{E}[X^2] - (\mu_X)^2$. Makes sense.)

- We know that $\mathrm{E}(aX+bY) = a\,\mathrm{E}(X) + b\,\mathrm{E}(Y)$, and $\mathrm{Var}(aX+b) = a^2\,\mathrm{Var}(X)$. But what about $\mathrm{Cov}(aX+bY, Z)$?

  → Well, $\mathrm{Cov}(aX+bY, Z) = \mathrm{E}[(aX+bY-\mu_{aX+bY})(Z-\mu_Z)]$ $= \mathrm{E}[(aX+bY-a\mu_X-b\mu_Y)(Z-\mu_Z)] = \mathrm{E}[(a(X-\mu_X)+b(Y-\mu_Y))(Z-\mu_Z)]$ $= a\,\mathrm{E}[(X-\mu_X)(Z-\mu_Z)] + b\,\mathrm{E}[(Y-\mu_Y))(Z-\mu_Z)] = a\,\mathrm{Cov}(X,Z) + b\,\mathrm{Cov}(Y,Z)$.

  → Similarly, $\mathrm{Cov}(X, aY+bZ) = a\,\mathrm{Cov}(X,Y) + b\,\mathrm{Cov}(X,Z)$. ("bilinear")

- Let $X \sim \mathrm{Uniform}[5,9]$, and $Y \sim \mathrm{Exponential}(3)$, with X and Y independent.

  → Then $\mathrm{Cov}(X,Y) = 0$ (by independence).

  → And if $Z = 3X + 2Y$ and $W = X - 5Y$, then $\mathrm{Cov}(Z,W) = \mathrm{Cov}(3X+2Y, X-5Y) = 3\,\mathrm{Cov}(X, X-5Y) + 2\,\mathrm{Cov}(Y, X-5Y)$ $= 3\,\mathrm{Cov}(X,X) - 15\,\mathrm{Cov}(X,Y) + 2\,\mathrm{Cov}(Y,X) - 10\,\mathrm{Cov}(Y,Y)$

$= 3\,\mathrm{Var}(X) - 15(0) + 2(0) - 10\,\mathrm{Var}(Y) = 3\,(4^2/12) - 10(1/3^2) = 26/9.$

• Fact: If $X \sim \mathrm{Normal}(\mu_1, \sigma_1^2)$, and $Y \sim \mathrm{Normal}(\mu_2, \sigma_2^2)$, with X and Y independent, then $X + Y$ is also normal (!). (Textbook Problem 2.9.14.)

→ What mean and variance?

→ By linearity and independence, $\mathrm{E}(X + Y) = \mathrm{E}(X) + \mathrm{E}(Y) = \mu_1 + \mu_2$, and $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) = \sigma_1^2 + \sigma_2^2$, so $X + Y \sim \mathrm{Normal}(\mu_1 + \mu_2,\, \sigma_1^2 + \sigma_2^2)$.

• Suppose now that $Y = cX$ for some constant $c \in \mathbf{R}$.

→ Then $\mathrm{Var}(Y) = c^2\,\mathrm{Var}(X)$, so $\mathrm{Sd}(Y) = |c|\,\mathrm{Sd}(X)$, and $\mathrm{Sd}(X)\,\mathrm{Sd}(Y) = |c|\,\mathrm{Var}(X)$.

→ Also, $\mathrm{Cov}(X, Y) = \mathrm{Cov}(X, cX) = c\,\mathrm{Cov}(X, X) = c\,\mathrm{Var}(X)$.

→ So, if $c \geq 0$, then $\mathrm{Cov}(X, Y) = \mathrm{Sd}(X)\,\mathrm{Sd}(Y)$.

→ Or, if $c < 0$, then $\mathrm{Cov}(X, Y) = -\mathrm{Sd}(X)\,\mathrm{Sd}(Y)$.

• Prop: these are the extremes, i.e. always $-\mathrm{Sd}(X)\,\mathrm{Sd}(Y) \leq \mathrm{Cov}(X, Y) \leq \mathrm{Sd}(X)\,\mathrm{Sd}(Y)$.

→ That is, we always have $-\sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)} \leq \mathrm{Cov}(X, Y) \leq \sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)}$.

• Proof: Use the "Cauchy-Schwarz Inequality" that $-||u||\;||v|| \leq u \cdot v \leq ||u||\;||v||$.

→ Here the "vector space" is all random variables with finite variance.

→ And, the "dot product" is $X \cdot Y = \mathrm{Cov}(X, Y)$.

→ So, $\|X\| = \sqrt{X \cdot X} = \sqrt{\mathrm{Cov}(X, X)} = \sqrt{\mathrm{Var}(X)} = \mathrm{Sd}(X)$.

→ So, the result follows by setting $u = X$ and $v = Y$. ∎

• The correlation of X and Y is $\mathrm{Corr}(X, Y) = \mathrm{Cov}(X, Y)\big/\sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)}$.

→ So we always have $-1 \leq \mathrm{Corr}(X, Y) \leq 1$.

→ $\mathrm{Corr}(X, Y)$ is a "normalised" version of $\mathrm{Cov}(X, Y)$.

→ Can also be written as $\mathrm{Corr}(X, Y) = \mathrm{Cov}(X, Y)\big/[\mathrm{Sd}(X)\,\mathrm{Sd}(Y)]$.

→ (Requires first computing μ_X, μ_Y, $\mathrm{Var}(X)$, $\mathrm{Var}(Y)$, $\mathrm{Cov}(X, Y)$, ...)

• Now suppose that Y is a constant r.v., e.g. $Y = 5$. Then what is $\mathrm{Cov}(X, 5)$?

→ Well, $\mathrm{Cov}(X, Y) := \mathrm{E}[(X - \mu_X)(Y - \mu_Y)] = \mathrm{E}[(X - \mu_X)(5 - 5)] = 0$.

→ Of course! And what about $\mathrm{Corr}(X, 5)$?

→ Well, $\mathrm{Var}(Y) = 0$, so $\mathrm{Corr}(X, Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)}} = \frac{0}{0}$. Undefined!

→ Correlation is only defined for non-constant r.v.: $\mathrm{Var}(X) > 0$ and $\mathrm{Var}(Y) > 0$.

• e.g. Suppose $Z = c\,Y$ for some $c > 0$. How is $\mathrm{Corr}(X, Z)$ related to $\mathrm{Corr}(X, Y)$?

→ Here $\mathrm{Var}(Z) = c^2\,\mathrm{Var}(Y)$, so $\mathrm{Sd}(Z) = \sqrt{\mathrm{Var}(Z)} = \sqrt{c^2\,\mathrm{Var}(Y)} = c\,\mathrm{Sd}(Y)$.

→ But also, $\mathrm{Cov}(X, Z) = \mathrm{Cov}(X, cY) = c\,\mathrm{Cov}(X, Y)$.

→ Hence, $\mathrm{Corr}(X, Z) = \frac{\mathrm{Cov}(X,Z)}{\mathrm{Sd}(X)\,\mathrm{Sd}(Z)} = \frac{c\,\mathrm{Cov}(X,Y)}{\mathrm{Sd}(X)\,c\,\mathrm{Sd}(Y)} = \frac{\mathrm{Cov}(X,Y)}{\mathrm{Sd}(X)\,\mathrm{Sd}(Y)} = \mathrm{Corr}(X, Y)$.

→ That is, $\mathrm{Corr}(X, cY) = \mathrm{Corr}(X, Y)$. Unaffected by the constant scale $c > 0$.

• If instead $Z = c\,Y$ where $c < 0$, then $\sqrt{c^2} = -c$, so $\mathrm{Corr}(X, cY) = -\mathrm{Corr}(X, Y)$.

→ So, the sign of c is still important! (But not its magnitude.)

• We always have $\mathrm{Corr}(X, X) = \frac{\mathrm{Cov}(X,X)}{\mathrm{Sd}(X)\,\mathrm{Sd}(X)} = \frac{\mathrm{Var}(X)}{\mathrm{Var}(X)} = 1$.

→ And, $\mathrm{Corr}(X, cX) = \mathrm{sign}(c)$, i.e. $= 1$ if $c > 0$, or $= -1$ if $c < 0$.

→ And what about if $c = 0$? ...

Suggested Homework: 3.3.1, 3.3.2, 3.3.3, 3.3.4, 3.3.7, 3.3.10, 3.3.11, 3.3.12, 3.3.13, 3.3.14, 3.3.15, 3.3.29, 3.3.30.

[Reminder: Extra Prof office hours tomorrow; see web page.]

[Reminder: Midterm #2 this Wednesday Nov 15 in EX200.]

END MONDAY #9

(Midterm #2.)

END WEDNESDAY #9

• e.g. Suppose $p_{X,Y}(5,1) = p_{X,Y}(5,9) = p_{X,Y}(7,3) = p_{X,Y}(7,7) = 1/4$, otherwise 0. What is $\mathrm{Cov}(X,Y)$? And, are X and Y independent? Diagram:

→ Here $\mu_X := \mathrm{E}(X) = \sum_{x\in\mathbf{R}} x\, p_X(x) = \sum_{x,y\in\mathbf{R}} x\, p_{X,Y}(x,y) = 5(1/4) + 5(1/4) + 7(1/4) + 7(1/4) = 6$.

→ And $\mu_Y := \mathrm{E}(Y) = \sum_{y\in\mathbf{R}} y\, p_Y(y) = \sum_{x,y\in\mathbf{R}} y\, p_{X,Y}(x,y) = 1(1/4) + 9(1/4) + 3(1/4) + 7(1/4) = 5$.

→ Also $\mathrm{E}(XY) = \sum_{x,y\in\mathbf{R}} xy\, p_{X,Y}(x,y) = (5)(1)(1/4)+(5)(9)(1/4)+(7)(3)(1/4)+(7)(7)(1/4) = 30$.

→ So, $\mathrm{Cov}(X,Y) = \mathrm{E}(XY) - \mu_X\mu_Y = 30 - (6)(5) = 0$, i.e. $\mathrm{E}(XY) = \mathrm{E}(X)\,\mathrm{E}(Y)$.

→ Hence, also, $\mathrm{Corr}(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)}} = 0$, too. ("Uncorrelated")

→ And also $\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$, since $\mathrm{Cov}(X,Y) = 0$.

→ So, does that mean that X and Y must be independent?

→ No, since e.g. $p_X(5) = 1/4 + 1/4 = 1/2 > 0$ and $p_Y(3) = 1/4 > 0$, but $p_{X,Y}(5,3) = 0 \neq p_X(5)\, p_Y(3)$. So, X and Y are not independent!

→ Conclusion: independent $\Rightarrow$ uncorrelated, but uncorrelated $\not\Rightarrow$ independent.

Markov's Inequality

• Suppose $X \geq 0$, and $\mathrm{E}(X) = 5$. Can $\mathrm{P}(X > 100)$ be very large?

→ No, since then we would have $\mathrm{E}(X) \geq (100)\,\mathrm{P}(X > 100) \gg 5$.

→ Indeed, to make $\mathrm{E}(X) = 5$, we need to have $(100)\,\mathrm{P}(X > 100) \leq 5$.

• Markov's Inequality: If $X \geq 0$, and $a > 0$, then $\mathrm{P}(X \geq a) \leq \mathrm{E}(X) \,/\, a$.

• Proof: Define a new random variable Z by $Z = a\, I_{X\geq a}$.

→ That is, $Z = a$ whenever $X \geq a$, otherwise $Z = 0$.

→ Then if $X \geq a$, then $Z = a$, so $X \geq Z$.

→ Or, if $X < a$, then $Z = 0$, so $X \geq Z$ (since we've assumed $X \geq 0$).

→ Either way, $X \geq Z$. So, by monotonicity, $\mathrm{E}(X) \geq \mathrm{E}(Z)$.

→ But $\mathrm{E}(Z) = \mathrm{E}[a\, I_{X \geq a}] = a\, \mathrm{P}(X \geq a)$. So, $\mathrm{E}(X) \geq a\ \mathrm{P}(X \geq a)$. ∎

- e.g. If $X \geq 0$ and $\mathrm{E}(X) = 5$, then must have $\mathrm{P}(X \geq 100) \leq 5/100 = 1/20$.

→ Also, $\mathrm{P}(X \geq 1000) \leq 5/1000 = 1/200$. Small!

- But this is only for non-negative random variables. Better is ...

Chebychev's Inequality

- Let Y be any random variable, with finite mean μ_Y.

→ If $\mathrm{Var}(Y)$ is small, then Y will usually be close to μ_Y. More precise?

- Chebychev's Inequality: For any $a > 0$, $\mathrm{P}(|Y - \mu_Y| \geq a) \leq \mathrm{Var}(Y) \,/\, a^2$.

- Proof: Let $X = (Y - \mu_Y)^2 \geq 0$. Then by Markov's Inequality, $\mathrm{P}(|Y - \mu_Y| \geq a) = \mathrm{P}((Y - \mu_Y)^2 \geq a^2) \leq \mathrm{E}((Y - \mu_Y)^2) \,/\, a^2 = \mathrm{Var}(Y) \,/\, a^2$. ∎

- e.g. Suppose Z has mean 5 and variance 9. Then, $\mathrm{P}(Z \geq 17) = \mathrm{P}(Z - 5 \geq 12) \leq \mathrm{P}(|Z - 5| \geq 12) \leq 9/12^2 = 9/144 = 1/16 = 0.0625$. Unlikely!

→ And, this is true for any random variable with this mean and variance.

→ If we also knew that $Z \geq 0$, then we could use Markov's inequality directly to get that $\mathrm{P}(Z \geq 17) \leq \mathrm{E}(Z)/17 = 5/17 \doteq 0.294$. (Weaker bound.)

Suggested Homework: 3.6.1, 3.6.2, 3.6.3, 3.6.4, 3.6.5, 3.6.6, 3.6.8, 3.6.9, 3.6.10, 3.6.11, 3.6.12, 3.6.13, 3.6.14, 3.6.15, 3.6.18.

[END OF TEXTBOOK CHAPTER #3]

Convergence of Random Variables

- Suppose we flip 100 coins.

→ Will the number of Heads be close to 50? How close?

→ Will the fraction of Heads be close to 0.5?

→ If we flip 1,000 coins, will it be closer to 0.5?

→ Maybe? Usually? For sure??

- [Try it in R: e.g. "mean(rbinom(1000,1,1/2))", "mean(rgeom(1000,1/5))", "mean(rpois(1000,3))", "mean(rexp(1000,3))"]

- If we flip n coins as $n \to \infty$, will the fraction get even closer to 1/2?

→ Will the fraction converge to 1/2? For sure? In what sense?

→ What does it mean for a random quantity to converge??

Convergence in Probability

• Defn: A sequence $X_1, X_2, X_3, \ldots$ of random variables converges in probability to another random variable (or constant) Y if: For all $\epsilon > 0$, $\lim_{n\to\infty} \mathrm{P}(|X_n - Y| \geq \epsilon) = 0$.

→ Or, equivalently: For all $\epsilon > 0$, $\lim_{n\to\infty} \mathrm{P}(|X_n - Y| < \epsilon) = 1$.

→ Sometimes written as: $\{X_n\} \xrightarrow{P} Y$, or just $X_n \xrightarrow{P} Y$.

• e.g. Suppose $X_n \sim$ Bernoulli($\frac{1}{n}$), i.e. $\mathrm{P}(X_n{=}1) = \frac{1}{n}$ and $\mathrm{P}(X_n{=}0) = 1 - \frac{1}{n}$.

→ Does $X_n \to 0$ in probability, i.e. $X_n \xrightarrow{P} 0$?

→ For any $\epsilon > 0$, $\mathrm{P}(|X_n - 0| \geq \epsilon) \leq \mathrm{P}(X_n{\neq}0) = \mathrm{P}(X_n{=}1) = \frac{1}{n}$, and this probability $\to 0$ as $n \to \infty$. So, yes, $X_n \xrightarrow{P} 0$.

• In general, for any $\epsilon > 0$, $\mathrm{P}(|X_n - Y| \geq \epsilon) \leq \mathrm{P}(X_n \neq Y)$.

→ So, if $\lim_{n\to\infty} \mathrm{P}(X_n \neq Y) = 0$, then $X_n \xrightarrow{P} Y$.

• e.g. Let $U \sim$ Uniform$[0, 1]$, and $X_n = I_{U \leq (1/2)+(1/2^n)}$, and $Y = I_{U \leq 1/2}$.

→ Does $X_n \to Y$ in probability?

→ For any $\epsilon > 0$, $\mathrm{P}(|X_n - Y| \geq \epsilon) \leq \mathrm{P}(X_n \neq Y) = \mathrm{P}(X_n = 1 \text{ and } Y = 0) = \mathrm{P}[1/2 < U \leq (1/2) + (1/2^n)] = 1/2^n$, and this probability $\to 0$ as $n \to \infty$. Yes!

• e.g. Let $Y \sim$ Uniform$[0, 5]$, and $X_n = (1 + \frac{1}{n})Y$. Does $X_n \to Y$ in probability?

→ Here $|X_n - Y| = |(1 + \frac{1}{n})Y - Y| = \frac{1}{n} Y \leq 5/n$.

→ Now, for any $\epsilon > 0$, if $n > 5/\epsilon$, then $5/n < \epsilon$.

→ Hence, for all $n > 5/\epsilon$, we must have $|X_n - Y| \leq 5/n < \epsilon$.

→ This means that for all $n > 5/\epsilon$, $\mathrm{P}(|X_n - Y| \geq \epsilon) = 0$.

→ So, yes, $\lim_{n\to\infty} \mathrm{P}(|X_n - Y| \geq \epsilon) = 0$, i.e. $X_n \to Y$ in probability. Yes!

• e.g. Flip an infinite sequence of fair coins.

→ Let $X_n = I_{n^{\text{th}} \text{ coin Heads}}$, i.e. $X_n = 1$ if the n^{th} coin is Heads, otherwise 0.

→ Does $X_n \to 1/2$ in probability?

→ No! For $0 < \epsilon < 1/2$, we have $\mathrm{P}(|X_n - (1/2)| \geq \epsilon) = 1$, not $\to 0$.

→ But suppose instead we let $M_n = \frac{1}{n}(X_1 + X_2 + \ldots + X_n)$.

→ Then M_n is the fraction of Heads in the first n coins.

→ Does $M_n \to 1/2$ in probability? Maybe!

Suggested Homework: 4.2.1, 4.2.2, 4.2.6, 4.2.7, 4.2.8, 4.2.14, 4.2.17.

END MONDAY #10

Weak Law of Large Numbers (WLLN)

• Theorem: For any sequence of random variables $X_1, X_2, X_3, \ldots$ which are independent, and each have the same mean μ, and each have variance $\leq v$ for some constant $v < \infty$, if $M_n = \frac{1}{n}(X_1 + X_2 + \ldots + X_n)$, then $M_n \to \mu$ in probability.

- Proof: We need to understand M_n better.

  → First, by linearity, $\mathrm{E}(M_n) = \frac{1}{n}[\mathrm{E}(X_1) + \mathrm{E}(X_2) + \ldots + \mathrm{E}(X_n)] = \frac{1}{n}[n\,\mu] = \mu$.

  → Then, since the $\{X_n\}$ are independent, $\mathrm{Var}(M_n) = (\frac{1}{n})^2[\mathrm{Var}(X_1) + \mathrm{Var}(X_2) + \ldots + \mathrm{Var}(X_n)] \leq (\frac{1}{n})^2[v + v + \ldots + v] = (\frac{1}{n})^2[n\,v] = v/n$. (Not just v.)

  → Now, let $\epsilon > 0$, and consider $\mathrm{P}(|M_n - \mu| \geq \epsilon)$.

  → Use Chebychev's Inequality! Since $\mathrm{E}(M_n) = \mu$, therefore $\mathrm{P}(|M_n - \mu| \geq \epsilon) \leq \mathrm{Var}(M_n)/\epsilon^2 \leq v/n\epsilon^2$, which $\to 0$ as $n \to \infty$.

  → So, $M_n \to \mu$ in probability. ∎

- Often assume the $\{X_n\}$ are i.i.d., i.e. independent and identically distributed.

  → "identically distributed" means the X_n all have the same probabilities.

  → That is, $\mathrm{P}(a \leq X_n \leq b)$ is the same for all n (for any $a < b$).

  → In particular, the X_n all have the same mean μ and variance v.

  → Fact: If $\{X_n\}$ i.i.d., then the WLLN doesn't even need $v < \infty$.

- e.g. Flip an infinite sequence of fair coins, with $X_n = I_{n^{\text{th}}\text{ coin Heads}}$.

  → Then $\{X_n\}$ independent (and i.i.d.), with $\mathrm{E}(X_n) = 1/2 =: \mu$, and $\mathrm{Var}(X_n) = (1/2)(1 - (1/2)) = 1/4 =: v < \infty$.

  → So, if $M_n = \frac{1}{n}(X_1 + X_2 + \ldots + X_n)$ is the fraction of Heads on the first n fair coin flips, then by WLLN, $M_n \to \mu = 1/2$ in probability.

  → Hence, $\mathrm{P}(|M_n - (1/2)| \geq \epsilon) \to 0$ for all $\epsilon > 0$.

  → e.g. $\epsilon = 0.003$: $\mathrm{P}(|M_n - (1/2)| \geq 0.003) \to 0$.

  → So, for all sufficiently large n, $\mathrm{P}(|M_n - (1/2)| \geq 0.003) < 0.01$ (say).

  → In particular, for those n, $\mathrm{P}(M_n - (1/2) \geq 0.003) < 0.01$, i.e. $\mathrm{P}(M_n \geq 0.503) < 0.01$, i.e. $\mathrm{P}(M_n < 0.503) > 0.99$, etc.

- e.g. Roll an infinite sequence of fair dice, with X_n the result of the n^{th} roll.

  → Then $\{X_n\}$ independent (and i.i.d.), and $\mathrm{E}(X_n) = 3.5 =: \mu$.

  → What about $\mathrm{Var}(X_n)$? Well, $\mathrm{E}(X_n^2) = \sum_{x \in \mathbf{R}} x^2\,\mathrm{P}(X_n = x) = \sum_{k=1}^{6} k^2\,(1/6) = 91/6$. So $\mathrm{Var}(X_n) = 91/6 - (3.5)^2 \doteq 2.92 =: v < \infty$.

  → (Or, simpler: We always have $1 \leq X_n \leq 6$, so $|X_n - 3.5| \leq 2.5$, so $\mathrm{Var}(X_n) = \mathrm{E}(|X_n - 3.5|^2) \leq (2.5)^2 =: v < \infty$, since we only need the variances to be bounded.)

  → (Or, even simpler: since $\{X_n\}$ i.i.d., don't need to check variance.)

  → So, if $M_n = \frac{1}{n}(X_1 + X_2 + \ldots + X_n)$ is the average value on the first n fair dice, then by WLLN, $M_n \to \mu = 3.5$ in probability.

- e.g. Repeatedly take free throws, with independent probability $\theta = 1/4$ of scoring each time. Let $X_n = I_{\text{score on } n^{\text{th}}\text{ attempt}}$.

  → Then $\{X_n\}$ independent, $\mathrm{E}(X_n) = \theta =: \mu$, and $\mathrm{Var}(X_n) = \theta(1 - \theta) =: v < \infty$.

  → So, if $M_n = \frac{1}{n}(X_1 + X_2 + \ldots + X_n)$ is the fraction of scores on the first n attempts, then by WLLN, $M_n \to \mu = 1/4$ in probability.

  → So, after e.g. 1,000 attempts, you will probably have about 250 scores.

• Again repeatedly take free throws, with independent probability $\theta = 1/4$ of scoring each time. How many attempts to score 500 times?

→ Let X_n be the number of misses just before the n^{th} score (i.e., in between the $(n-1)^{\text{th}}$ and n^{th} scores).

→ Then $X_n \sim$ Geometric(1/4), so $\mathrm{E}(X_n) = (1-\theta)/\theta = (3/4)/(1/4) = 3$.

→ Let $Z_n = X_n + 1$, so Z_n is the total number of attempts for the n^{th} score.

→ So, $\mathrm{E}(Z_n) = \mathrm{E}(X_n) + 1 = 4$.

→ Then, W := "# attempts to score 500 times" $= Z_1 + Z_2 + \ldots + Z_{500}$.

→ Fact: Geometric(1/4) has finite variance, which we'll call $v < \infty$.

→ (Actually $v = (1-\theta)/\theta^2$, see Problem 3.3.18, but we don't need that.)

→ (Or, even simpler: since $\{X_n\}$ and $\{Z_n\}$ i.i.d., don't need to check variance.)

→ So, if $M_n = \frac{1}{n}(Z_1 + Z_2 + \ldots + Z_n)$, then by WLLN, $M_n \to 4$ in probability.

→ So, $M_{500} \approx 4$, i.e. $W := Z_1 + Z_2 + \ldots + Z_{500} \approx (4)(500) = 2000$.

→ So, it will probably take about 2000 attempts to score 500 free throws.

Suggested Homework: 4.2.3, 4.2.4, 4.2.5, 4.2.10, 4.2.11. Optional: 4.2.12, 4.2.13.

Convergence Almost Surely (a.s.) (with Probability 1)

• Why is the above called just the "weak" law of large numbers?

→ e.g. For a sequence of fair coins, we know $M_n \xrightarrow{P} 1/2$.

→ This means that for large n, probably $M_n \approx 1/2$.

→ But does this mean the random sequence M_n actually converges to 1/2?

→ What does that sort of convergence even mean?

• e.g. Define a sequence of r.v. $X_1, X_2, X_3, \ldots$ as follows.

→ Most of the X_n are equal to 5.

→ However, one of variables $X_1, X_2, \ldots, X_9$ is selected (uniformly at random) and is instead set to be equal to 7. (But the rest are still equal to 5.)

→ And, one of variables $X_{10}, X_{11}, \ldots, X_{99}$ is selected (uniformly at random) and is instead set to be equal to 7. (But the rest are still equal to 5.)

→ And, one of variables $X_{100}, X_{101}, \ldots, X_{999}$ is selected (uniformly at random) and is instead set to be equal to 7. (But the rest are still equal to 5.)

→ And, one of variables $X_{1000}, X_{1001}, \ldots, X_{9999}$ is selected (uniformly at random) and is instead set to be equal to 7. (But the rest are still equal to 5.)

→ And so on. For each $k = 1, 2, 3, \ldots$, one of the X_n for those n which have exactly k digits is selected (uniformly at random) and is instead set to be equal to 7.

→ So, when we're done, the sequence $X_1, X_2, X_3, \ldots$ looks something like:
5,5,5,7,5,5,5,5,5,5,5,5,. . . ,5,5,7,5,5,. . . . . ,5,5,7,5,5, ... ,5,5,7,5,5,. . . . . .

• Does this sequence $X_1, X_2, X_3, \ldots$ converge to 5 in probability?

→ Well, for $1 \le n \le 9$, $\mathrm{P}(X_n = 7) = 1/9$ and $\mathrm{P}(X_n = 5) = 1 - [1/9]$.

→ And, for $10 \le n \le 99$, $\mathrm{P}(X_n = 7) = 1/90$ and $\mathrm{P}(X_n = 5) = 1 - [1/90]$.

→ And, for $100 \le n \le 999$, $\mathrm{P}(X_n = 7) = 1/900$ and $\mathrm{P}(X_n = 5) = 1 - [1/900]$.

→ And, for $1000 \le n \le 9999$, $\mathrm{P}(X_n = 7) = 1/9000$ and $\mathrm{P}(X_n = 5) = 1 - [1/9000]$.

→ In general, if n has k digits (in base 10), then we compute that: $\mathrm{P}(X_n = 7) = 1/(9 \cdot 10^{k-1})$ and $\mathrm{P}(X_n = 5) = 1 - [1/(9 \cdot 10^{k-1}]$.

→ [To be fancy, we could write this as: $\mathrm{P}(X_n = 7) = 1/(9 \cdot 10^{\lfloor \log_{10}(n) \rfloor})$.]

→ The key is that $\lim_{n\to\infty} \mathrm{P}(X_n = 7) = 0$ and $\lim_{n\to\infty} \mathrm{P}(X_n = 5) = 1$.

→ Hence, for any $\epsilon > 0$, $\lim_{n\to\infty} \mathrm{P}(|X_n - 5| \ge \epsilon) \le \lim_{n\to\infty} \mathrm{P}(|X_n - 5| \neq 0) = \lim_{n\to\infty} \mathrm{P}(X_n = 7) = 0$.

→ So, yes, $\{X_n\} \to 5$ in probability, i.e. $X_n \xrightarrow{P} 5$.

• Okay, great. But does the actual sequence $\{X_n\}$ actually converge to 5?

→ Recall that it looks something like:
5,5,5,7,5,5,5,5,5,5,5,5,5,. . . ,5,5,7,5,5,. . . . . ,5,5,7,5,5, ... ,5,5,7,5,5,. . . . . .

→ So, even though it usually equals 5, it still equals 7 infinitely often.

→ But $X_n \to 5$ as a sequence means: For all $\epsilon > 0$, there is $N \in \mathbf{N}$ such that for all $n \ge N$, we have $|X_n - 5| \le \epsilon$.

→ This cannot ever hold (for any $0 < \epsilon < 2$), since an infinite number of the X_n equal 7, with $|X_n - 5| = |7 - 5| = 2 > \epsilon$. That is, $X_n \to 5$ as a sequence is impossible!

→ Conclusion: $\mathrm{P}(X_n \to 5$ as a sequence of numbers$) = 0$. Can never happen!

• So, just because $X_n \xrightarrow{P} 5$, that does not mean that $\mathrm{P}(X_n \to 5$ as a sequence$) = 1$; that probability could still be 0. In this sense, convergence in probability is "weak".

• Defn: A sequence $X_1, X_2, X_3, \ldots$ of r.v. converges almost surely or converges a.s. or converges with probability 1 to another r.v. Y if $\mathrm{P}(X_n \to Y$ as a sequence$) = 1$, i.e. $\mathrm{P}(\lim_{n\to\infty} X_n = Y) = 1$. This is sometimes written as: $X_n \xrightarrow{a.s.} Y$.

• So, in the above example $X_n \xrightarrow{P} 5$, but $X_n \not\xrightarrow{a.s.} 5$. i.e. we do not have $X_n \xrightarrow{a.s.} 5$.

• However, the converse always holds – convergence almost surely is "stronger":

• Theorem: If $X_n \xrightarrow{a.s.} Y$, then $X_n \xrightarrow{P} Y$. [That is, if $\{X_n\}$ converges to Y almost surely (i.e. with probability 1), then it also converges to Y in probability.]

• Proof: Fix $\epsilon > 0$, and let A_n be the event that there is some $m \ge n$ with $|X_m - Y| \ge \epsilon$. That is, $A_n = \{\exists\ m \ge n$ with $|X_m - Y| \ge \epsilon\}$.

→ Or, as functions: $A_n = \{s \in S : \exists\ m \ge n$ with $|X_m(s) - Y(s)| \ge \epsilon\}$.

→ If $s \in \bigcap_{n=1}^{\infty} A_n$, this means we can always find some $m \ge n$ with $|X_m(s) - Y(s)| \ge \epsilon$, i.e. the sequence $\{X_n(s)\}$ does not converge as a sequence to $Y(s)$.

→ This shows: $\mathrm{P}\big(\{X_n\}$ does not converge as a sequence to $Y\big) \ge \mathrm{P}\big(\bigcap_{n=1}^{\infty} A_n\big)$.

→ But if $X_n \xrightarrow{a.s.} Y$, then $\mathrm{P}\big(\{X_n\}$ does converge as a sequence to $Y\big) = 1$, so $\mathrm{P}\big(\{X_n\}$ does not converge as a sequence to $Y\big) = 0$. Hence, $\mathrm{P}\big(\bigcap_{n=1}^{\infty} A_n\big) = 0$.

→ So what? Well, here $A_{n+1} \subseteq A_n$, i.e. the $\{A_n\}$ are decreasing.

→ So, by Continuity of Probabilities, $\lim_{n\to\infty} \mathrm{P}(A_n) = \mathrm{P}(\bigcap_{n=1}^{\infty} A_n) = 0$.

→ But $\mathrm{P}(|X_n - Y| \geq \epsilon) \leq \mathrm{P}(A_n)$, so $\lim_{n\to\infty} \mathrm{P}(|X_n - Y| \geq \epsilon) = 0$.

→ Since this is true for any $\epsilon > 0$, we must have $X_n \xrightarrow{P} Y$. ∎

• Intuition from the proof: For all $\epsilon > 0$, as $n \to \infty$, ...

→ For $X \xrightarrow{P} Y$, just need $\mathrm{P}(|X_n - Y| \geq \epsilon) \to 0$.

→ But for $X \xrightarrow{a.s.} Y$, need $\mathrm{P}(\exists\ m \geq n \text{ with } |X_m - Y| \geq \epsilon) \to 0$. (Stronger.)

Suggested Homework: 4.3.1, 4.3.2, 4.3.5, 4.3.10, 4.3.16, 4.3.17, 4.3.18, 4.3.19, 4.3.21, 4.3.22.

Strong Law of Large Numbers (SLLN)

• Theorem: For any sequence of random variables $X_1, X_2, X_3, \ldots$ which are i.i.d., each with the same mean μ, if $M_n = \frac{1}{n}(X_1 + X_2 + \ldots + X_n)$, then $M_n \to \mu$ almost surely (i.e., a.s.) (i.e., with probability 1) (i.e., $M_n \xrightarrow{a.s.} \mu$).

→ Proof in more advanced books, e.g. http://probability.ca/grprob

→ Then, of course, also $M_n \xrightarrow{P} \mu$, too. (WLLN)

• e.g. Flip an infinite sequence of fair coins, with $X_n = I_{n^{\text{th}} \text{ coin Heads}}$.

→ Then $\{X_n\}$ i.i.d., with $\mathrm{E}(X_n) = 1/2 =: \mu$.

→ So, if $M_n = \frac{1}{n}(X_1 + X_2 + \ldots + X_n)$ is the fraction of Heads on the first n fair coin flips, then by WLLN, $M_n \to \mu = 1/2$ in probability.

→ Hence, for all $\epsilon > 0$, $\mathrm{P}(|M_n - (1/2)| \geq \epsilon) \to 0$.

→ So, for all sufficiently large n, i.e. $\mathrm{P}(M_n < 0.503) > 0.99$, etc.

→ But the SLLN says more: $\mathrm{P}(M_n \to 1/2) = 1$.

→ So, for all $\epsilon > 0$, $\mathrm{P}(|M_n - 0.5| \leq \epsilon$ for all sufficiently large $n) = 1$.

→ So e.g. $\mathrm{P}(M_n < 0.503$ for all sufficiently large $n) = 1$.

→ In particular, $\mathrm{P}(\exists\ n : M_n < 0.503) = 1$.

→ That is, $\mathrm{P}(\exists\ n : X_1 + X_2 + \ldots + X_n < (0.503)n) = 1$. etc.

• Try it out in R! File http://probability.ca/Rslln (first choose theta):

```
N = 1000; M = rep(NA, N); X = rbinom(N, 1, theta)
for (i in 1:N) M[i] = mean(X[1:i])
plot(M, type='l', col="blue", ylim=c(0,1), xlab="n", ylab="Mn")
abline(h=theta, col="red", lty="dotted")
```

Suggested Homework: 4.3.3, 4.3.4, 4.3.6, 4.3.7, 4.3.8, 4.3.9, 4.3.11, 4.3.12.

END WEDNESDAY #10

Central Limit Theorem (CLT)

• Suppose $X_1, X_2, \ldots$ are independent and identically distributed, each with finite mean μ and finite variance σ^2. What can we say about the probabilities of their sum?

→ Let $S_n = X_1 + X_2 + \ldots + X_n$. So the average is $\frac{1}{n}S_n$.

→ We know that $\frac{1}{n}S_n \to \mu$. But how close?

→ What is the probability distribution of $\frac{1}{n}S_n - \mu$?

• Frequency histograms in R – file http://probability.ca/Rclt (first choose theta):

```
numrep=1000; N=1000; D = rep(NA,numrep)
for (i in 1:numrep) { X = rbinom(N, 1, theta); D[i] = mean(X) - theta }
hist(D, col="blue", xlab="Mn - mean", ylab="frequency", main="", breaks="Free")
```

• How does the frequency distribution look?

→ Usually centered near 0 (makes sense).

→ Width is fairly small (how small?).

→ Shape is approximately ... normal!?!

• For center, the mean is $\mathrm{E}[\frac{1}{n}S_n - \mu] = \frac{1}{n}(n\mu) - \mu = \mu - \mu = 0$. (Of course.)

• For width, let's compute the standard deviation:

→ Well, since the $\{X_i\}$ are i.i.d., $\mathrm{Var}(\frac{1}{n}S_n - \mu) = (\frac{1}{n})^2 \mathrm{Var}(S_n) = \frac{1}{n^2}\mathrm{Var}(X_1 + X_2 + \ldots + X_n) = \frac{1}{n^2}[\mathrm{Var}(X_1) + \mathrm{Var}(X_2) + \ldots + \mathrm{Var}(X_n)] = \frac{1}{n^2}[n\,\mathrm{Var}(X_i)] = \frac{1}{n}\mathrm{Var}(X_i)$.

→ So, if $\mathrm{Var}(X_i) = \sigma^2$, then $\mathrm{Var}(\frac{1}{n}S_n - \mu) = \sigma^2/n$. Small! Narrow!

→ So, $\mathrm{Var}\Big([\frac{1}{n}S_n - \mu]/\sqrt{\sigma^2/n}\Big) = \mathrm{Var}\big(\frac{1}{n}S_n - \mu\big)/(\sqrt{\sigma^2/n})^2 = \big(\sigma^2/n\big)/(\sigma^2/n) = 1$.

• So, let $Z_n = [\frac{1}{n}S_n - \mu]/\sqrt{\sigma^2/n} = \frac{S_n - n\mu}{\sqrt{n}\,\sigma}$. Then $\mathrm{E}(Z_n) = 0$, and $\mathrm{Var}(Z_n) = 1$.

→ Check: $\mathrm{E}(S_n) = n\mu$, and $\mathrm{Sd}(S_n) = \sqrt{n}\sigma$, so $Z_n := \frac{S_n - n\mu}{\sqrt{n}\,\sigma}$ has mean 0, var 1.

→ But is it really approximately normal??

• Theorem (CLT): The probabilities of Z_n converge to those of $Z \sim \mathrm{Normal}(0,1)$.

→ This means that for each $z \in \mathbf{R}$, $\lim_{n\to\infty} \mathrm{P}(Z_n \le z) = \mathrm{P}(Z \le z)$.

→ i.e. $F_{Z_n}(z) \to F_Z(z) =: \Phi(z)$ for all $z \in \mathbf{R}$. (Convergence in distribution)

→ Equivalently, $\lim_{n\to\infty} \mathrm{P}(S_n \le n\mu + \sqrt{n}\,\sigma z) = \mathrm{P}(Z \le z) \equiv \Phi(z)$.

→ Or, $\lim_{n\to\infty} \mathrm{P}(\frac{1}{n}\,S_n \le \mu + \frac{\sigma}{\sqrt{n}}z) = \mathrm{P}(Z \le z) \equiv \Phi(z)$. (e.g. $z = 0$: lim $= 1/2$)

→ Equivalently, $\frac{S_n - n\mu}{\sqrt{n}\sigma} \approx Z$, and $\frac{1}{n}\,S_n \approx \mu + \frac{\sigma}{\sqrt{n}}Z$, where $Z \sim \mathrm{Normal}(0,1)$.

→ So, not only does $\frac{1}{n}S_n$ converge to μ (which we already knew from the Laws of Large Numbers), but its deviations from μ are $O(1/\sqrt{n})$, with normal probabilities.

• Idea of proof: Use "moment-generating functions". (Textbook: Section 3.4.)

→ For any random variable X, its moment-generating function is the function $m_X(s)$ defined by $m_X(s) = \mathrm{E}[e^{sX}]$ for all $s \in \mathbf{R}$.

→ Assume that $m_X(s) < \infty$, at least in a neighbourhood of $s = 0$.

→ (If not, can instead use the characteristic function $c_X(s) = \mathrm{E}[e^{isX}]$ where $i = \sqrt{-1}$ ... similar but more complicated ...)

→ Useful properties, e.g. $m'_X(s) = \frac{d}{ds}m_X(s) = \frac{d}{ds}\mathrm{E}[e^{sX}] = \mathrm{E}[\frac{\partial}{\partial s}e^{sX}] = \mathrm{E}[Xe^{sX}]$, so $m'_X(0) = \mathrm{E}[X]$. Similarly $m''_X(0) = \mathrm{E}[X^2]$, $m'''_X(0) = \mathrm{E}[X^3]$, and in general for any $k \in \mathbf{N}$ we have $m^{(k)}_X(0) = \mathrm{E}[X^k]$. ("moments")

→ We need one key property: If $\lim_{n\to\infty} m_{X_n}(s) = m_X(s)$ for all s, at least in a neighbourhood of $s = 0$, then for all $x \in \mathbf{R}$, $\lim_{n\to\infty} \mathrm{P}(X_n \le x) = \mathrm{P}(X \le x)$, i.e.

$\lim_{n\to\infty} F_{X_n}(x) = F_X(x)$, i.e. X_n converges to X in distribution.

→ Oh, also, if X and Y are <u>independent</u>, then $m_{X+Y}(s) = \mathrm{E}[e^{s(X+Y)}] = \mathrm{E}[e^{sX}\, e^{sY}] = \mathrm{E}[e^{sX}]\,\mathrm{E}[e^{sY}] = m_X(s)\, m_Y(s)$.

- So, how can we prove the Central Limit Theorem?

  → Show that $\mathrm{E}(e^{s\,Z_n}) \to \mathrm{E}(e^{sZ})$ for all $s \in \mathbf{R}$, where $Z \sim \mathrm{Normal}(0,1)$.

- For starters, if $Z \sim \mathrm{Normal}(0,1)$, then $m_Z(s) = \mathrm{E}[e^{sZ}] = \int_{-\infty}^{\infty} e^{sz}\, \frac{1}{\sqrt{2\pi}}\, e^{-z^2/2}\, dz = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{sz-(z^2/2)}\, dz = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-(z-s)^2/2+(s^2/2)}\, dz = e^{s^2/2}\, \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-(z-s)^2/2}\, dz$.

  → $\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-(z-s)^2/2}\, dz = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-w^2/2}\, dw = 1$, so $m_Z(s) = e^{s^2/2}\,(1) = e^{s^2/2}$.

- So, we need to show that $m_{Z_n}(s) := \mathrm{E}(e^{s\,Z_n}) \to e^{s^2/2}$ for all $s \in \mathbf{R}$.

- Let $Y_i = (X_i - \mu)/\sigma$, so also i.i.d., with $\mathrm{E}(Y_i) = 0$, and $\mathrm{Var}(Y_i) = \sigma^2/\sigma^2 = 1$.

  → Then $Z_n = \frac{S_n - n\mu}{\sqrt{n}\,\sigma} = \frac{(X_1+X_2+\ldots+X_n)-n\mu}{\sqrt{n}\,\sigma} = \frac{1}{\sqrt{n}}(Y_1 + Y_2 + \ldots + Y_n)$.

  → So, $m_{Z_n}(s) = m_{\frac{1}{\sqrt{n}}(Y_1+Y_2+\ldots+Y_n)}(s) = m_{\frac{1}{\sqrt{n}}Y_1}(s) \ldots m_{\frac{1}{\sqrt{n}}Y_n}(s)$.

  → Then, since $\{Y_n\}$ are i.i.d., $m_{Z_n}(s) = \left[m_{\frac{1}{\sqrt{n}}Y_1}(s)\right]^n$.

  → But $m_{\frac{1}{\sqrt{n}}Y_1}(s) = \mathrm{E}[e^{s(\frac{1}{\sqrt{n}}Y_1)}] = \mathrm{E}[e^{(s/\sqrt{n})Y_1}] = m_{Y_1}(s/\sqrt{n})$.

  → So, $m_{Z_n}(s) = \left[m_{\frac{1}{\sqrt{n}}Y_1}(s)\right]^n = \left[m_{Y_1}(s/\sqrt{n})\right]^n$.

- Now, $m_{Y_1}(0) = \mathrm{E}[e^{0\,Y_1}] = \mathrm{E}[e^0] = 1$.

  → And, $m'_{Y_1}(0) = \mathrm{E}[Y_1] = 0$.

  → And, $m''_{Y_1}(0) = \mathrm{E}[(Y_1)^2] = \mathrm{Var}(Y_1) = 1$.

  → Then we can use a Taylor series expansion around $s = 0$:

  → For <u>small</u> s, $m_{Y_1}(s) \approx 1 + 0 \cdot s + 1 \cdot \frac{s^2}{2!} + O(s^3) \approx 1 + \frac{s^2}{2} + O(s^3)$.

  → Hence, as $n \to \infty$, $m_{Y_1}(s/\sqrt{n}) \approx 1 + \frac{(s/\sqrt{n})^2}{2} = 1 + \frac{s^2}{2n} + O(n^{-3/2})$.

  → So, $m_{Z_n}(s) = \left[m_{Y_1}(s/\sqrt{n})\right]^n \approx \left[1 + \frac{s^2}{2n} + O(n^{-3/2})\right]^n$.

- Finally, for any $a \in \mathbf{R}$, as $n \to \infty$, $[1 + \frac{a}{n}]^n \to e^a$.

  → Hence, $m_{Z_n}(s) = \left[m_{Y_1}(s/\sqrt{n})\right]^n \approx \left[1 + \frac{s^2}{2n}\right]^n \to e^{s^2/2}$, as required. ∎

END MONDAY #11

Normal Approximations

- Okay, so we know that as $n \to \infty$, $\mathrm{P}(\frac{S_n - n\mu}{\sqrt{n}\,\sigma} \le z) \to \Phi(z)$.

- Hence, for "reasonably large" n, we must have $\mathrm{P}(\frac{S_n - n\mu}{\sqrt{n}\,\sigma} \le z) \approx \Phi(z)$.

  → How large? Depends on the distribution of the X_i.

  → Rough "rule of thumb": Pretty good approximation if $n \ge 30$ ...

- Example: Suppose $\{X_n\}$ are i.i.d. $\sim$ Poisson(4).

  → What is a good approximation to $\mathrm{P}(X_1 + X_2 + \ldots + X_{900} \ge 3700)$?

  → Here $\mu := \mathrm{E}(X_i) = \lambda = 4$, and $\sigma := \mathrm{Sd}(X_i) = \sqrt{\mathrm{Var}(X_i)} = \sqrt{\lambda} = 2$.

→ Let $S_{900} = X_1 + X_2 + \ldots + X_{900}$.

→ Then $\mathrm{P}(X_1 + X_2 + \ldots + X_{900} \geq 3700) = \mathrm{P}(S_{900} \geq 3700)$

$$= \mathrm{P}\left(\frac{S_{900} - 900(4)}{\sqrt{900}\,(2)} \geq \frac{3700 - 900(4)}{\sqrt{900}\,(2)}\right) = \mathrm{P}\left(\frac{S_{900} - 900(4)}{\sqrt{900}\,(2)} \geq 5/3\right)$$

$$= \mathrm{P}\,(Z_{900} \geq 5/3) \approx \mathrm{P}\,(Z \geq 5/3) = \mathrm{P}\,(Z \leq -5/3) = \Phi(-5/3) \doteq 0.0478\,.$$

→ Here the value of $\Phi(-5/3)$ can found from software [e.g. "pnorm(−5/3)" in R], or from a table like Table D.2. (Both use numerical integration.)

→ [On an exam, if there is no table, you could just leave it as "$\Phi(-2)$".]

- Example: Suppose $\{X_n\}$ are independent, each $\sim$ Uniform$[2, 5]$.

→ What is a good approximation to $\mathrm{P}(X_1 + X_2 + \ldots + X_{400} \leq 1420)$?

→ Here $\mu := \mathrm{E}(X_i) = (2 + 5)/2 = 3.5$, and $\sigma := \mathrm{Sd}(X_i) = \sqrt{\mathrm{Var}(X_i)} = \sqrt{(5-2)^2/12} \doteq 0.866$.

→ Let $S_{400} = X_1 + X_2 + \ldots + X_{400}$.

→ Hence, $\mathrm{P}(X_1 + X_2 + \ldots + X_{400} \leq 1420) = \mathrm{P}(S_{400} \leq 1420)$

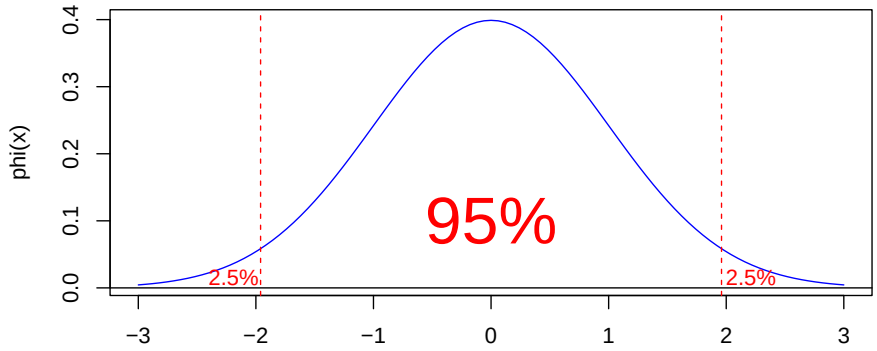
$$= \mathrm{P}\left(\frac{S_{400} - 400(3.5)}{\sqrt{400}\,(0.866)} \leq \frac{1420 - 400(3.5)}{\sqrt{400}\,(0.866)}\right) \doteq \mathrm{P}\left(\frac{S_{400} - 400(3.5)}{\sqrt{400}\,(0.866)} \leq 1.15\right)$$

$$\approx \mathrm{P}\,(Z \leq 1.15) = \Phi(1.15) = 1 - \Phi(-1.15) \doteq 1 - 0.1251 = 0.8749\,.$$

Suggested Homework: 4.4.5, 4.4.6, 4.4.7, 4.4.12, 4.4.13, 4.4.22, 4.4.23.

Estimation and Confidence Intervals

- Fact: $\Phi(-1.96) \doteq 0.025$. So, if $Z \sim$ Normal$(0, 1)$, then $\mathrm{P}(Z \leq -1.96) \doteq 0.025$, and $\mathrm{P}(Z \geq +1.96) \doteq 0.025$, so $\mathrm{P}(-1.96 \leq Z \leq +1.96) \doteq 1 - 0.025 - 0.025 = 0.95$:



→ That is, Z will be between -1.96 and $+1.96$ with probability 0.95, or 95%, or "19 times out of 20".

- So, if $\frac{S_n - n\mu}{\sqrt{n}\sigma} \approx Z$, then $\mathrm{P}(-1.96 \leq \frac{S_n - n\mu}{\sqrt{n}\sigma} \leq +1.96) \approx 0.95$, too.
- Probability interpretation: $\mathrm{P}(n\mu - 1.96\sqrt{n}\sigma \leq S_n \leq n\mu + 1.96\sqrt{n}\sigma) \approx 0.95$.

→ Tells us the probabilities for S_n, if we know μ and σ.

- e.g. If $\{X_n\}$ i.i.d. $\sim$ Exponential(5), then $\mu = 1/5$ and $\sigma = 1/5$, so if $S_n = X_1 + X_2 + \ldots + X_n$, then $\mathrm{P}(\frac{1}{5}(n - 1.96\sqrt{n}) \leq S_n \leq \frac{1}{5}(n + 1.96\sqrt{n}) \approx 0.95$.

→ So e.g. with $n = 200$, we get $\mathrm{P}(34.45 \le X_1 + X_2 + \ldots + X_{200} \le 45.54) \approx 0.95$.

→ That is, $X_1 + X_2 + \ldots + X_{200}$ will "usually" be in the interval [34.5, 45.5].

→ Try it in R: `sum( rexp(200,5) )`

- Statistics interpretation: $\mathrm{P}(\frac{1}{n}S_n - 1.96\frac{\sigma}{\sqrt{n}} \le \mu \le \frac{1}{n}S_n + 1.96\frac{\sigma}{\sqrt{n}}) \approx 0.95$.

→ Different perspective: Trying to "estimate" μ, if we know S_n (and σ?).

→ Statistics: Observe the variable values, then estimate the parameter(s).

→ By LLN, a good estimate of μ is $M_n := \frac{1}{n}S_n$. But how accurate is it?

- Well, if $M_n := \frac{1}{n}S_n$, then $\mathrm{P}(M_n - 1.96\frac{\sigma}{\sqrt{n}} \le \mu \le M_n + 1.96\frac{\sigma}{\sqrt{n}}) \approx 0.95$.

→ Sometimes write $\overline{X}_n := \frac{1}{n}S_n$, so $\mathrm{P}(\overline{X}_n - 1.96\frac{\sigma}{\sqrt{n}} \le \mu \le \overline{X}_n + 1.96\frac{\sigma}{\sqrt{n}}) \approx 0.95$.

- Example: Suppose $X_1, X_2, \ldots, X_{500} \sim \mathrm{Uniform}[a-1, a+1]$.

→ Suppose we observe the values $X_1, X_2, \ldots, X_{500}$, but a is unknown.

→ Well, here $n = 500$, and $\mu = \mathrm{E}[X_i] = [(a-1) + (a+1)]/2 = a$.

→ Also $\sigma = \mathrm{Sd}(X_i) = \sqrt{[R-L]^2/12} = \sqrt{[(a+1) - (a-1)]^2/12} = \sqrt{1/3} \doteq 0.577$.

→ But if $M_n := \frac{1}{n}S_n$, then $\mathrm{P}(M_n - 1.96\frac{\sigma}{\sqrt{n}} \le \mu \le M_n + 1.96\frac{\sigma}{\sqrt{n}}) \approx 0.95$.

→ Hence, $\mathrm{P}(M_{500} - 1.96\,\frac{0.577}{\sqrt{500}} \le a \le M_{500} + 1.96\,\frac{0.577}{\sqrt{500}}) \approx 0.95$.

→ That is, $\mathrm{P}(M_{500} - 0.051 \le a \le M_{500} + 0.051) \approx 0.95$.

→ Hence, a will "usually" be in the interval $[M_{500} - 0.051,\ M_{500} + 0.051]$.

- In the above example, suppose we observe that $X_1 + X_2 + \ldots + X_{500} = 29$.

→ Then $M_{500} = \frac{29}{500} \doteq 0.058$, so $[M_{500} - 0.051,\ M_{500} + 0.051] = [0.007,\ 0.109]$.

→ Can we say that $\mathrm{P}(0.007 \le a \le 0.109) \approx 0.95$?

→ Not really, since a is not random (just unknown) – so no probabilities!

→ And yet, we're still fairly "confident" that a is in $[0.007,\ 0.109]$.

→ Here, $[0.007,\ 0.109]$ is called a 95% confidence interval for a.

→ [Aside: Alternative "Bayesian" perspective treats parameters like a as random.]

- In general, recall that $\mathrm{P}(\frac{1}{n}S_n - 1.96\frac{\sigma}{\sqrt{n}} \le \mu \le \frac{1}{n}S_n + 1.96\frac{\sigma}{\sqrt{n}}) \approx 0.95$.

→ Hence, $[\frac{1}{n}S_n - 1.96\frac{\sigma}{\sqrt{n}},\ \frac{1}{n}S_n + 1.96\frac{\sigma}{\sqrt{n}}]$ is a 95% confidence interval for μ.

- The value 95% is "usual", but other values are also possible. (e.g. 99%, etc.)

→ e.g. $\Phi(-3) \doteq 0.00135$, so $\mathrm{P}(-3 \le Z \le 3) \doteq 1 - 0.00135 - 0.00135 = 0.9973$.

→ So, $\mathrm{P}(\frac{1}{n}S_n - 3\frac{\sigma}{\sqrt{n}} \le \mu \le \frac{1}{n}S_n + 3\frac{\sigma}{\sqrt{n}}) \approx 0.9973$. (textbook: "virtual certainty")

→ Hence, $[\frac{1}{n}S_n - 3\frac{\sigma}{\sqrt{n}},\ \frac{1}{n}S_n + 3\frac{\sigma}{\sqrt{n}}]$ is a 99.73% confidence interval for μ.

- Suppose now that $Y \sim \mathrm{Binomial}(n, \theta)$.

→ Then we can think of Y as $Y = X_1 + X_2 + \ldots + X_n$ where each $X_i \sim \mathrm{Bernoulli}(\theta)$ and they are independent. (e.g. $X_i = 1$ if you score on the i^{th} free throw, otherwise 0)

→ So, $M_n = \frac{1}{n}Y$, and $\mu = \theta$, and $\sigma = \sqrt{\theta(1-\theta)}$.

→ Suppose θ is unknown. 95% confidence interval?

→ Well, we know that $\mathrm{P}(M_n - 1.96\frac{\sigma}{\sqrt{n}} \le \mu \le M_n + 1.96\frac{\sigma}{\sqrt{n}}) \approx 0.95$.

→ That is, $\mathrm{P}(M_n - 1.96\sqrt{\theta(1-\theta)/n} \le \theta \le M_n + 1.96\sqrt{\theta(1-\theta)/n}) \approx 0.95$.

→ So, $[M_n - 1.96\sqrt{\theta(1-\theta)/n},\ M_n + 1.96\sqrt{\theta(1-\theta)/n}]$ is 95% confidence interval.

→ Problem: θ is unknown! What to do?

→ Usual solution: By LLN, probably $M_n \approx \theta$. So, approximate the true standard deviation $\sigma = \sqrt{\theta(1-\theta)}$ by the estimate $\sigma_n := \sqrt{M_n(1-M_n)}$. ("standard error")

→ So, use the interval $[M_n - 1.96\sqrt{M_n(1-M_n)/n},\ M_n + 1.96\sqrt{M_n(1-M_n)/n}]$.

→ [Aside: sometimes "standard error" is taken to mean the estimate of σ divided by $\sqrt{n}$, e.g. $\sqrt{M_n(1-M_n)/n}$. So, best to just say "estimate of σ".]

- Aside. Alternative solution: always have $\theta(1-\theta) \le (1/2)(1/2) = 1/4$.

→ So, use the "conservative" interval $[M_n - 1.96/2\sqrt{n},\ M_n + 1.96/2\sqrt{n}]$ instead.

→ (Here "conservative" means the interval is a little bit larger than necessary, i.e. the probability that μ will be within the interval is a little bit more than 95%. So, the interval is slightly "wasteful", but still okay and useful, and more reliable.)

- Now, the above discussion is in terms of general n and S_n (or $M_n := S_n/n$).

→ If we observe a specific value of S_n for some specific n, then we can get a specific quantitative confidence interval.

- Example: Suppose you're shooting free throws, and score 86 out of 250 of them.

→ The number of scores is $S_{250} \sim \text{Binomial}(250, \theta)$, with θ unknown.

→ Here $n = 250$, and $\mu = \theta$ (unknown).

→ Also $\sigma = \sqrt{\theta(1-\theta)}$, unknown. (But $\le 1/2$.)

→ So, if $M_n := \frac{1}{n}S_n$, then $\mathrm{P}(M_n - 1.96\frac{\sigma}{\sqrt{n}} \le \theta \le M_n + 1.96\frac{\sigma}{\sqrt{n}}) \approx 0.95$.

→ Hence, $\mathrm{P}(M_{250} - 1.96\sqrt{\theta(1-\theta)/250} \le \theta \le M_{250} + 1.96\sqrt{\theta(1-\theta)/250}) \approx 0.95$.

→ 95% confidence interval: $[M_{250} - 1.96\sqrt{\theta(1-\theta)/250},\ M_{250} + 1.96\sqrt{\theta(1-\theta)/250}]$.

→ Usual solution: $\theta \approx 86/250 \doteq 0.344$, so $\theta(1-\theta) \doteq 0.344(1-0.344) \doteq 0.226$.

→ Then $M_{250} - 1.96\sqrt{\theta(1-\theta)/250} \doteq (86/250) - 1.96\sqrt{0.226/250} \doteq 0.285$.

→ And, $M_{250} + 1.96\sqrt{\theta(1-\theta)/250} \doteq (86/250) + 1.96\sqrt{0.226/250} \doteq 0.403$.

→ Hence, $[0.285,\ 0.403]$ is a 95% confidence interval for θ.

- Aside. Alternative conservative solution in above example: Use that $\theta(1-\theta) \le 1/4$.

→ So, $M_{250} - 1.96\sqrt{\theta(1-\theta)/250} \ge (86/250) - 1.96/2\sqrt{250} \doteq 0.282$.

→ And, $M_{250} + 1.96\sqrt{\theta(1-\theta)/250} \le (86/250) - 1.96/2\sqrt{250} \doteq 0.406$.

→ Hence, $[0.282,\ 0.406]$ is a "conservative" 95% confidence interval for θ.

Suggested Homework: 4.5.4, 4.5.7, 4.5.8, 4.5.9, 4.5.10, and the following.
Q1. Suppose $Y \sim \text{Binomial}(600, \theta)$, where θ is unknown. Suppose we observe that there were 483 out of 600 successes. Based on these observations, compute a 95% confidence interval for θ, and also a 99.73% confidence interval for θ.
Q2. Suppose $\{X_n\}$ are i.i.d. $\sim \text{Uniform}[\mu - 5, \mu + 5]$, where μ is unknown. Compute a 95% confidence interval for μ, both:
(a) in terms of general n and S_n.
(b) based on the observation that $X_1 + X_2 + \ldots + X_{64} = 300$.

Q3. Suppose $\{X_n\}$ are i.i.d. $\sim$ Exponential(λ), where λ is unknown. Compute a 95% confidence interval for λ. [Hint: What is μ? And, how to approximate σ?]
Q4. Suppose $\{X_n\}$ are i.i.d. $\sim$ Poisson(λ), where λ is unknown. Compute a 95% confidence interval for λ. [Recall that Poisson(λ) has mean λ and variance λ.]
Q5. Suppose $\{X_n\}$ are i.i.d. $\sim$ Uniform$[0, a]$, where a is unknown. Compute a 95% confidence interval for a. [Hint: What are μ and σ in terms of a?]

Monte Carlo Approximations

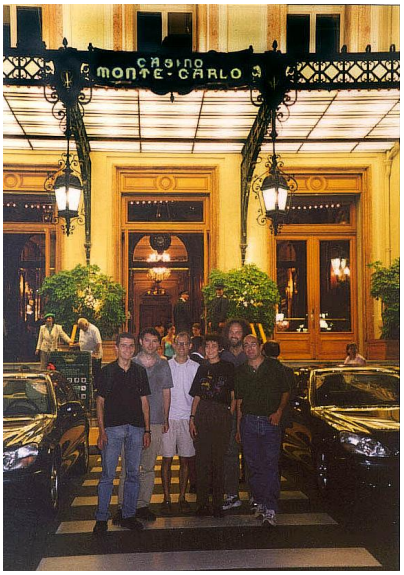
- e.g. Suppose $U \sim$ Uniform$[0, 1]$. What is $\mu := \mathrm{E}\Big(U^3\big[\sin(U^4) + \cos(U^5)\big]e^{-U^6}\Big)$?

→ In principle, this equals $\int_0^1 u^3[\sin(u^4) + \cos(u^5)]e^{-u^6}\,du$. How to compute??

→ One method: Use a "Monte Carlo algorithm". What is that?

→ A wealthy region in Monaco with yachts and a big casino?

→ A nice place for a conference?



→ Well, yes ... but also a method of computing by using randomness.

→ To compute $\mu := \mathrm{E}\big(U^3[\sin(U^4) + \cos(U^5)]e^{-U^6}\big)$, first generate i.i.d. random values $U_1, U_2, \ldots, U_n \sim$ Uniform$[0, 1]$ on a computer.

→ Then set $X_i = U_i^3[\sin(U_i^4) + \cos(U_i^5)]e^{-U_i^6}$, for $i = 1, 2, 3, \ldots$.

→ Since the $\{U_i\}$ are i.i.d., therefore the $\{X_i\}$ are i.i.d. too.

→ Now, recall that $\mathrm{E}[g(X)] = \int_{-\infty}^{\infty} g(x)\, f_X(x)\, dx$. Hence, $\mathrm{E}(X_i) := \mathrm{E}\big(U_i^3[\sin(U_i^4)+\cos(U_i^5)]e^{-U_i^5}\big)] = \int_0^1 u^3[\sin(u^4) + \cos(u^5)]e^{-u^6}\,(1)\, du \equiv \mu$ for each i.

→ Hence, if $M_n = \frac{1}{n}S_n := \frac{1}{n}(X_1 + X_2 + \ldots + X_n)$, then $M_n \approx \mu$ for large n.

→ That is, M_n (observed) is a good estimate of μ (unknown).

→ I ran it in R, with $n = 50,000$:

```
U = runif(50000); sum( U^3*(sin(U^4)+cos(U^5))*exp(-U^6) ) / 50000
```

→ I got $S_{50000} = 11319.6$, which gives estimate $= M_n = 11319.6/50000 \doteq 0.2264$.

→ Accurate??

- Well, $[\frac{1}{n}S_n - 1.96\frac{\sigma}{\sqrt{n}},\ \frac{1}{n}S_n + 1.96\frac{\sigma}{\sqrt{n}}]$ is a 95% confidence interval for μ.

→ σ unknown, but $|X_n| \le 2$, so $\sigma^2 := \mathrm{Var}(X_n) \le \mathrm{E}[(X_n)^2] \le 4$, and $\sigma \le 2$.

→ So, $[\frac{1}{n}S_n - 1.96\frac{2}{\sqrt{n}},\ \frac{1}{n}S_n + 1.96\frac{2}{\sqrt{n}}]$ is a 95% confidence interval.

→ In our case, this works out to:
$= [\frac{1}{50000}(11319.6) - 1.96\frac{2}{\sqrt{50000}},\ \frac{1}{50000}(11319.6) + 1.96\frac{2}{\sqrt{50000}}] \doteq [0.209, 0.244]$.

→ So, 95% confident that $\mu := \mathrm{E}[U^3(\sin(U^4) + \cos(U^5))e^{-U^6}] \in [0.209, 0.244]$.

→ Of course, μ isn't really random. Good estimate? Inside interval??

→ Numerical integration in *Mathematica*: $\mu \doteq 0.2258 \approx M_n$. Yes, inside! Good!

- Can also use Monte Carlo to estimate the value of integrals!

→ Idea: first re-write the integral as an expected value.

- e.g. Compute $I := \int_0^1 e^{\cos(x)}\, dx$.

→ Use calculus? Too hard! (No closed-form solution?)

→ Instead, note that $I = \mathrm{E}[e^{\cos(U)}]$ where $U \sim \mathrm{Uniform}[0, 1]$.

→ So, as before, first generate random i.i.d. values $U_1, U_2, \ldots, U_n \sim \mathrm{Uniform}[0, 1]$.

→ Then set $X_i = e^{\cos(U_i)}$, so $\mu := \mathrm{E}[X_i] = I$. And $\sigma \le \sqrt{\mathrm{E}[(X_i)^2]} \le \sqrt{e^2} = e$.

→ Then $\frac{1}{n}S_n \approx \mu$, so $\frac{1}{n}S_n$ gives a good estimate of I.

→ And, $[\frac{1}{n}S_n - 1.96\frac{e}{\sqrt{n}},\ \frac{1}{n}S_n + 1.96\frac{e}{\sqrt{n}}]$ is a conservative 95% conf. int. for I.

- Many other integrals can also be converted to expected values:

→ e.g. $\int_5^8 \cos(x^7)\, dx = \int_5^8 [3\, \cos(x^7)]\, \frac{1}{3}\, dx = \mathrm{E}[3\, \cos(X^7)]$ where $X \sim \mathrm{Uniform}[5, 8]$.

→ e.g. $\int_0^\infty \cos(x^7)\, e^{-5x}\, dx = \int_0^\infty [\frac{1}{5}\, \cos(x^7)]\, 5e^{-5x}\, dx = \mathrm{E}[\frac{1}{5}\, \cos(Y^7)]$ where $Y \sim \mathrm{Exponential}(5)$.

→ e.g. $\int_{-\infty}^\infty \cos(x^7)\, e^{-x^2/2}\, dx = \int_{-\infty}^\infty [\sqrt{2\pi}\, \cos(x^7)]\, \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx = \mathrm{E}[\sqrt{2\pi}\, \cos(Z^7)]$ where $Z \sim \mathrm{Normal}(0, 1)$.

→ And sums too, e.g. $\sum_{j=0}^\infty \cos(j^7)\, (2/3)^j = \sum_{j=0}^\infty [3\, \cos(j^7)]\, [1 - (1/3)]^j (1/3) = \mathrm{E}[3\, \cos(X^7)]$ where $X \sim \mathrm{Geometric}(1/3)$.

→ etc. And then each one can be approximated by similar Monte Carlo, too!

Suggested Homework: 4.5.1, 4.5.2, 4.5.3, 4.5.5, 4.5.6, 4.5.11, 4.5.12.

World's Oldest Monte Carlo: Buffon's Needle

- A Monte Carlo method from 1733, to compute the value of π!
- Suppose we toss a needle randomly onto a lined surface.
  - → Suppose the needle length L is equal to the space between the lines.
  - → Try it out in R: `source("http://probability.ca/mc/Rbuffon"); buffon()`
- What is the probability that a needle will touch a line?
  - → Well, let α be the angle that the needle makes with the line direction.
  - → Then in terms of α, the needle covers vertical distance $L\sin(\alpha)$.
  - → So, the probability it touches a line is $\frac{L\sin(\alpha)}{L} = \sin(\alpha)$.
  - → e.g. If $\alpha = 0°$, then prob $= 0$. If $\alpha = 90°$, prob $= 1$. If $\alpha = 30°$, prob $= 1/2$.
  - → But this depends on α, which is random. Need to average!
- That is, the probability that the needle will touch the line is equal to the average value of $\sin(\alpha)$, as α ranges over all of its possible (random) values.
  - → Here $\alpha \sim \text{Uniform}[0°, 180°]$, i.e. $\alpha \sim \text{Uniform}[0, \pi]$ in radians.
  - → So, $\text{P}(\text{needle touches line}) = \text{E}[\sin(\alpha)] = \frac{1}{\pi}\int_0^\pi \sin(x)\,dx = \frac{1}{\pi}[-\cos(x)]\Big|_{x=0}^{x=\pi}$ $= \frac{1}{\pi}[-\cos(\pi) + \cos(0)] = \frac{1}{\pi}[-(-1) + (1)] = 2/\pi$. (Depends on π!)
- Suppose we throw a large number N of needles, of which M touch a line.
  - → Then, we know that each one had success probability $\theta = 2/\pi$.
  - → So, for large N, we should have $M/N \approx \theta = 2/\pi$.
  - → This means that $\pi \approx 2N/M$, so $2N/M$ is a possible estimate of π.
  - → This is a Monte Carlo method to approximately compute π!
  - → Try it out in R: `source("http://probability.ca/mc/Rbuffon"); buffon()`
- First proposed by George-Louis Leclerc, Comte de Buffon, back in 1733 (!).
- In 1864, injured civil war Captain O.C. Fox experimented three times:
  - → #1: N=500, est=3.1780. #2: N=530, est=3.1423. #3: N=590, est=3.1416.
- (See the textbook Challenge 4.5.25 and Discussion 4.5.28.)
- Aside: There are other, better ways to estimate π:
  - → $\pi/4 = \arctan(1) = 1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \ldots$. [trigonometry / calculus]
  - → $\pi = 3 + \frac{4}{2\cdot3\cdot4} + \frac{4}{4\cdot5\cdot6} + \frac{4}{6\cdot7\cdot8} + \frac{4}{8\cdot9\cdot10} + \ldots$. [Nilakantha, India, 1444–1550]
  - → But Buffon's Needle is more fun. And it uses probabilities!

Distributions Related to the Normal

- Because of the CLT, the normal distribution is extremely important!
  - → Nearly everything becomes approximately normal for large n.

- So, other distributions related to the normal also become important:

- If $X_1, X_2, \ldots, X_n \sim \text{Normal}(0,1)$ are i.i.d., then the distribution of their sum of squares $X_1^2 + X_2^2 + \ldots + X_n^2$ is called the chi-squared distribution with n degrees of freedom, also written $\chi^2(n)$.

- If $Z, X_1, X_2, \ldots, X_n \sim \text{Normal}(0,1)$ are i.i.d., the distribution of $\frac{Z}{\sqrt{(X_1^2+X_2^2+\ldots+X_n^2)/n}}$ is called the t-distribution with n "degrees of freedom", sometimes written $t(n)$.

- If $X_1, X_2, \ldots, X_m, Y_1, Y_2, \ldots, Y_n, \sim \text{Normal}(0,1)$ are i.i.d., then the distribution of $\frac{(X_1^2+X_2^2+\ldots+X_m^2)/m}{(Y_1^2+Y_2^2+\ldots+Y_n^2)/n}$ is called the F-distribution with m and n degrees of freedom.

- The above distributions all have corresponding densities, and expected values, and variances, and various interesting properties. (See textbook Section 4.6.)

  → And their probabilities can be computed by statistical software (e.g. R).

  → And some statistics textbooks even have tables of their values.

  → And they are used for lots of statistical tests and analyses. (See e.g. the second half of the textbook, and the follow-up course STA261.)

END MONDAY #12

Final Announcements

- **No lecture** this Wednesday. (I will still come to class in case you have questions.)
- **Please complete the online course evaluation!**
- During the coming days: TA tutorials and office hours (and Piazza).
- Instructor Office Hours: Fri Dec 8 from 1:10 to 2:30 in SS 2125. [Exam Jam]
- **AND MOST IMPORTANT OF ALL:**

FINAL EXAM: Sat Dec 9 from 2:00 to 5:00 pm, in **NEW ROOMS** by Last Name:

**** MP 102 A–HU; MP 103 HUA–NA; MP 202 NE–WA; MP 203 WE–Z.**

All in the **MP (physics) building**, NOT in other buildings!

******* Good luck on the exam, and with all of your future studies! *******