STA257 (Probability and Statistics I) Lecture Notes, Fall 2024

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

(Last updated: November 27, 2024)

Note: I will update these notes regularly, posting them on the course web page each evening after lectures (though without annotations). 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, and doing the suggested homework exercises.

Introduction

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

• Register for PollEverywhere: probability.ca/sta257/pollinfo.html **USE UofT EMAIL!**

• 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 the most likely? How 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 time you roll a 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 find solutions to these questions (later!), and mathematical probability theory was born!

POLL: If you have independent probability 1/2 of winning each game, and you are up 3 games to 1, what do you think is the probability that you will win 4 games first?
(A) 1/2. **(B)** 2/3. **(C)** 3/4. **(D)** 7/8. **(E)** No idea. [Best guess only – later.]

- 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 (§1.2) (i.e. Section 1.2 of the textbook)

- 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. Canada's next Olympic medal: $S = \{\text{Gold, Silver, Bronze}\}$.

→ 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: $S = (0, \infty)$.

→ 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 (§1.2)

- 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. Olympic medal: maybe P(Gold)=0.40, P(Silver)=0.15, and P(Bronze)=0.45.

→ (Note: We could also write P(Bronze) = 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" ("Uniform[0,1]"):
$\mathrm{P}([0, 1/2]) = 1/2$, $\mathrm{P}([1/2, 1]) = 1/2$, $\mathrm{P}([0, 1/3]) = 1/3$, $\mathrm{P}([1/3, 2/3]) = 1/3$,
and in general $\mathrm{P}([a, b]) = b - a$ whenever $0 \le a \le b \le 1$. Diagram:

Basic Properties of Probabilities (§1.2)

- Let's begin with a specific example (and then we will generalise):

- e.g. Olympic medal, with P(Gold)=0.40, P(Silver)=0.15, and P(Bronze)=0.45.

→ Probability of Gold or Silver = P({Gold, Silver}) = P({Gold}) + P({Silver}) $= 0.40 + 0.15 = 0.55$.

→ Probability of any medal = Probability of Gold or Silver or Bronze = P({Gold, Silver, Bronze}) = P({Gold}) + P({Silver}) + P({Bronze}) $= 0.40 + 0.15 + 0.45 = 1$.

→ Probability next medal not Gold nor Silver nor Bronze $= \mathrm{P}(\emptyset) = 0$.

- In general, certain properties must hold for any probability model ("axioms"):

- If A is an event, then $0 \le \mathrm{P}(A) \le 1$.

- 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. $A = \{\text{Gold}\}$ and $B = \{\text{Silver}\}$, then $P(A \cup B) = P(A) + P(B)$.

• 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 $P\left(\bigcup_i A_i\right) = \sum_i P(A_i)$.

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

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

→ e.g. P(Gold) + P(Silver) + P(Bronze) = 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, 1.2.13, 1.2.14, 1.2.15.

END WEDNESDAY #1

Derived Properties of Probabilities (§1.3)

• 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 $P(A^C) = 1 - P(A)$. (Important! Remember this! Use this!)

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

→ e.g. P(Bronze) = P(not Gold or Silver) = 1−P(Gold or Silver) = 1−0.55 = 0.45.

• Fact: For any events A and B, $P(A) = P(A \cap B) + P(A \cap B^C)$. (∗)
Diagram:

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

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

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

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

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

→ e.g. integer between 1 and 10: $P(\leq 7) = P(\leq 4) + P(\leq 7 \text{ but } \geq 5)$.

• Monotonicity: If $A \supseteq B$, then $P(A) \geq P(B)$. (Remember this!)

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

→ e.g. $P(\{\text{Gold, Silver}\}) = 0.55 \geq 0.40 = P(\{\text{Gold}\})$.

• 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, $\mathrm{P}(A \cup B) = \mathrm{P}(A \cap B) + \mathrm{P}(A \cap B^C) + \mathrm{P}(A^C \cap B)$.

But 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) = $\mathrm{P}(\{2, 4, 6, 8, 10\}) + \mathrm{P}(\{1, 4, 9\}) - \mathrm{P}(\{4\})$.

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

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

- 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 (§1.7): 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:

Then the $\{B_i\}$ are disjoint by construction, and $\bigcup_i B_i = \bigcup_i A_i$.
[Formally, the above construction ensures that $\bigcup_{i=1}^n B_i = \bigcup_{i=1}^n A_i$ for each finite n. Then, in the infinite case, $\bigcup_{i=1}^\infty B_i = \bigcup_{n=1}^\infty (\bigcup_{i=1}^n B_i) = \bigcup_{n=1}^\infty (\bigcup_{i=1}^n A_i) = \bigcup_{i=1}^\infty A_i$.]
Also $B_i \subseteq A_i$ so $\mathrm{P}(B_i) \le \mathrm{P}(A_i)$. Hence, $\mathrm{P}(A_1 \cup A_2 \cup \ldots) = \mathrm{P}(B_1 \cup B_2 \cup \ldots) = \mathrm{P}(B_1) + \mathrm{P}(B_2) + \ldots \le \mathrm{P}(A_1) + \mathrm{P}(A_2) + \ldots$. ■

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

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

[Note that we do not have "uncountable" subadditivity, e.g. for uniform on $S = [0,1]$, if $A_x = \{x\}$ for each $x \in S$, then $\mathrm{P}(\bigcup_{x \in S} A_x) = \mathrm{P}(S) = \mathrm{P}([0,1]) = 1$, even though $\mathrm{P}(A_x) = \mathrm{P}(\{x\}) = 0$ for each individual $x \in S$, so also $\sum_{x \in S} \mathrm{P}(A_x) = \sum_{x \in S}(0) = 0$.]

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.

Uniform Probabilities on Finite Spaces (§1.4)

• 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 $\mathrm{P}(\ge 5)$?

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

→ Also $A = \{5,6\}$ so $|A| = 2$.

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

• Flip two fair coins. What is $\mathrm{P}(\#\text{ Heads} = 1)$?

POLL: **(A)** 1/4. **(B)** 1/3. **(C)** 1/2. **(D)** 3/4. **(E)** 1. **(F)** No idea.

→ Here $S = \{HH, HT, TH, TT\}$, all equally likely. So, $|S| = 4$.

→ And, $A = \{HT, TH\}$. So, $|A| = 2$.

→ Hence, $\mathrm{P}(A) = |A| \,/\, |S| = 2/4 = 1/2$. Easy!

• e.g. Roll one fair six-sided die, and flip two fair coins.
What is P(# Heads = Number Showing On The Die)? (Best guess?)

POLL: **(A)** 1/6. **(B)** 1/8. **(C)** 1/12. **(D)** 1/16. **(E)** 1/24. **(F)** No idea.

→ 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) = $|A| \,/\, |S| = 3/24 = 1/8$.

→ [Alternatively (later): $(1/6)(1/2)+(1/6)(1/4) = (1/12)+(1/24) = 3/24 = 1/8$.]

- e.g. Roll three fair six-sided dice. What is P(sum ≥ 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, P(sum ≥ 17) = P(A) = $|A| \,/\, |S| = 4/216 = 1/54$.

→ Exercise: What about P(sum ≥ 16)? P(sum ≥ 15)?

- Chevalier de Méré's historical 1654 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 = {no sixes in four rolls} = $\{1,2,3,4,5\}^4$, so $|A^C| = 5^4 = 625$.

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

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

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

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

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

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

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

END MONDAY #1

- (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 unchanged.)

  → 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.)

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 $S = \{\text{all ordered pairs of two dice}\}$, 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!

- Also, note that 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.

  → But on a discrete infinite space, cannot ever have a uniform distribution!

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

More Finite Uniform Probabilities (§1.4)

- Distinct, in order: 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 $n(n-1)(n-2)\ldots(n-k+1) = n!/(n-k)!$. ("permutations")

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

→ (List their birthdays 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 $P(A) = 1 - P(A^C)$.)

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

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

→ Intuition: Even with just 40 people, have $\binom{40}{2} = 780$ pairs of people – lots!

→ Or, if 23 people, $P(A^C) = \big(365!/342!\big)\,/365^{23} \doteq 0.493$, so $P(A) \doteq 0.507 > 50\%$.

POLL: With 60 people, what is P(some pair have same birthday)? (guess)
(A) 92.8%. **(B)** 95.1%. **(C)** 99.4%. **(D)** 99.86%. **(E)** 99.993%.

→ With 60 people: $P(A^C) = \big(365!/305!\big)\,/365^{60} \doteq 0.059$; $P(A) \doteq 0.994 = 99.4\%$.

→ (For discussion with "C" people, see the textbook's Challenge 1.4.21.)

• Distinct, unordered: Suppose we are still picking k distinct objects, but now we don't care about the order. Then, we have to divide by the number of different orderings of k items, which is: $k! = k(k-1)(k-2)\ldots(2)(1)$.

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

POLL: 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)?
(A) 1/60. **(B)** 1/120. **(C)** 1/240. **(D)** 1/360. **(E)** 1/720. **(F)** No idea.

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

→ But what is $|S|$?

→ Here $|S| = \binom{10}{3} = \frac{10!}{7!\,3!} = 120$.

→ 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.

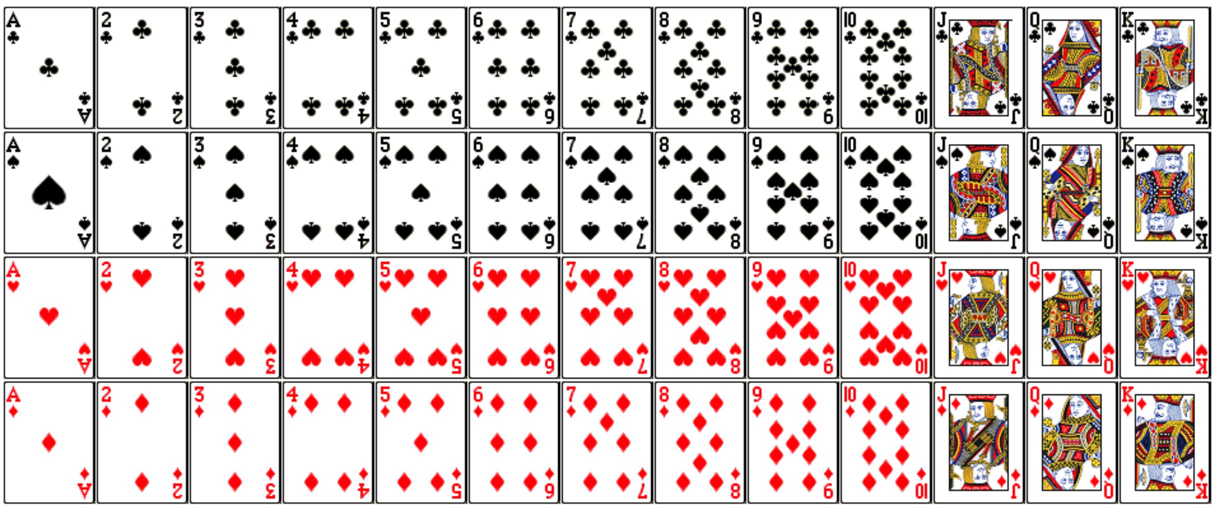
→ So, $|S| = \frac{50!}{43!\,7!} = 99,884,400 \doteq 100$ million.

→ 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 separate 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) $= 1/4 + 1/13 - 1/52 = 16/52 = 4/13$.

POLL: 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)?
(A) $(3/52)^2$. **(B)** $(12/52)^2$. **(C)** $12/\binom{52}{2}$. **(D)** $\binom{12}{2}/\binom{52}{2}$. **(E)** No idea.

→ 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, $P(A) = |A|/|S| = \binom{12}{2}/\binom{52}{2} = 66/1326 \doteq 0.0498 \doteq 1/20$.

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

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

Suggested Homework: 1.3.6, 1.4.4, 1.4.6, 1.4.7, 1.4.8. Trickier: 1.4.5.

END WEDNESDAY #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: 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) [three random samples, without replacement]

→ sample(1:6, 3, replace=TRUE) [three samples, with replacement]

→ sample(c("Gold","Silver","Bronze"), 1, prob=c(0.40,0.15,0.45)) [with probs]

→ rgeom(1, 1/2) + 1 [sample where $P(i) = 2^{-i}$]

A Bit More Finite Uniform Probabilities (§1.4)

POLL: Suppose we flip 4 fair coins. What is P(exactly 2 Heads)?
(A) 1/2. **(B)** 1/4. **(C)** 1/8. **(D)** 3/8. **(E)** 5/8. **(F)** No idea.

→ 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? (More coins?)

→ 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: 1.4.2, 1.4.3, 1.4.15, 1.4.16, 1.4.19, 1.4.21.

Conditional Probability (§1.5)

- 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.

POLL: Now what is the probability that the first coin was Heads?
(B) 1/2. **(C)** 2/3. **(D)** 3/4. **(E)** No idea.

→ 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 $P(A\,|\,B)$, and represents the fraction of the times when B occurs, in which A also occurs. [Diagram.] So, it is equal to:

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

- Note: If $\mathrm{P}(B) = 0$, then $\mathrm{P}(A \mid B)$ is …

undefined! It <u>only</u> 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 $B = \{\text{2 coins Heads}\}$.

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

<u>**POLL:**</u> Roll three fair six-sided dice. What is P(first die is 3 | at least one 3)? (guess)
(A) Less than 1/6. **(B)** 1/6. **(C)** Between 1/6 and 1/3. **(D)** More than 1/3.

→ 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, $\mathrm{P}(A \mid B) = \mathrm{P}(A \cap B)/\mathrm{P}(B) = (1/6)/(91/216) = (36/216)/(91/216) = 36/91 \doteq 0.396$. Much more than $1/6 \doteq 0.167$, or even $1/3 \doteq 0.333$. Surprising?

<u>**POLL:**</u> Roll three fair six-sided dice. What is P(first die is 3 | sum is ≤ 5)? (guess)
(A) Less than 1/6. **(B)** 1/6. **(C)** Between 1/6 and 1/3. **(D)** More than 1/3.

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

→ Here $A = \{\text{first die is 3}\}$, 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 $A \cap B$? Here $A \cap B = \{311\}$, so $P(A \cap B) = 1/216$.

→ Then $\mathrm{P}(A \mid B) = \mathrm{P}(A \cap B) \,/\, \mathrm{P}(B) = (1/216) / (10/216) = 1/10 = 10\% < 1/6$.

- Or, what is P(at least one 3 | sum is ≤ 5)?

→ Here $A = \{\text{at least one 3}\}$, and $B = \{\text{sum is} \leq 5\}$. So, $|B| = 10$ as above.

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

- Conditional Multiplication Formula: Since $\mathrm{P}(A \mid B) = \mathrm{P}(A \cap B)/\mathrm{P}(B)$, therefore $\mathrm{P}(A \cap B) = \mathrm{P}(B)\,\mathrm{P}(A \mid B)$. Similarly, $\mathrm{P}(A \cap B) = \mathrm{P}(A)\,\mathrm{P}(B \mid A)$. 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 $\text{P}(B \mid A)$?

→ 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, $\text{P}(B \mid A) = 11/51$.

→ Then $\text{P}(A \cap B) = \text{P}(A)\,\text{P}(B \mid A) = (12/52)\,(11/51)$. 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 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. Then $\text{P}(B) = \sum_i \text{P}(A_i)\ \text{P}(B \mid A_i)$, or equivalently $\text{P}(B) = \text{P}(A_1)\ \text{P}(B \mid A_1) + \text{P}(A_2)\ \text{P}(B \mid A_2) + \ldots$.

- 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 $\text{P}(A_1) = \text{P}(A_2) = 1/2$. Need $\text{P}(B \mid A_i)$.

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

→ Also, $\text{P}(B \mid A_2) =$ P(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, $\text{P}(B \mid A_2) = 1 - (25/36) = 11/36$.

→ Then, from the above Law of Total Probability,

$$\text{P}(B) = \sum_i \text{P}(A_i)\ \text{P}(B \mid A_i) = \text{P}(A_1)\ \text{P}(B \mid A_1) + \text{P}(A_2)\ \text{P}(B \mid A_2)$$

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

- 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 $\text{P}(B)$?

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

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

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

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

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

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

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

• Related question: The Monty Hall Problem!
See Challenge 1.5.18, and/or my article at probability.ca/monty.

• In the above "two Face Cards" question, suppose we ignore the first card. Then what is P(second card is Face Card)?

→ Well, if B = {second card is Face Card}, and A_1 = {first card is Face Card} and A_2 = {first card is NOT Face Card} then $\{A_1, A_2\}$ is a partition, so $P(B) = P(A_1)P(B \mid A_1) + P(A_2)P(B \mid A_2) = (12/52)(11/51) + (40/52)(12/51) = 12/52 = 3/13$, exactly the same as if it was the only card picked.

→ Makes sense, since ignoring the first card is the same as not picking it at all.

• e.g. Suppose a disease affects one person in a thousand, and a test for the disease has 99% accuracy.

→ This means that P(test positive | have disease) = 0.99, P(test negative | have disease) = 0.01, P(test positive | do NOT have disease) = 0.01, and P(test negative | do NOT have disease) = 0.99.

→ Suppose someone is selected at random, and is tested for the disease.

POLL: (i) What is P(they test positive)? **(A)** 1/1000. **(B)** (1/1000) (0.99). **(C)** (1/1000) (0.99) + (999/1000) (0.01). **(D)** (999/1000) (0.99) + (1/1000) (0.01).

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

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

POLL: (ii) What is P(they test positive and have the disease)?
(A) 1/1000. **(B)** (1/1000) (0.99). **(C)** (1/1000) (0.99) + (999/1000) (0.01).
(D) (999/1000) (0.99) + (1/1000) (0.01).

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

POLL: (iii) Given that they tested positive (i.e., conditional on them testing positive), what is the conditional probability that they have the disease?
(A) (0.00099) / (0.01098). **(B)** (0.01098) / (0.00099).
(C) (0.00099) / (0.00099 +0.01098). **(D)** (0.01098) / (0.00099 +0.01098).

→ This is $P(A_1 \mid B) = P(A_1 \cap B)/P(B)$. And we know these!

→ So, $P(A_1 \mid B) = P(A_1 \cap B)/P(B) = (0.00099)/(0.01098) = 0.0901639 \doteq 9\% \doteq 1/11$. 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 $P(B \mid A_1)$ (it was 99%), but we wanted $P(A_1 \mid B)$.

→ What is the connection between them?

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

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

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

  → In particular, if $P(A) \neq P(B)$, then $P(A \mid B) \neq P(B \mid A)$. Different!

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 (§1.5)

- 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 $P(A \mid B) = P(A)$, and $P(B \mid A) = P(B)$.

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

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

END MONDAY #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. Roll two dice. Are the two results independent?

  → Yes of course, since they are physically unrelated.

- Can two events be independent even if they are not physically separated, i.e. they deal with the same objects? Maybe!

- 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}.

POLL: Which pairs of these events are independent?
(A) A and B, only. **(B)** A and C, only. **(C)** B and C, only. **(D)** All three pairs (A and B, A and C, and B and C). **(E)** None of the pairs are independent.

→ Well, let's see ...

→ 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, i.e. each pair is 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 finite subcollection of the events.

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

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

• e.g. Flip 3 fair coins. Let A = {first coin Heads}, B = {second coin Heads}, and $C = \{HHH, THH, THT, TTH\}$. Then $\mathrm{P}(A \cap B \cap C) = \mathrm{P}(HHH) = 1/8 = \mathrm{P}(A)\,\mathrm{P}(B)\,\mathrm{P}(C)$, but $\mathrm{P}(A \cap C) = \mathrm{P}(HHH) = 1/8 \neq \mathrm{P}(A)\,\mathrm{P}(C)$.

→ So, A, B, C are not independent, although $\mathrm{P}(A \cap B \cap C) = \mathrm{P}(A)\,\mathrm{P}(B)\,\mathrm{P}(C)$.

POLL: Suppose A and B are independent. Does this necessarily imply that A and B^C are independent? **(A)** Yes, always. **(B)** Yes, but only if $\mathrm{P}(B) > 0$. **(C)** No, not necessarily. **(D)** No idea.

→ Well, let's see ...

→ We know from (∗∗) before, that $\mathrm{P}(A \cap B^C) = \mathrm{P}(A) - \mathrm{P}(A \cap B)$.

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

→ So,
$$\begin{aligned}\mathrm{P}(A \cap B^C) &= \mathrm{P}(A) - \mathrm{P}(A \cap B)\\ &= \mathrm{P}(A) - \mathrm{P}(A)\,\mathrm{P}(B) = \mathrm{P}(A)[1 - \mathrm{P}(B)] = \mathrm{P}(A)\,\mathrm{P}(B^C).\end{aligned}$$

→ So, yes, A and B^C must be independent, always!

• Can A and B be both independent and disjoint?

→ Well, yes, but if so, then $A \cap B = \emptyset$, so $\mathrm{P}(A \cap B) = \mathrm{P}(\emptyset) = 0$, but $\mathrm{P}(A \cap B) = \mathrm{P}(A)\,\mathrm{P}(B)$, so $\mathrm{P}(A)\,\mathrm{P}(B) = 0$, so either $\mathrm{P}(A) = 0$ or $\mathrm{P}(B) = 0$ (or both).

→ If $P(A) > 0$ and $P(B) > 0$, then A and B not both independent and disjoint.

Suggested Homework: 1.5.9, 1.5.14, 1.5.15, 1.5.20.

• 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

Continuity of Probabilities (§1.6)

POLL: 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 $P\{1, 2, \ldots, n\} = 1$?
(C) Yes. **(D)** No. **(E)** Not sure.

→ No! e.g. in above example with $P(i) = 2^{-i}$, we have $P\{1, 2, \ldots, n\} = \sum_{i=1}^{n} 2^{-i} = \frac{2^{-1}-2^{-n-1}}{1-2^{-1}} = 1 - 2^{-n}$, which is always < 1. (e.g. if $n = 10$, it equals $1023/1024 < 1$.)

POLL: For any probabilities P on $S = \{1, 2, 3, \ldots\}$, does there necessarily exist some finite $n \in \mathbf{N}$ with $P\{1, 2, \ldots, n\} > 0.99$? **(C)** Yes. **(D)** No. **(E)** Not sure.

→ Let's see ...

• 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 $P(A_n)$? Sort of ...

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

→ Let $A_n = \{1, 2, 3, \ldots, n\}$. Does A_n "converge" to S?

→ If so, then does $P(A_n)$ converge to $P(S) = 1$?

• 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 $\lim_{n\to\infty} \mathrm{P}(A_n) = \mathrm{P}(A)$.

END WEDNESDAY #3

→ Proof (§1.7): 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, $\mathrm{P}(A) = \sum_{i=1}^{\infty} \mathrm{P}(B_i) \equiv \lim_{n\to\infty} \sum_{i=1}^{n} \mathrm{P}(B_i)$.

→ 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:

POLL: If $\{A_n\} \searrow A$, does it necessarily follow that $\lim_{n\to\infty} \mathrm{P}(A_n) = \mathrm{P}(A)$?
(C) Yes. **(E)** No. **(F)** Not sure.

→ Well, $\{A_n\} \searrow A$ 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. above example with $\mathrm{P}(i) = 2^{-i}$: always have $\mathrm{P}\{1, 2, \ldots, n\} < 1$.

→ 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.

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

POLL: What is P(all the coins are all Heads)?
(A) 1/2. **(B)** 0. **(C)** Undefined. **(D)** Not sure.

→ How to even think about this?

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

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

→ Hence, P(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.

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

→ Suppose we only know that $\mathrm{P}\big([a,b]\big) = b - a$ whenever $0 \le a < b \le 1$.
Diagram:

POLL: Which fact follows logically from this?
(A) $\mathrm{P}(\{x\}) = 0$ for each individual $x \in [0,1]$.
(B) $\mathrm{P}((a,b)) = b - a$ whenever $0 \le a < b \le 1$.
(C) $\mathrm{P}([a,b)) = b - a$ whenever $0 \le a < b \le 1$.
(D) $\mathrm{P}((a,b]) = b - a$ whenever $0 \le a < b \le 1$.
(E) All of the above.
(F) None of the above.

- Start with an example. Know that 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 $A_n = \big[\frac{1}{2} + \frac{1}{n}, \frac{2}{3} - \frac{1}{n}\big]$ (for sufficiently large n). 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, using $A_n = [a + \frac{1}{n}, b - \frac{1}{n}]$ shows $\mathrm{P}\big((a,b)\big) = b - a$ for $0 \le a < b \le 1$.

→ Or, for e.g. $[a,b)$, use $A_n = [a, b - \frac{1}{n}]$ instead, then have $\{A_n\} \nearrow A := [a,b)$.

- What about $\mathrm{P}(\{x\})$, for $x \in \mathbf{R}$? Zero? Let $A_n = [x - \frac{1}{n}, x + \frac{1}{n}]$. Then …

→ Here $\{A_n\} \searrow A := \{x\}$. Hence, by Continuity of Probabilities,
$\mathrm{P}(\{x\}) = \lim_{n\to\infty} \mathrm{P}([x - \frac{1}{n}, x + \frac{1}{n}]) = \lim_{n\to\infty}((x + \frac{1}{n}) - (x - \frac{1}{n})) = \lim_{n\to\infty} \frac{2}{n} = 0$.

→ So, yes, it's (E) All of the above!

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 (§2.1)

- 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 = (# 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 "$X \le Y$".
- Fun Fact: 1950s Doob/Feller argument, "random variable" or "chance variable"?

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 (§2.2)

- 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. Olympic medal, with S = {Gold, Silver, Bronze}, and P(Gold)=0.40, P(Silver)=0.15, and P(Bronze)=0.45. Let $X(\text{Gold}) = 1$, $X(\text{Silver}) = 2$, $X(\text{Bronze}) = 5$.

POLL: What is $P(X \le 3)$? **(A)** 0.40. **(B)** 0.15. **(C)** 0.45. **(D)** 0.55. **(E)** 1.

→ Probabilities for X? Here $P(X = 1) = P\{\text{Gold}\} = 0.40$, and $P(X = 2) = P\{\text{Silver}\} = 0.15$, and $P(X = 5) = P\{\text{Bronze}\} = 0.45$. What about $P(X \le 3)$?

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

→ And $P(X < 20) = P\{\text{Gold, Silver, Bronze}\} = 0.40 + 0.15 + 0.45 = 1$.

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

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

→ e.g. If B is the event "≤ 3", then $B = \{x \in \mathbf{R} : x \le 3\}$, so $P(X \in B) = P(X \le 3) = P\big(X \in (-\infty, 3]\big) = P\big(X^{-1}(-\infty, 3]\big)$, which equals 0.55 in this case.

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

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

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 (§2.3)

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

→ 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 $P(X = x) = 0$ for all values of x, so $\sum_{x \in \mathbf{R}} P(X = x) = 0 < 1$.

- If it's true, there's a distinct sequence $x_1, x_2, x_3, \ldots \in \mathbf{R}$, and corresponding probabilities $p_1, p_2, p_3, \ldots \ge 0$, with $\sum_i p_i = 1$, such that $P(X = x_i) = p_i$ 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 $p_X(x) := P(X = x)$.

→ 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 $P(X = 0) = 1/2$, and $P(X = 1) = 1/2$.

→ 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 $x \ne 0, 1$.

- e.g. Flip two fair coins, and let $X = \#$ Heads.

POLL: The probability function $p_X(x)$ for this X is given by:
(A) $p_X(1) = p_X(2) = 1/2$, otherwise $p_X(x) = 0$.
(B) $p_X(0) = p_X(1) = p_X(2) = 1/3$, otherwise $p_X(x) = 0$.
(C) $p_X(0) = 1/4$ and $p_X(1) = 1/2$ and $p_X(2) = 1/4$, otherwise $p_X(x) = 0$.
(D) $p_X(0) = 1/4$ and $p_X(1) = 3/4$ and $p_X(2) = 1/4$, otherwise $p_X(x) = 0$.
(E) $p_X(0) = 1/4$ and $p_X(1) = 2/3$ and $p_X(2) = 1/4$, otherwise $p_X(x) = 0$.

• We know that $P(X = k) = \binom{n}{k}/2^n = \binom{2}{k}/4$. So, $P(X = 0) = \binom{2}{0}/2^2 = 1/4$, and $P(X = 1) = \binom{2}{1}/2^2 = 2/4 = 1/2$, and $P(X = 2) = \binom{2}{2}/2^2 = 1/4$.

→ 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 First Discrete Distributions (§2.3.1)

• 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 ...).

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

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

→ 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 ...

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

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

<u>**POLL:**</u> What is $P(X = 1)$?
(A) $\theta(1 - \theta)$. **(B)** $2\theta(1 - \theta)$. **(C)** $\theta + (1 - \theta)$. **(D)** $\theta - 2(1 - \theta)$. **(E)** Not sure.

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

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

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

• e.g. Shoot "n" free throws, each <u>independent</u> 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).

POLL: Let $X = \#$ Successes. What is the probability function for X?

(A) $p_X(k) = \theta^k$, for any $k \in \{0, 1, 2, \ldots, n\}$, otherwise 0.

(B) $p_X(k) = \theta^k (1-\theta)^{n-k}$, for any $k \in \{0, 1, 2, \ldots, n\}$, otherwise 0.

(C) $p_X(k) = \binom{n}{k}\theta^k$, for any $k \in \{0, 1, 2, \ldots, n\}$, otherwise 0.

(D) $p_X(k) = \binom{n}{k}\, \theta^k\, (1-\theta)^{n-k}$, for any $k \in \{0, 1, 2, \ldots, n\}$, otherwise 0.

(E) No idea.

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

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

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

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

→ 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 <u>any</u> integer $k \in \{0, 1, 2, \ldots, n\}$?

→ 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 $k \in \{0, 1, 2, \ldots, n\}$.

→ 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: $\sum_{k=0}^{n} \mathrm{P}(X=k) = \sum_{k=0}^{n} \binom{n}{k}\theta^k(1-\theta)^{n-k} = ??$

→ Well, recall the "Binomial Theorem": $(a+b)^n = \sum_{k=0}^{n} \binom{n}{k} a^k b^{n-k}$.

→ Set $a = \theta$ and $b = 1-\theta$: $\sum_{k=0}^{n} \binom{n}{k}\theta^k(1-\theta)^{n-k} = [\theta + (1-\theta)]^n = 1^n = 1$. Yep!

- 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 coins before.

- Special case: Binomial$(1, \theta)$ is the <u>same</u> as Bernoulli(θ).

- Suppose $X_1, X_2, \ldots, X_n \sim \text{Bernoulli}(\theta)$, for <u>independent</u> trials.

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

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

POLL: 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)?
(A) $(1/4)^2$. **(B)** $(3/4)^2$. **(C)** $(1/4)^2(3/4)^2$. **(D)** $3(1/4)^2(3/4)^2$. **(E)** $6(1/4)^2(3/4)^2$.

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

Suggested Homework: 2.3.7, 2.3.11, 2.3.14, 2.3.24.

END MONDAY #3

Geometric Distribution (§2.3.1)

POLL: e.g. Repeatedly shoot free throws, each independent with probability θ of scoring. What is P(miss exactly 3 times before first score)?
(A) $\theta/4$. **(B)** θ^3. **(C)** $(1-\theta)^3$. **(D)** $\theta^3(1-\theta)$. **(E)** $(1-\theta)^3\theta$. **(F)** No idea.

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

  → And, $\mathrm{P}(Z=3) = \mathrm{P}(\text{miss-miss-miss-score}) = (1-\theta)^3\theta$. (E)

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

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

- 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.]

POLL: 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)?
(A) $(1/4)(3/4)$. **(B)** $(1/4)^5(3/4)$. **(C)** $(1/4)(3/4)^5$. **(D)** $(1/4)^6(3/4)$.
(E) $(1/4)(3/4)^6$.

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

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

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

→ So, $P(X \geq m) = (3/4)^m > 0$ for any $m \in \mathbf{N}$. ("unbounded random variable")

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

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

- If $X \sim \text{Geometric}(\theta)$ for any $0 < \theta < 1$, and any $m \in \mathbf{N}$, then we still have $P(X \geq m) = (1-\theta)^m > 0$, unbounded, but still $P(X = \infty) = 0$.

Suggested Homework: 2.3.6, 2.3.10, 2.3.15, 2.3.16(a,b), 2.3.23, 2.3.27.

- Suppose $X \sim \text{Geometric}(\theta)$, and $a, b \in \mathbf{N}$. Then what is $P(X \geq a+b \,|\, X \geq a)$?

→ $P(X \geq a+b \,|\, X \geq a) = \frac{P(X \geq a+b \text{ and } X \geq a)}{P(X \geq a)} = \frac{P(X \geq a+b)}{P(X \geq a)} = \frac{(1-\theta)^{a+b}}{(1-\theta)^a} = (1-\theta)^b$.

→ So what? Well, this is equal to $P(X \geq b)$.

→ Suppose your waiting time (for a bus, or an elevator, or ...) is Geometric(θ).

→ Suppose you've already waited for a minutes.

→ Then the probabilities for how long you still have to wait, are the same as they were when you started waiting!

→ This is the "memoryless" or "forgetfulness" property of Geometric(θ).

Poisson Distribution (§2.3.1)

- 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

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

→ Now, what happens as $n \to \infty$, for a fixed value of k?

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

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

→ Also, from calculus, $e^x = 1 + x + \frac{x^2}{2!} + \dots$, 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 $P(\#\text{fires} = k) \;\to\; \frac{1}{k!}\lambda^k e^{-\lambda} \;=\; e^{-\lambda}\frac{\lambda^k}{k!}$.

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

- Check: $\sum_{k=0}^{\infty} e^{-\lambda}\frac{\lambda^k}{k!} = e^{-\lambda}\,[1 + \lambda + \frac{\lambda^2}{2!} + \frac{\lambda^3}{3!} + \dots] = e^{-\lambda}\,[e^{\lambda}] = 1$. 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 $Y \sim \text{Poisson}(3)$. What is $P(Y = 4)$?

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

POLL: e.g. Suppose $Y \sim \text{Binomial}(20000, 0.0001)$. Then the actual value of $P(Y = 4)$, and the Poisson approximation value of $P(Y = 4)$, are:
(A) $20000\,(0.0001)^4$, and $e^{-20000}\frac{(20000)^4}{4!}$.
(B) $\binom{20000}{4}(0.0001)^4$, and $e^{-20000}\frac{(2)^4}{4!}$.
(C) $\binom{20000}{4}(0.0001)^4(0.9999)^{19996}$, and $e^{-2}\frac{(2)^4}{4!}$.
(D) $\binom{20000}{4}(0.0001)^4(0.9999)^{19996}$, and $e^{-20000}\frac{(2)^4}{4!}$.

- Here $Y \sim \text{Binomial}(n, \theta)$ where $n = 20000$ and $\theta = 0.0001$.

  → So, $P(Y = 4) = \binom{n}{4}\theta^4(1-\theta)^{n-4} = \binom{20000}{4}(0.0001)^4(0.9999)^{19996}$

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

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

POLL: To how many decimal points do these two values agree? (Don't compute it, just guess.) **(A)** 3. **(B)** 4. **(C)** 5. **(D)** 6. **(E)** 7. **(F)** 8.

→ Here $P(Y = 4) = \binom{20000}{4}(0.0001)^4(0.9999)^{19996} \doteq 0.09022352216$.

→ And, the approximation is $P(Y = 4) \approx e^{-2}\frac{(2)^4}{4!} \doteq 0.09022352178$.

→ (Agree to 8 decimal places! Or even 9, with rounding!)

- Or, if $Y \sim \text{Binomial}(200, 0.01)$, then still $\lambda = 200 \cdot 0.01 = 2$, so Poisson approximation is the same, but how close is it now?

POLL: Same question as above, but now for Binomial(200, 0.01).

→ Now $P(Y = 4) = \binom{200}{4}(0.01)^4(0.99)^{200-4} \doteq 0.0902197$.

→ Still pretty close: 4 decimals! (Or 5 with rounding!) But not as close.

→ Binomial(20,0.1): P(Y=4) $\doteq$ 0.0897788. (3 decimals with rounding)

→ Binomial(10,0.2): P(Y=4) $\doteq$ 0.0881; Binomial(5,0.4): $\doteq$ 0.0768; worse.

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

- We'll omit some other common discrete distributions (save for next year!).

  → e.g. Negative-Binomial(r, θ) and Hypergeometric(N, M, n).

Law of Total Probability (again) (§2.3)

- 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
$P(B) = \sum_i P(X = x_i)\, P(B \,|\, X = x_i) = \sum_i p_i\, P(B \,|\, X = x_i)$.

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

END WEDNESDAY #4

POLL: 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. Then $\mathrm{P}(X = 3)$ equals: **(A)** $\sum_{y=3}^{6}(1/6)\,[\binom{y}{3}/2^y]$. **(B)** $\binom{6}{3}/2^y$. **(C)** $\frac{6!}{3!}/2^3$. **(D)** $\sum_{y=3}^{6}(1/6)\,[y(y-1)(y-2)/2^y]$. **(E)** $\sum_{y=1}^{6}(1/6)\,[y(y-1)(y-2)/6(2^y)]$. **(F)** No idea.

- 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 ...

$\mathrm{P}(X=3) = \sum_{y\in\mathbf{R}} \mathrm{P}(Y=y)\ \mathrm{P}(X=3\,|\,Y=y) = \sum_{y=1}^{6} \mathrm{P}(Y=y)\ \mathrm{P}(X=3\,|\,Y=y)$
$= \sum_{y=3}^{6}(1/6)\ [\binom{y}{3}\ /\ 2^y]$. (A)

→ This equals $\frac{1}{6}\left(\frac{1}{8}+\frac{4}{16}+\frac{10}{32}+\frac{20}{64}\right) = \frac{1}{6}(1) = \frac{1}{6}$. (Why? Coincidence!)

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

Understanding Distributions Using the Computer Software "R"

- Recall – basic info and links at: 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 get more info with e.g. "?rbinom", etc.
- 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.]

Continuous Random Variables (§2.4)

- 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):
  - → $X \sim \mathrm{Uniform}[0,1]$ if $\mathrm{P}(a \le X \le b) = b - a$ whenever $0 \le a \le b \le 1$.
  - → Then e.g. $\mathrm{P}(X \in [0,1]) = \mathrm{P}(0 \le X \le 1) = 1 - 0 = 1$,

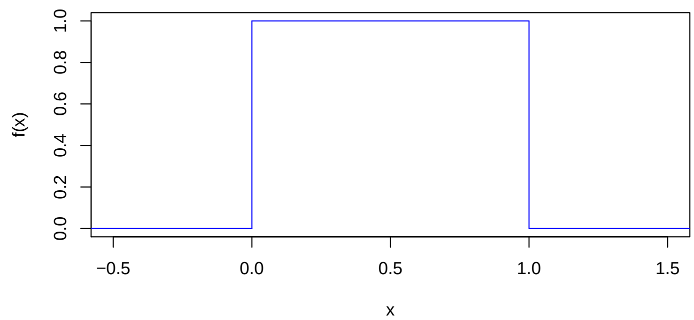
$\mathrm{P}(1/3 \le X \le 3/4) = (3/4) - (1/3) = 5/12$,
$\mathrm{P}(X \ge 2/3) = \mathrm{P}(2/3 \le X \le 1) = 1 - (2/3) = 1/3$, etc.

→ Also, $\text{P}(X > 1) = 0$, and $\text{P}(X < 0) = 0$, so e.g. $\text{P}(1/3 \le X \le 5) = \text{P}(1/3 \le X \le 1) = 1 - (1/3) = 2/3$, etc.

→ And, we previously showed (using Continuity Of Probabilities) that we can always replace "$\le$" with "$<$", or "$>$" by "$\ge$", etc. (Also true since $\text{P}(X = x) = 0$.)

- 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> $a \le b$,

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

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

  → This defines a new distribution! Very general! ("absolutely continuous")

- Follows that $\text{P}(X = a) = \text{P}(a \le X \le a) = \int_a^a f(x)\, dx = 0$, i.e. X is continuous.
- If $f(x)$ is the density function for a random variable X, write it as $f_X(x)$.

<u>Some First Continuous Distributions (§2.4.1)</u>

- e.g. the Uniform[5,12] distribution has density: $f_X(x) = \begin{cases} 0, & x < 5 \\ 1/7, & 5 \le x \le 12 \\ 0, & x > 12 \end{cases}$

Diagram:

→ 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.

<u>POLL:</u> Then for any $5 \le a \le b \le 12$, the probability $P(a \le X \le b)$ is equal to:
(A) $b - a$. **(B)** $\frac{1}{7}(b-a)$. **(C)** $\frac{2}{7}(12-a)$. **(D)** $\frac{2}{7}(b-5)$. **(E)** $\frac{1}{7}(b-a-5)$.
(F) No idea.

- For <u>any</u> $L < R$, 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. if $L = 5$ and $R = 12$, then $P(a \le X \le b) = \frac{b-a}{R-L} = \frac{1}{7}(b-a)$. (Of course.)

- If $X \sim \text{Uniform}[L, R]$, then $\text{P}(L \le X \le R) = 1$. (Bounded distribution.)
- e.g. Let $f(x) = e^{-x}$ for $x \ge 0$, otherwise $f(x) = 0$. Diagram:

→ 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 $0 \le a \le b$, $\text{P}(a \le X \le b) = \int_a^b e^{-x}\,dx = e^{-a} - e^{-b}$.

→ Also $\text{P}(X \ge a) = e^{-a}$. 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 $0 \le a \le b$, $\text{P}(a \le X \le b) = e^{-\lambda a} - e^{-\lambda b}$.

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

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

POLL: What property does Exponential(λ) have, just like a previous distribution?
(A) It gives probabilities for coin flips, just like Binomial(n, θ).
(B) It's a bounded distribution, just like Uniform[L, R].
(C) It has the memoryless property, just like Geometric(θ).
(D) It is a limit of Binomials, just like Poisson(λ).
(E) All of the above.
(F) Exactly two of the above.

- Suppose $X \sim \text{Exponential}(\lambda)$, and $a, b > 0$. Then what is $\text{P}(X \ge a + b \,|\, X \ge a)$?

→ $\text{P}(X \ge a + b \,|\, X \ge a) = \frac{\text{P}(X \ge a+b \text{ and } X \ge a)}{\text{P}(X \ge a)} = \frac{\text{P}(X \ge a+b)}{\text{P}(X \ge a)} = \frac{e^{-\lambda(a+b)}}{e^{-\lambda a}} = e^{-\lambda b}$.

→ So what? Well, this is equal to $\text{P}(X \ge b)$.

→ If your waiting time is Exponential(λ), and you've already waited for a minutes, then the probabilities for how long you still have to wait are the same as they were when you started waiting. Just like for Geometric(θ).

→ This is the "memoryless" or "forgetfulness" property of Exponential(λ).

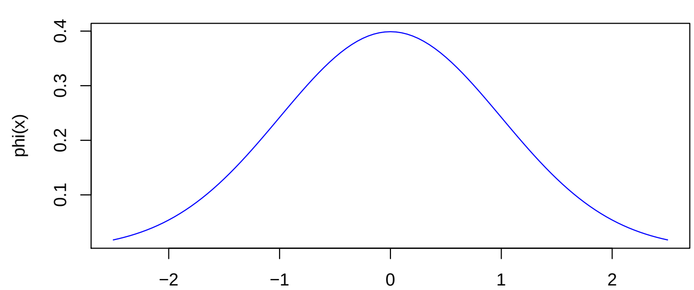
Suggested Homework: 2.4.1, 2.4.2, 2.4.3, 2.4.4, 2.4.5, 2.4.6, 2.4.7, 2.4.8, 2.4.9, 2.4.10, 2.4.11, 2.4.12, 2.4.14.

The Normal Distribution (§2.4.1)

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

→ "Standard normal density"

→ "bell curve", "Gaussian"



→ Clearly $\phi(x) \geq 0$.

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

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

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

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

→ Then $f(x) \geq 0$. 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.)

→ Arises naturally when there are lots of small influences.

→ See e.g. https://www.mathsisfun.com/data/quincunx.html

Suggested Homework: 2.4.13, 2.4.26.

- We'll <u>omit</u> some other common continuous distributions, e.g. Gamma(α, λ).

Cumulative Distribution Functions (cdf) (§2.5)

- For <u>any</u> random variable X, the cumulative distribution function (cdf) is the function F_X defined by $F_X(x) = \mathrm{P}(X \leq x)$ for all $x \in \mathbf{R}$.

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

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

<u>POLL:</u> If $a < b$, the expression $F_X(b) - F_X(a)$ is equal to:
(A) $\mathrm{P}[X \leq \min(a,b)]$. **(B)** $\mathrm{P}[X \geq \max(a,b)]$. **(C)** $\mathrm{P}[a < X < b]$.
(D) $\mathrm{P}[a < X \leq b]$. **(E)** $\mathrm{P}[a \leq X < b]$. **(F)** $\mathrm{P}[a \leq X \leq b]$.

- Well, for any $a < b$, let $A = \{X \leq a\}$ and $B = \{X \leq b\}$.

→ Then $A \cap B = A = \{X \leq a\}$.

→ Then, $\{a < X \le b\} = \{X \le b\} \cap \{X > a\} = \{X \le b\} \cap \{X \le a\}^C = B \cap A^C$.

→ Hence, $\mathrm{P}(a < X \le b) = \mathrm{P}(B \cap A^C) = \mathrm{P}(B) - \mathrm{P}(A \cap B)$
$= \mathrm{P}(B) - \mathrm{P}(A) = \mathrm{P}(X \le b) - \mathrm{P}(X \le a) = F_X(b) - F_X(a)$. So: (D).

POLL: If X has cdf F_X, then $\mathrm{P}(a \le X \le b)$ must always be equal to:
(A) $F_X(b) - F_X(a)$.
(B) $F_X(b) - \lim_{n\to\infty} F_X(a - \frac{1}{n})$.
(C) $F_X(b) - \lim_{n\to\infty} F_X(a + \frac{1}{n})$.
(D) $\lim_{n\to\infty} F_X(b - \frac{1}{n}) - F_X(a)$.
(E) $\lim_{n\to\infty} F_X(b + \frac{1}{n}) - F_X(a)$.
(F) $\lim_{n\to\infty} F_X(b - \frac{1}{n}) - \lim_{n\to\infty} F_X(a - \frac{1}{n})$.

- Indeed, by Continuity Of Probabilities, $\mathrm{P}(a \le X \le b) = \mathrm{P}(X \le b) - \mathrm{P}(X < a) = \mathrm{P}(X \le b) - \lim_{n\to\infty} \mathrm{P}(X \le a - \frac{1}{n}) = F_X(b) - \lim_{n\to\infty} F_X(a - \frac{1}{n})$. (B)

  → If F_X is a continuous function, then $\mathrm{P}(a \le X \le b) = F_X(b) - F_X(a)$.

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

  → Might equal 0, but might be positive!

  → If F_X is continuous, then $\mathrm{P}(X = a) = \mathrm{P}(a \le X \le a)$
  $= F_X(a) - \lim_{n\to\infty} F_X(a - \frac{1}{n}) = F_X(a) - F_X(a) = 0$.

- And, e.g. $\mathrm{P}(3 < X \le 5$ or $6 < X \le 9) = [F_X(5) - F_X(3)] + [F_X(9) - F_X(6)]$, etc.

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

POLL: The cumulative distribution function (cdf) F_X of any real-valued random variable X must always satisfy the following property:
(A) $0 \le F_X(x) \le 1$ for all $x \in \mathbf{R}$.
(B) If $x \le y$, then $F_X(x) \le F_X(y)$, i.e. F_X is a non-decreasing function.
(C) $\lim_{x\to-\infty} F_X(x) = 0$.
(D) $\lim_{x\to\infty} F_X(x) = 1$.
(E) All of the above.
(F) Exactly two of the above.

END MONDAY #4

- Well, let's see …

  → $F_X(x) = \mathrm{P}(X \le x)$ is a probability, so $0 \le F_X(x) \le 1$ for all $x \in \mathbf{R}$.

  → If $x \le y$, and we set $A = \{X \le x\}$ and $B = \{X \le y\}$, then $A \subseteq B$. Hence, $\mathrm{P}(A) \le \mathrm{P}(B)$, i.e. $F_X(x) \le F_X(y)$. (non-decreasing)

  → If $A_n = \{X \le -n\}$, then $\{A_n\} \searrow \{X = -\infty\} = \emptyset$, so by Continuity of Probabilities, $\lim_{n\to\infty} \mathrm{P}(A_n) = \mathrm{P}(\bigcap_n A_n) = \mathrm{P}(\emptyset) = 0$.

  → Similarly, if $A_n = \{X \le +n\}$, then $\{A_n\} \nearrow \{X < \infty\} = S$, so by Continuity of Probabilities, $\lim_{n\to\infty} \mathrm{P}(A_n) = \mathrm{P}(\bigcup_n A_n) = \mathrm{P}(S) = 1$.

  → So: (E) All of the above!

POLL: Are cumulative distribution functions (cdfs) <u>continuous</u> functions?
(A) Yes, they must always be continuous functions.
(B) They must be <u>left-continuous</u>, but might not be <u>right-continuous</u>.
(C) They must be <u>right-continuous</u>, but might not be <u>left-continuous</u>.
(D) They might be neither left- nor right-continuous.
(E) No idea.

- Well, 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 $\mathrm{P}(X \leq x)$.]

  → i.e. $F_X(x - \frac{1}{n}) \to \mathrm{P}(X < x) = \mathrm{P}(X \leq x) - \mathrm{P}(X = x) = F_X(x) - \mathrm{P}(X = x)$.

- If $\mathrm{P}(X = x) > 0$, then $F_X(x)$ is discontinuous at x. Just right-continuous. (C)

- If $\mathrm{P}(X = x) = 0$, then $F_X(x - \frac{1}{n}) \to F_X(x)$, so F_X is also "left-continuous".

  → Then, since it's right-continuous <u>and</u> left-continuous, then it is continuous!

  → So, if X is a <u>continuous</u> random variable, i.e. $\mathrm{P}(X = x) = 0$ for all x, then F_X is a continuous function for all x. (This is actually "if and only if".)

- In general, the jump-size of F_X at x is equal to $\mathrm{P}(X = x)$.

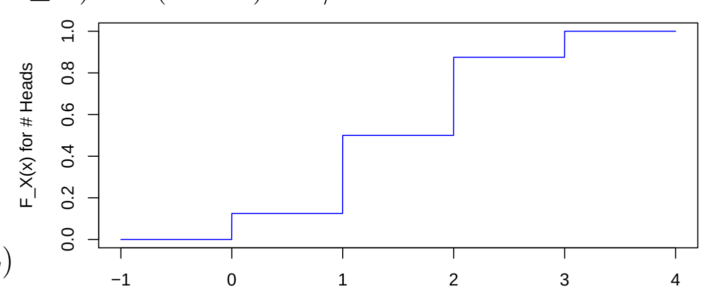
- e.g. Flip 3 coins, $X = \#$ Heads.

POLL: In this example, what is the cdf value $F_X(2.5)$?
(A) 1/8. **(B)** 3/8. **(C)** 1/2. **(D)** 5/8. **(E)** 7/8. **(F)** 1.

→ 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 $x < 0$, $F_X(x) = \mathrm{P}(X \leq x) = 0$.

→ And, for $0 \leq x < 1$, $F_X(x) = \mathrm{P}(X \leq x) = \mathrm{P}(X = 0) = 1/8$.



→ And, for $1 \leq x < 2$, $F_X(x) = \mathrm{P}(X \leq x) = \mathrm{P}(X = 0) + \mathrm{P}(X = 1) = 1/8 + 3/8 = 4/8 = 1/2$.

→ And, for $2 \leq x < 3$, $F_X(x) = \mathrm{P}(X \leq x) = \mathrm{P}(X = 0) + \mathrm{P}(X = 1) + \mathrm{P}(X = 2) = 1/8 + 3/8 + 3/8 = 7/8$. (E)

→ And, for $x \geq 3$, $F_X(x) = \mathrm{P}(X \leq x) = \mathrm{P}(X = 0) + \mathrm{P}(X = 1) + \mathrm{P}(X = 2) + \mathrm{P}(X = 3) = 1/8 + 3/8 + 3/8 + 1/8 = 1$.

→ [Graph.] All properties satisfied!

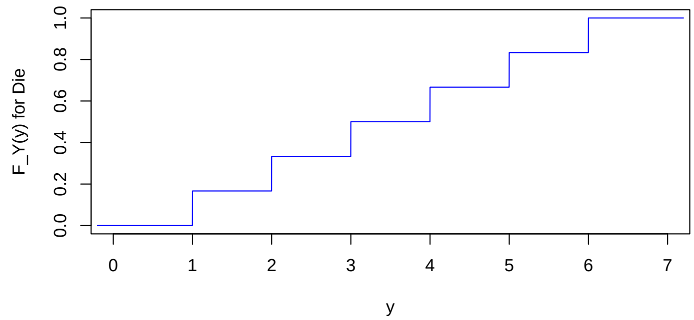
POLL: In this example, what is the value of $F_X(1) - \lim_{n\to\infty} F_X(1 - \frac{1}{n})$?
(A) 1/8. **(B)** 3/8. **(C)** 1/2. **(D)** 5/8. **(E)** 7/8. **(F)** 1.

→ $F_X(1) - \lim_{n\to\infty} F_X(1 - \frac{1}{n}) = \mathrm{P}(X \leq 1) - \mathrm{P}(X < 1) = \mathrm{P}(X = 1) = 3/8$. (B)

- All <u>discrete</u> distributions have somewhat similar cdfs. (piecewise-constant)

- e.g. Y = roll of one fair six-sided die. Check that all properties of F_Y are satisfied!

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



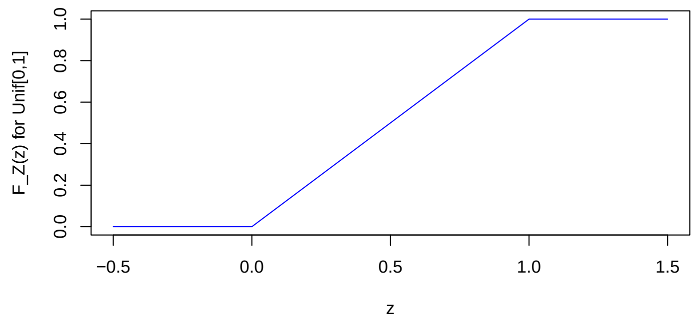
- Continuous? e.g. $X \sim$ Uniform$[0,1]$.

  $\rightarrow$ Then $P(X \le x) = 0$ for $x < 0$.

  $\rightarrow$ And, $P(X \le x) = 1$ for $x > 1$.

  $\rightarrow$ For $0 \le x \le 1$, $P(X \le x) = P(0 \le X \le x) = x - 0 = x$.

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



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

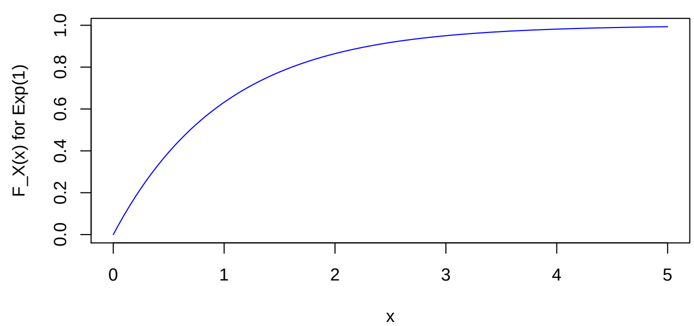
  $\rightarrow$ So e.g. if $L = 2$ and $R = 5$, then $F_Z(z) = \frac{z-2}{3}$ for $2 \le z \le 5$.

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

  $\rightarrow$ Then $P(X < 0) = 0$.

  $\rightarrow$ So, for $x < 0$, $F_X(x) = 0$.

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



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

  $\rightarrow$ Then $P(Y < 0) = 0$. So, for $y < 0$, $F_Y(y) = 0$.

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

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

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

  $\rightarrow$ Can use previous knowledge that $P(X \ge a) = e^{-\lambda a}$. Or ...

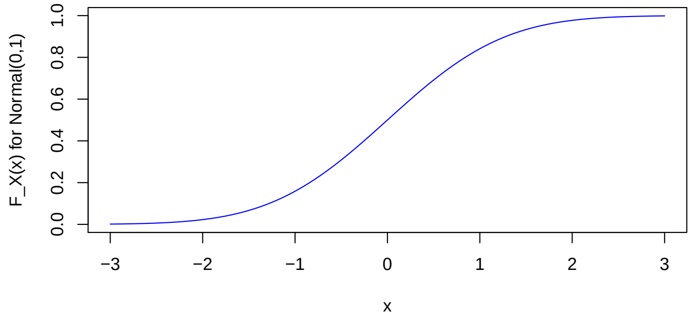
  $\rightarrow$ Here F_X is continuous, so $P(X \ge 2.6) = 1 - P(X < 2.6) = 1 - P(X \le 2.6) = 1 - F_X(2.6) = 1 - [1 - e^{-3(2.6)}] = e^{-3(2.6)} = e^{-7.8} \doteq 0.00041$.

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

→ [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.

- Furthermore, the bell curve is symmetric, i.e. $\phi(-u) = \phi(u)$ for all u.

→ This implies that $\text{P}(Z \le x) = \text{P}(Z \ge -x)$, i.e. $\text{P}(Z \le x) = 1 - \text{P}(Z \le -x)$.

→ So, $\Phi(x) = 1 - \Phi(-x)$ for all $x \in \mathbf{R}$, i.e. $\Phi(x) + \Phi(-x) = 1$.

→ It then also follows that $\Phi(0) = 1/2$.

END WEDNESDAY #5

POLL: e.g. Suppose $Z \sim \text{Normal}(0,1)$. What is $\text{P}(Z \le 1.43)$?
(A) $\Phi(1.43)$. **(B)** $1 - \Phi(-1.43)$. **(C)** $\int_{-\infty}^{1.43} \phi(x)\,dx$. **(D)** $(1/2) + \int_0^{1.43} \phi(x)\,dx$.
(E) $1 - \int_{1.43}^{\infty} \phi(x)\,dx$. **(F)** All of the above.

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

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

→ And, since $\Phi(x) = \int_{-\infty}^{x} \phi(x)\,dx$ and $\int_{-\infty}^{\infty} \phi(x)\,dx = 1$ and $\int_{-\infty}^{0} \phi(x)\,dx = 1/2$, the other expressions all equal this, too! So, (F)!

POLL: e.g. Suppose $W \sim \text{Normal}(5, 4^2)$. What is $\text{P}(6 \le W \le 8)$? **(A)** $\Phi(1/4)$.
(B) $\Phi(3/4)$. **(C)** $\Phi(3/4) + \Phi(1/4)$. **(D)** $\Phi(3/4) - \Phi(1/4)$. **(E)** $\Phi(7/8) - \Phi(1/8)$.

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

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

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

→ Then, this also 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 $\text{P}(6 \le W \le 8) \doteq 0.1747$.

Suggested Homework: 2.5.4, 2.5.5.

- 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 $F_X(x) := \text{P}(X \le x) = \int_{-\infty}^{x} f_X(u)\,du$.

→ 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 $X \sim$ 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 $Y \sim$ Exponential(λ), 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 $Z \sim$ 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. $\text{P}(3 < X \leq 5.5) = F_X(5.5) - F_X(3) = (5.5 - 5)^4 - 0 = 0.0625$.

  → 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 (§2.5.4): 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.

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

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

  → 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).

POLL: Then what is, say, $F_X(4.4)$? **(A)** $F_Y(4.4) + F_Z(4.4)$. **(B)** $[F_Y(4.4) + F_Z(4.4)]/2$. **(C)** $(1/3)F_Y(4.4) + (1/3)F_Z(4.4)$. **(D)** $(1/3)F_Y(4.4) + (2/3)F_Z(4.4)$.

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

→ 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. $\text{P}(X = 2) = \text{P}(W = 1,\ Y = 2) = \text{P}(W = 1)\ \text{P}(Y = 2) = (1/3)(1/6) = 1/18 > 0$. Not zero, like for the continuous case.

- Ah, so then is X discrete?

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

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

  $\rightarrow$ (In this course we'll usually stick with either discrete or absolutely continuous. But there are other kinds of random variables too. Even "singular", 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) (§2.6)

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

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

  $\rightarrow$ Then Y is another random variable. ("function of a random variable")

  $\rightarrow$ 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 $\sum_i p_i = 1$.

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

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

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

  $\rightarrow$ Discrete Change-of-Variable Theorem.

POLL: e.g. X = roll of fair die, and $Y = (X-3)^2$. What is $\mathrm{P}(Y = 4)$?
(A) 0. **(B)** 1/6. **(C)** 1/3. **(D)** 1/2. **(E)** 2/3. **(F)** 5/6.

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

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

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

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

$\rightarrow$ 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 $Y = h(X)$.

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

POLL: Will Y necessarily be absolutely continuous at all?
(A) Yes, Y must be absolutely continuous (i.e., have a density).
(B) Well, Y must be continuous (i.e. $\mathrm{P}(Y{=}y) = 0$ for all y), but not necessarily absolutely continuous.
(C) Well, Y might not be continuous, but cannot be a discrete random variable.
(D) Actually, Y could even be a discrete random variable.
(E) No idea.

- Well, let's consider an example ...

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

→ Then if $Y = h(X)$, then $\mathrm{P}(Y = 2) = \mathrm{P}(X \le 1/3) = 1/3$, 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 satisfies certain conditions?

→ Then must Y be absolutely continuous, i.e. have a density $f_Y(y)$?

→ And if yes, then what must $f_Y(y)$ equal?

[Reminder: MIDTERM #1, Wednesday Oct 9, at regular lecture time, in Exam Centre (EX) room 320 or 100. Bring TCard, basic calculator.]

[Reminder: Monday Oct 14 is THANKSGIVING – no classes.]

END MONDAY #5

(Midterm #1.)

END WEDNESDAY #6

(Thanksgiving – no class.)

END MONDAY #6

POLL: Suppose X is absolutely continuous, with density function $f_X(x)$, and $Y = h(X)$. Then Y must also be absolutely continuous, i.e. also have a density function, provided that h is: **(A)** Continuous. **(B)** Non-decreasing. **(C)** Strictly increasing. **(D)** Constant. **(E)** None of the above. **(F)** No idea.

- 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 $f_Y(y) = f_X\big(h^{-1}(y)\big) \,/\, \big|h'\big(h^{-1}(y)\big)\big|$.

→ That is, $f_Y(y) = f_X(x)/|h'(x)|$, where $y = h(x)$ so $x = h^{-1}(y)$.

- Proof: Assume h is strictly increasing.

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

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

- Method #1:

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

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

→ 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:

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

→ 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 $X \sim \mathrm{Uniform}[0,1]$, and $Y = 5X + 4$.

POLL: Will Y be absolutely continuous? **(C)** Yes. **(D)** No. **(E)** No idea.

POLL: What distribution do you think Y will have?
(A) Uniform[0,1]. **(B)** Uniform[0,5]. **(C)** Uniform[0,9]. **(D)** Uniform[4,9].
(E) Some other Uniform distribution. **(F)** Some non-Uniform distribution.

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

- 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 $X \sim \mathrm{Uniform}[0,1]$, and $Y = X^2$.

POLL: Will Y be absolutely continuous? **(C)** Yes. **(D)** No. **(E)** No idea.

POLL: What distribution do you think Y will have?
(A) Uniform[0,1]. **(B)** Uniform[0,2]. **(C)** Uniform[0,4]. **(D)** Uniform[1,4].
(E) Some other Uniform distribution. **(F)** Some non-Uniform distribution.

→ 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: (F).]

→ Alternatively: For $0 \le y \le 1$, $F_y(y) = P(Y \le y) = P(X^2 \le y) = P(X \le \sqrt{y}) = \sqrt{y}$, 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.

- e.g. Suppose $X \sim$ Exponential(5), and $Y = X^2$.

POLL: Will Y be absolutely continuous? **(C)** Yes. **(D)** No. **(E)** No idea.

POLL: What distribution do you think Y will have?
(A) Uniform[0,1]. **(B)** Uniform[0,5]. **(C)** Exponential(10). **(D)** Exponential(25). **(E)** Some other Uniform or Exponential distribution. **(F)** Some non-Uniform nor Exponential distribution.

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

→ 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 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! (F)

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

POLL: Will Y be absolutely continuous? **(C)** Yes. **(D)** No. **(E)** No idea.

POLL: What distribution do you think Y will have?
(A) Normal(0,1). **(B)** Normal(0,9). **(C)** Normal$(3, 6^2)$. **(D)** Normal$(6, 3^2)$.
(E) Some other Normal distribution. **(F)** Some non-Normal distribution.

END WEDNESDAY #7

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

→ 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, $Y \sim$ Normal$(6, 3^2)$, as might be expected. (D)

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

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

Joint Distributions (§2.7)

- 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 $Y_3 = I_{\text{second coin Heads}}$.

POLL: What are the distributions of Y_1 and Y_2 and Y_3?
(A) $Y_1 \sim \text{Bernoulli}(1/2)$; $Y_2 \sim \text{Bernoulli}(1/2)$; $Y_3 \sim \text{Bernoulli}(1/2)$.
(B) $Y_1 \sim \text{Bernoulli}(1/2)$; $Y_2 \sim \text{Bernoulli}(-1/2)$; $Y_3 \sim \text{Bernoulli}(3/2)$.
(C) $Y_1 \sim \text{Bernoulli}(1/2)$; $Y_2 \sim \text{Bernoulli}(0)$; $Y_3 \sim \text{Bernoulli}(1)$.
(D) $Y_1 \sim \text{Bernoulli}(1/2)$; $Y_2 \sim \text{Bernoulli}(1)$; $Y_3 \sim \text{Bernoulli}(1/2)$.
(E) Some other Bernoulli distributions.
(F) Some other non-Bernoulli distributions.

- Here each of the Y_i is equally likely to equal 0 or 1.

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

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

POLL: What are $\text{P}(X{=}1,\ Y_1{=}1)$; and $\text{P}(X{=}1,\ Y_2{=}1)$; and $\text{P}(X{=}1,\ Y_3{=}1)$?
(A) 1/4; 1/4; 1/4. **(B)** 1/2; 1/2; 1/2. **(C)** 1/2; 1/2; 0. **(D)** 1/2; 0; 1/2.
(E) 1/2; 0; 1/4. **(F)** 1/4; 0; 1/2.

→ Here $\text{P}(X{=}1,\ Y_1{=}1) = 1/2$ [since $Y_1 = X$, same], and $\text{P}(X{=}1,\ Y_2{=}1) = 0$ [since $Y_2 = 1 - X$, opposite], and $\text{P}(X{=}1,\ Y_3{=}1) = 1/4$ [since Y_3, X indep.]. (E)

→ All different! Despite same individual distributions!

- 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; first).

Joint Cumulative Distribution Functions (§2.7.1)

- 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 $F_{X,Y}(x,y) = \text{P}(X \le x,\ Y \le y) \equiv \text{P}(X \le x \text{ and } Y \le y)$.

→ Like before, cdf's provide all information about all joint probabilities, e.g. $\text{P}(a < X \le b,\ c < Y \le d) = F_{X,Y}(b,d) - F_{X,Y}(a,d) - F_{X,Y}(b,c) + F_{X,Y}(a,c)$. [Why?]

→ However, joint cdf's can be quite tricky, and difficult to work with.

→ So, we will omit them here. (But feel free to ask about them!)

Joint Probability Functions (§2.7.3)

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

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

POLL: What is $p_{X,Y_i}(x,y)$ with $i=1$?
(A) $p_{X,Y_i}(1,1) = 1/2$ and $p_{X,Y_i}(0,0) = 1/2$ (otherwise $p_{X,Y_i}(x,y)=0$).
(B) $p_{X,Y_i}(1,0) = 1/2$ and $p_{X,Y_i}(0,1) = 1/2$ (otherwise $p_{X,Y_i}(x,y)=0$).
(C) $p_{X,Y_i}(1,0) = 1/2$ and $p_{X,Y_i}(0,1) = p_{X,Y_i}(1,1) = 1/4$ (otherwise $p_{X,Y_i}(x,y)=0$).
(D) $p_{X,Y_i}(1,0) = p_{X,Y_i}(0,1) = p_{X,Y_i}(1,1) = 1/3$ (otherwise $p_{X,Y_i}(x,y)=0$).
(E) $p_{X,Y_i}(1,0) = p_{X,Y_i}(0,1) = p_{X,Y_i}(1,1) = p_{X,Y_i}(0,0) = 1/4$ (o.w. $p_{X,Y_i}(x,y)=0$).
(F) Other.

POLL: Same question (and answers), except with $i=2$.

POLL: Same question (and answers), except with $i=3$.

- 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	

POLL: If we switch from Y_3 to Y_1, which entries in the above table will change?
(A) The blue ones, only. **(B)** The green ones, only. **(C)** The red ones, only.
(D) The blue and green but not red ones. **(E)** The blue and red but not green ones.
(F) The green and red but not blue ones.

- Well, the marginal distribution of X (green) will not change.

  → And, the marginal distribution of the Y_i (red) are all Bernoulli(1/2) so they will not change.

  → But the joint (blue) probabilities will change, as discussed above. (A)

- Can we find other joint probabilities from $p_{X,Y}(x,y)$?

  → e.g. can we find $\mathrm{P}(a \le X \le b,\ c \le Y \le d)$, for any $a<b$ and $c<d$?

  → Yes! $\mathrm{P}(a \le X \le b,\ c \le Y \le d) = \sum_{a \le x \le b} \sum_{c \le y \le d} p_{X,Y}(x,y)$, etc.

Suggested Homework: 2.7.3, 2.7.6.

Joint Density Functions (§2.7.4)

• 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 $P(a \leq X \leq b,\ c \leq Y \leq d) = \int_c^d \int_a^b f_{X,Y}(x,y)\,dx\,dy$ for all $a \leq b$ and $c \leq d$.

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

→ 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 examples below).

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

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

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

→ But $P(a \leq Y \leq b) = \int_a^b f_Y(y)\,dy$, for all $a \leq b$.

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

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

• SIMPLE EXAMPLE: $f_{X,Y}(x,y) = \frac{4}{3}x + y^2$ for $0 \leq x \leq 1$ and $0 \leq y \leq 1$, otherwise 0.

→ Check: ≥ 0 (yes). And, $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dx\,dy = \int_0^1 \Big(\int_0^1 (\frac{4}{3}x + y^2)\,dx \Big) dy = \int_0^1 (\frac{2}{3} + y^2)\,dy = \frac{2}{3} + \frac{1}{3} = 1$. Yes.

→ Or, in the other order: $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dy\,dx = \int_0^1 \Big(\int_0^1 (\frac{4}{3}x + y^2)\,dy \Big) dx = \int_0^1 (\frac{4}{3}x + \frac{1}{3})\,dx = \frac{4}{3}\frac{1}{2} + \frac{1}{3} = 1$. Yes.

→ $f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dy = \int_0^1 (\frac{4}{3}x + y^2)\,dy = \frac{4}{3}x + \frac{1}{3}$ for $0 \leq x \leq 1$, o.w. 0.

→ $f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dx = \int_0^1 (\frac{4}{3}x + y^2)\,dx = \frac{2}{3} + y^2$ for $0 \leq y \leq 1$, o.w. 0.

→ Check: $\int_0^1 (\frac{4}{3}x + \frac{1}{3})\,dx = \frac{2}{3} + \frac{1}{3} = 1$, and $\int_0^1 (\frac{2}{3} + y^2)\,dy = \frac{2}{3} + \frac{1}{3} = 1$.

→ And, $P(X < \frac{1}{2},\ Y < \frac{2}{3}) = \int_0^{\frac{2}{3}} \Big(\int_0^{\frac{1}{2}} (\frac{4}{3}x + y^2)\,dx \Big) dy = \int_0^{\frac{2}{3}} (\frac{4}{3}\frac{1}{8} + y^2\,\frac{1}{2})\,dy = (\frac{4}{3}\frac{1}{8}\frac{2}{3} + [(\frac{2}{3})^3/3]\,\frac{1}{2}) = 13/81$.

→ Or, $P(X < \frac{1}{2},\ Y < \frac{2}{3}) = \int_0^{\frac{1}{2}} \Big(\int_0^{\frac{2}{3}} (\frac{4}{3}x + y^2)\,dy \Big) dx = \int_0^{\frac{1}{2}} \Big(\frac{4}{3}x\,\frac{2}{3} + [(\frac{2}{3})^3/3] \Big)\,dx = (\frac{4}{3}\frac{1}{8}\frac{2}{3} + [(\frac{2}{3})^3/3]\,\frac{1}{2}) = 13/81$. Same! Phew!

• RUNNING EXAMPLE: $f_{X,Y}(x,y) = \frac{15}{32}xy^2$ for $0 \leq y \leq x \leq 2$, otherwise 0. Diagram:

• Valid joint density function?

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

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

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

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

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

  $\rightarrow$ 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 $x > 2$.)

  $\rightarrow$ 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!

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

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

  $\rightarrow$ 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 $y > 2$.)

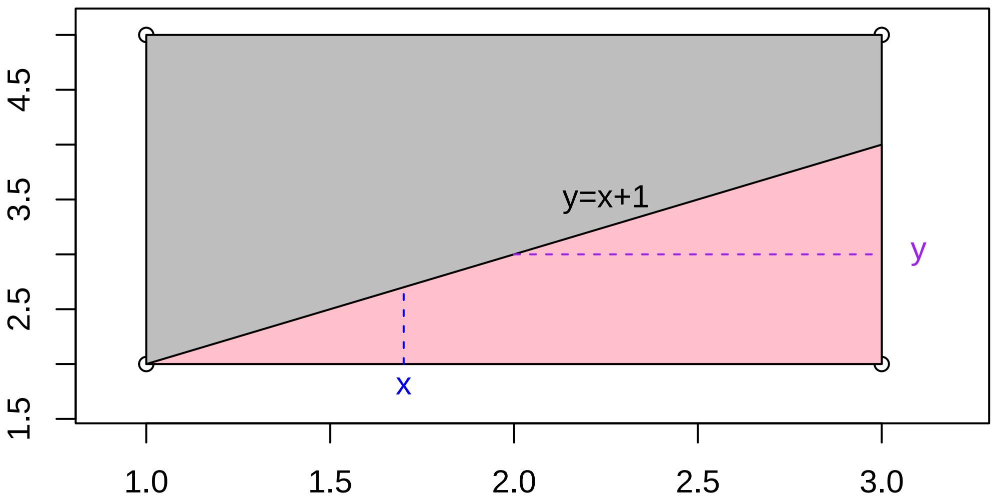
  $\rightarrow$ Check: $\int_{-\infty}^{\infty} f_Y(y)\,dy = \int_0^2 \frac{15}{64}(4y^2 - y^4)\,dy = \frac{15}{64}[4\frac{1}{3}y^3 - \frac{1}{5}y^5)]\Big|_{y=0}^{y=2} = \frac{15}{64}[4\frac{1}{3}(2^3 - 0^3) - \frac{1}{5}(2^5 - 0^5)] = 1$. Phew!

- BONUS EXAMPLE: Suppose X and Y have joint density function $f_{X,Y}(x,y) = \frac{1}{780}x^3y^2$ for $1 \le x \le 3$ and $2 \le y \le 5$, otherwise 0. What is $P(Y < X + 1)$?

- SOLUTION #1: Integrate in the order $dy\,dx$.

POLL: Then $P(Y < X + 1)$ is equal to: **(A)** $\int_1^3\left(\int_2^5 \frac{1}{780}x^3y^2\,dy\right)dx$. **(B)** $\int_1^3\left(\int_2^{x+1} \frac{1}{780}x^3y^2\,dy\right)dx$. **(C)** $\int_1^{y-1}\left(\int_2^{x+1} \frac{1}{780}x^3y^2\,dy\right)dx$. **(D)** $\int_1^{y-1}\left(\int_2^5 \frac{1}{780}x^3y^2\,dy\right)dx$.

- Need to integrate $f_{X,Y}(x,y)$ over the pink triangle:



$\rightarrow$ So x goes from 1 to 3.

→ And, for each x, y goes from 2 to $x+1$ (blue dashed line). (B) So,

$$\mathrm{P}(Y < X+1) \;=\; \int_1^3 \Big(\int_2^{x+1} \frac{1}{780} x^3 y^2 \, dy \Big) dx \;=\; \int_1^3 \Big(\frac{1}{780} x^3 \frac{y^3}{3} \Big|_{y=2}^{y=x+1} \Big) dx$$

$$= \int_1^3 \Big(\frac{1}{2340} x^3 [(x+1)^3 - 2^3] \Big) dx \;=\; \frac{1}{2340} \int_1^3 [x^6 + 3x^5 + 3x^4 - 7x^3] dx$$

$$= \frac{1}{2340} \Big[\frac{x^7}{7} + 3\frac{x^6}{6} + 3\frac{x^5}{5} - 7\frac{x^4}{4} \Big] \Big|_{x=1}^{x=3}$$

$$= \frac{1}{2340} \Big[\frac{3^7-1}{7} + 3\frac{3^6-1}{6} + 3\frac{3^5-1}{5} - 7\frac{3^4-1}{4} \Big]$$

$$= \frac{1}{2340} \Big[\frac{26382}{35} \Big] \;=\; \frac{5963}{20475} \;\doteq 0.291233\,.$$

- SOLUTION #2: Integrate in the order $dx\,dy$.

POLL: Then $\mathrm{P}(Y < X+1)$ is equal to: **(A)** $\int_2^5 \Big(\int_1^3 \frac{1}{780} x^3 y^2 \, dx \Big) dy$. **(B)** $\int_2^4 \Big(\int_{y-1}^3 \frac{1}{780} x^3 y^2 \, dx \Big) dy$. **(C)** $\int_2^5 \Big(\int_{y-1}^3 \frac{1}{780} x^3 y^2 \, dx \Big) dy$. **(D)** $\int_2^{x+1} \Big(\int_2^5 \frac{1}{780} x^3 y^2 \, dx \Big) dy$.

- Here y goes from 2 to 4 (not 5!).

→ And, for each y, x goes from $y-1$ to 3 (purple dashed line). (B) So,

$$\mathrm{P}(Y < X+1) \;=\; \int_2^4 \Big(\int_{y-1}^3 \frac{1}{780} x^3 y^2 \, dx \Big) dy \;=\; \int_2^4 \Big(\frac{1}{780} \frac{x^4}{4} y^2 \Big|_{x=y-1}^3 \Big) dy$$

$$= \int_2^4 \Big(\frac{1}{780} \frac{3^4 - (y-1)^4}{4} y^2 \Big) dy \;=\; \frac{1}{3120} \int_2^4 \Big[3^4 - (y-1)^4 \Big] y^2 \, dy$$

$$= \frac{1}{3120} \int_2^4 \Big[(3^4-1)y^2 - y^6 + 4y^5 - 6y^4 + 4y^3 \Big] \, dy$$

$$= \frac{1}{3120} \Big[(3^4-1)\frac{y^3}{3} - \frac{y^7}{7} + 4\frac{y^6}{6} - 6\frac{y^5}{5} + 4\frac{y^4}{4} \Big] \Big|_{y=2}^{y=4}$$

$$= \frac{1}{3120} \Big[(3^4-1)\frac{4^3-2^3}{3} - \frac{4^7-2^7}{7} + 4\frac{4^6-2^6}{6} - 6\frac{4^5-2^5}{5} + 4\frac{4^4-2^4}{4} \Big]$$

$$= \frac{1}{3120} \Big[\frac{95408}{105} \Big] \;=\; \frac{5963}{20475} \;\doteq 0.291233\,.$$

- So, we get the same answer either way, and either method is fine.

→ Both ways are a bit messy, but hopefully not too bad.

Suggested Homework: 2.7.4, 2.7.7, 2.7.8(a–c), 2.7.9, 2.7.14, 2.7.15, 2.7.16.

Conditioning and Independence for Discrete Random Variables (§2.8.1)

- 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.)

POLL: In this example, what is $\mathrm{P}(Y=5\,|\,X=3)$?
(A) 1/6. **(B)** 1/5. **(C)** 1/4. **(D)** 1/3. **(E)** 1/2. **(F)** 1.

- We compute here that $\mathrm{P}(Y=5\,|\,X=3)=\frac{\mathrm{P}(X=3,\ Y=5)}{\mathrm{P}(X=3)}=\frac{0.1}{0.3}=1/3$. (D)

  → 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 $p_{Y|X}(y\,|\,3)=0$.

  → 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.

END MONDAY #7

	$Y=5$	$Y=6$	$p_X(x)$
$X=2$	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	

POLL: In the above example, what is $\mathrm{P}(X\geq 3\,|\,Y=6)$?
(A) 2/3. **(B)** 3/4. **(C)** 4/5. **(D)** 5/6. **(E)** 6/7. **(F)** 7/8.

- What about independence?

- Most general 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. Optional.)

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

POLL: If X and Y are discrete and independent, which of these must be true?
(A) $\mathrm{P}(a \le X \le b,\ c \le Y \le d) = \mathrm{P}(a \le X \le b)\ \mathrm{P}(c \le Y \le d)$ for $a < b$ and $c < d$.
(B) $p_{X|Y}(x|y) = p_X(x)$ for all x, y with $p_Y(y) > 0$.
(C) $p_{Y|X}(y|x) = p_Y(y)$ for all x, y with $p_X(x) > 0$.
(D) $p_{X|Y}(x|y) = p_{Y|X}(y|x)$ for all x, y with $p_X(x), p_Y(y) > 0$.
(E) All of the above.
(F) Just three of the above.

- Well, (A) follows by taking $B = [a, b]$ and $C = [c, d]$ above.
- And, if X and Y are discrete and independent, then $p_{X|Y}(x \mid y) = \frac{\mathrm{P}(X=x,\ Y=y)}{\mathrm{P}(Y=y)} = \frac{\mathrm{P}(X=x)\ \mathrm{P}(Y=y)}{\mathrm{P}(Y=y)} = \mathrm{P}(X = x)$, showing (B).

  → Similarly, $p_{Y|X}(y \mid x) = \mathrm{P}(Y = y)$, showing (C).
- But (D) is false (and crazy!). So, the answer is (F): just three of the above.
- Independence 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 (§2.8.2)

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

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

  → Trick: Do it anyway!

  → We first consider certain limits …

  → Intuitively, imagine replacing the event $\{X = x\}$ by the event $\{x \le X \le x + \epsilon\}$ for some small $\epsilon > 0$, so that $\mathrm{P}(x \le X \le x + \epsilon) > 0$.

POLL: Suppose X and Y have continuous joint density $f_{X,Y}(x, y)$, and X has marginal density $f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)\, dy > 0$ for some x. Then for $a < b$,

$$\lim_{\epsilon \searrow 0} \mathrm{P}(a \le Y \le b \mid x \le X \le x + \epsilon)$$

is equal to: **(A)** $\int_{-\infty}^{\infty} \int_a^b f_{X,Y}(x, y)\, dx\, dy$. **(B)** $\int_{-\infty}^{\infty} \int_a^b f_{X,Y}(x, y)\, dy\, dx$. **(C)** $\int_a^b \frac{f_{X,Y}(x,y)}{f_X(x)}\, dy$. **(D)** $\frac{\int_a^b f_{X,Y}(x,y)\, dy}{\int_a^b f_X(x)\, dx}$. **(E)** $\frac{\int_a^b f_Y(y)\, dy}{\int_a^b f_X(x)\, dx}$. **(F)** No idea.

- We have that $\mathrm{P}(x \le X \le x + \epsilon) = \int_x^{x+\epsilon} f_X(u)\, du$.

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

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

- What about independence?

→ Idea: X and Y being independent should imply that $\mathrm{P}(a \le Y \le b \,|\, X = x) = \mathrm{P}(a \le Y \le b)$ for all $a < b$.

POLL: To ensure this, it suffices that:
(A) $f_{X,Y}(x,y) = 0$ for all x, y. **(B)** $f_{X,Y}(x,y) > f_X(x)$ for all x, y. **(C)** $f_{X,Y}(x,y) < f_Y(y)$ for all x, y. **(D)** $f_{X,Y}(x,y) = f_X(x)$ for all x, y. **(E)** $f_{X,Y}(x,y) = f_Y(y)$ for all x, y. **(F)** $f_{X,Y}(x,y) = f_X(x)\, f_Y(y)$ for all x, y.

- Definition: X and Y are independent if $f_{X,Y}(x,y) = f_X(x)\, f_Y(y)$ for "all" $x, y \in \mathbf{R}$.

→ Then $f_{Y|X}(y \,|\, x) = \frac{f_{X,Y}(x,y)}{f_X(x)} = \frac{f_X(x) f_Y(y)}{f_X(x)} = f_Y(y)$ whenever $f_X(x) > 0$.

→ And, $\mathrm{P}(a \le Y \le b \,|\, X = x) = \int_a^b f_{Y|X}(y \,|\, x)\, dy = \int_a^b f_Y(y)\, dy = \mathrm{P}(a \le Y \le b)$.

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

- 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.

POLL: In this example, are X and Y independent?
(A) Yes. **(B)** No. **(C)** No idea.

→ 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.

→ Indeed, 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 $0 \le Y \le X$.

Summary – Independence of Random Variables (§2.8)

- 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; optional)

→ $p_{X,Y}(x,y) = p_X(x)\, p_Y(y)$ 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.

• Note: We are omitting a few topics from the end of Chapter 2, including:

→ Order Statistics (sorted sample values, from smallest to largest). (§2.8.4)

→ Multivariable Change-Of-Variable Theorem. (§2.9)

→ Computer algorithms to simulate probability distributions. (§2.10)

→ All interesting! Check them out! Try the exercises! Ask me questions!

[**END OF TEXTBOOK CHAPTER #2**]

Expected Values: Discrete Case (§3.1)

• 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.

[**Reminder: Next week is READING WEEK – no classes!.**]

END WEDNESDAY #8

• 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 $p_i \ge 0$ 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 $\mathrm{E}(X) = c(1) = c$.

POLL: e.g. If X is the number showing on a fair six-sided die, then $\mathrm{E}(X) =$
(A) 2. **(B)** 2.5. **(C)** 3. **(D)** 3.5. **(E)** 4.

• Here $\mathrm{E}(X) = \sum_{x\in\mathbf{R}} x\,\mathrm{P}(X=x) = \sum_{k=1}^{6} k\,(1/6) = (1+2+3+4+5+6)/6 = 21/6 = 3.5$. (Not 3!)

POLL: e.g. If $X \sim \text{Bernoulli}(\theta)$, then $\mathrm{E}(X) =$
(A) 0. **(B)** θ. **(C)** 0.5. **(D)** $1-\theta$. **(E)** 1.

• Here $\mathrm{E}(X) = 0(1-\theta) + 1(\theta) = \theta$.

POLL: e.g. Suppose $Y \sim \text{Binomial}(n, \theta)$. What is $\mathrm{E}(Y)$? [Best guess.]
(A) θ. **(B)** $n + \theta$. **(C)** $n\theta$. **(D)** $n^2\theta$. **(E)** $n\theta(1-\theta)$. **(F)** No idea.

- Here $\mathrm{E}(Y) = \sum_{y \in \mathbf{R}} y \, \mathrm{P}(Y = y) = \sum_{k=0}^{n} k \binom{n}{k} \theta^k (1-\theta)^{n-k} = \sum_{k=0}^{n} k \frac{n!}{(n-k)!\,k!} \theta^k (1-\theta)^{n-k} = \sum_{k=1}^{n} n \frac{(n-1)!}{(n-k)!\,(k-1)!} \theta^k (1-\theta)^{n-k} = n\theta \sum_{k=1}^{n} \binom{n-1}{k-1} \theta^{k-1} (1-\theta)^{n-k}$.
  → 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, $\mathrm{E}(Z) = \frac{1-\theta}{\theta} = \frac{1}{\theta} - 1$. 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 $\mathrm{E}(X) = \sum_{x \in \mathbf{R}} x \, \mathrm{P}(X = x) = \sum_{k=0}^{\infty} k \, e^{-\lambda} \lambda^k / k! = e^{-\lambda} \lambda \left[\sum_{k=1}^{\infty} \lambda^{k-1}/(k-1)!\right] = e^{-\lambda} \lambda \left[\sum_{\ell=0}^{\infty} \lambda^\ell / \ell!\right] = e^{-\lambda} \lambda \, [e^\lambda] = \lambda$.

POLL: 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 $\mathrm{E}(X)$ equals ...
(A) 1. **(B)** 2. **(C)** 4. **(D)** 8. **(E)** ∞.

- Here $\mathrm{E}(X) = \sum_{k=1}^{\infty} (2^k)(2^{-k}) = \sum_{k=1}^{\infty} (1) = 1 + 1 + 1 + \ldots = \infty$.

POLL: In this same example, what is $\mathrm{P}(X < \infty)$?
(A) 1/2. **(B)** 3/4. **(C)** 7/8. **(D)** 1.

- Here $\mathrm{P}(X < \infty) = \sum_{k=1}^{\infty} \mathrm{P}(X = 2^k) = \sum_{k=1}^{\infty} 2^{-k} = (1/2) + (1/4) + (1/8) + \ldots = 1$.
  → 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 $\mathrm{E}(Z) = \mathrm{E}(g(X)) = \sum\limits_{z \in \mathbf{R}} z \, \mathrm{P}(Z{=}z) = \sum\limits_{z \in \mathbf{R}} z \, \mathrm{P}(g(X){=}z) = \sum\limits_{z \in \mathbf{R}} z \sum_{x: g(x)=z} \mathrm{P}(X{=}x) = \sum\limits_{z \in \mathbf{R}} \sum_{x: g(x)=z} g(x) \, \mathrm{P}(X{=}x) = \sum\limits_{x \in \mathbf{R}} g(x) \, \mathrm{P}(X{=}x)$.
  → Or, if $Z = h(X, Y)$, $\mathrm{E}(Z) = \sum\limits_{z \in \mathbf{R}} z \, \mathrm{P}(Z{=}z) = \sum\limits_{x, y \in \mathbf{R}} h(x, y) \, \mathrm{P}(X{=}x, Y{=}y)$.
  → (Here Z is also discrete; and get the same expected value either way.)

POLL: e.g. if $X \sim \text{Binomial}(2, 1/3)$, then what is $\mathrm{E}(X^2)$?
(A) 2/3. **(B)** 4/3. **(C)** 4/9. **(D)** 8/9. **(E)** 10/9. **(F)** 16/9.

- Since $X \sim \text{Binomial}(2, 1/3)$, then know $\mathrm{E}(X) = n\theta = 2(1/3) = 2/3$, but also $\mathrm{E}(X^2) = \sum_{x \in \mathbf{R}} x^2 \, \mathrm{P}(X = x) = \sum_{k=0}^{2} k^2 \binom{2}{k} (1/3)^k (2/3)^{2-k} = 0 + 1^2 \cdot 2(1/3)(2/3) + 2^2 \cdot 1(1/3)^2 = 4/9 + 4/9 = 8/9$.
- e.g. if $X \sim \text{Binomial}(3, 1/4)$, then what is $\mathrm{E}(5X^2)$?

→ Since $X \sim \text{Binomial}(3, 1/4)$, then know $\text{E}(X) = 3(1/4) = n\theta = 3/4$, but also $\text{E}(5X^2) = \sum_{x \in \mathbf{R}} 5x^2 \, \text{P}(X = x) = \sum_{k=0}^{3} 5k^2 \binom{3}{k} (1/4)^k (3/4)^{3-k}$
$= 5(0)^2 \binom{3}{0}(1/4)^0(3/4)^3 + 5(1)^2\binom{3}{1}(1/4)^1(3/4)^2 + 5(2)^2\binom{3}{2}(1/4)^2(3/4)^1 + 5(3)^2\binom{3}{3}(1/4)^3(3/4)^0$
$= 0 + 5 \cdot 1 \cdot 3 \cdot 3^2/4^3 + 5 \cdot 4 \cdot 3 \cdot 3/4^3 + 5 \cdot 9 \cdot 1 \cdot 1/4^3 = 45/8 = 5.625.$

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, $\text{E}(Z) = \sum_{z \in \mathbf{R}} z \, \text{P}(Z = z) = \sum_{x,y \in \mathbf{R}} (ax + by) \, \text{P}(X = x, Y = y)$
$= a \sum_{x,y \in \mathbf{R}} x \, \text{P}(X = x, Y = y) + b \sum_{x,y \in \mathbf{R}} y \, \text{P}(X = x, Y = y)$
$= a \sum_{x \in \mathbf{R}} x \sum_{y \in \mathbf{R}} \text{P}(X = x, Y = y) + b \sum_{y \in \mathbf{R}} y \sum_{x \in \mathbf{R}} \text{P}(X = x, Y = y)$
$= a \sum_{x \in \mathbf{R}} x \, \text{P}(X = x) + b \sum_{y \in \mathbf{R}} y \, \text{P}(Y = y) = a \, \text{E}(X) + b \, \text{E}(Y)$. 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, $\text{E}(Y) = \text{E}(X_1) + \text{E}(X_2) + \ldots + \text{E}(X_n) = \theta + \theta + \ldots + \theta = n\theta$.

→ Same answer as before! Easier!

POLL: e.g. Suppose $X \sim \text{Binomial}(5, 1/4)$, and $Y \sim \text{Geometric}(1/3)$, and $Z = 2X - 3Y$. Then $\text{E}(Z)$ equals: **(A)** 5. **(B)** 3.5. **(C)** -1. **(D)** -3.5. **(E)** -7.

→ By linearity and the above, $\text{E}(Z) = \text{E}(2X - 3Y) = 2\,\text{E}(X) - 3\,\text{E}(Y) = 2[(5)(1/4)] - 3[\frac{2/3}{1/3}] = (10/4) - 6 = (10/4) - (24/4) = -14/4 = -3.5$. (Negative!)

- Caution: This is only for linear functions! e.g. If $X \sim \text{Bernoulli}(1/2)$, then $\text{E}(X^2) = \text{E}(X) = 1/2$, which is not the same as $\big(\text{E}(X)\big)^2 = (1/2)^2 = 1/4$.

- 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 $\text{P}(X \le Y) = 1$.

→ Let $Z = Y - X$. Then Z is discrete, and $\text{P}(Z \ge 0) = 1$.

→ So, $\text{P}(Z = z) = 0$ whenever $z < 0$.

→ Hence, $\text{E}(Z) = \sum_{z \in \mathbf{R}} z \, \text{P}(Z = z) = \sum_{z \in [0,\infty)} z \, \text{P}(Z = z) \ge 0$.

→ But $\text{E}(Z) = \text{E}(Y - X) = \text{E}(Y) - \text{E}(X)$, so $\text{E}(Y) - \text{E}(X) \ge 0$, i.e. $\text{E}(X) \le \text{E}(Y)$.

→ This is the monotonicity property: If $\text{P}(X \le Y) = 1$, then $\text{E}(X) \le \text{E}(Y)$.

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 $\text{E}(XY) = \sum_{x,y \in \mathbf{R}} xy \, \text{P}(X = x, \, Y = y) = \sum_{x,y \in \mathbf{R}} xy \, \text{P}(X = x) \, \text{P}(Y = y) = \Big(\sum_{x \in \mathbf{R}} x \, \text{P}(X = x)\Big)\Big(\sum_{y \in \mathbf{R}} y \, \text{P}(Y = y)\Big) = \text{E}(X) \, \text{E}(Y)$. Useful!

POLL: 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 $\text{E}(Z)$ equals:
(A) 0. **(B)** 0.5. **(C)** 1.5. **(D)** 2.5. **(E)** 3.5. **(F)** 4.5.

- Here $\text{E}(Z) = \text{E}(XY) = \text{E}(X)\,\text{E}(Y) = [(5)(1/4)]\,[\frac{2/3}{1/3}] = [5/4]\,[2] = 10/4 = 2.5$.

- e.g. Suppose $X \sim$ 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 $\mathrm{E}(XY) \neq \mathrm{E}(X)\,\mathrm{E}(Y)$. 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 (§3.2)

- 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 $X \sim$ Uniform$[0, 1]$, then what is $\mathrm{E}(X)$? We compute that:

$\mathrm{E}(X) = \int_{-\infty}^{\infty} x\, f_X(x)\,dx = \int_0^1 x\,(1)\,dx = \frac{1}{2}x^2\Big|_{x=0}^{x=1} = \frac{1}{2}(1^2 - 0^2) = \frac{1}{2}$.

POLL: e.g. If $X \sim$ Uniform$[L, R]$, then what is $\mathrm{E}(X)$?
(A) L. **(B)** R. **(C)** $L + R$. **(D)** $(L + R)/2$. **(E)** LR. **(F)** $LR/2$.

- We compute that: $\mathrm{E}(X) = \int_{-\infty}^{\infty} x\ f_X(x)\,dx = \int_L^R x\ (\frac{1}{R-L})\,dx = \frac{1}{2}x^2\ (\frac{1}{R-L})\Big|_{x=L}^{x=R} =$ $\frac{1}{2}\ (\frac{1}{R-L})(R^2 - L^2) = \frac{1}{2}\ (\frac{1}{R-L})(R - L)(R + L) = \frac{1}{2}(R + L) = (L + R)/2$.
  - → e.g. If $X \sim$ Uniform$[-8, 2]$, then $\mathrm{E}(X) = \frac{1}{2}(-8 + 2) = -3$. Negative!
- If $Y \sim$ Exponential(λ), 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.)

POLL: If $Z \sim$ Normal(0, 1), then $\mathrm{E}(Z) = \int_{-\infty}^{\infty} z\ \phi(z)\,dz = \int_{-\infty}^{\infty} z\ \frac{1}{\sqrt{2\pi}}\,e^{-z^2/2}\,dz = ??$
(A) $-\sqrt{2\pi}$. **(B)** -1. **(C)** 0. **(D)** 1. **(E)** $\sqrt{2\pi}$. **(F)** No idea.

  - → This integrand is an "odd" function, so by symmetry, $\mathrm{E}(Z) = 0$.
  - → (Alternatively, $-e^{-z^2/2}$ is an anti-derivative of $z\,e^{-z^2/2}$.)
- Now suppose $W \sim$ Normal(μ, σ^2). Then what is $\mathrm{E}(W)$?
  - → Well, this means that $W = \mu + \sigma Z$ where $Z \sim$ 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 (§3.3)

• 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 $\mu_X := \mathrm{E}(X)$.

→ 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, $\mathrm{Sd}(X) := \sqrt{\mathrm{Var}(X)} = \sqrt{\mathrm{E}[(X - \mu_X)^2]}$.

END MONDAY #8

• e.g. $X \sim \mathrm{Bernoulli}(\theta)$. Then $\mu_X := \mathrm{E}(X) = \theta$, 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 $\mathrm{Var}(X) := \mathrm{E}[(X - \mu_X)^2] = \mathrm{E}[X^2 - 2X(\mu_X) + (\mu_X)^2] = \mathrm{E}[X^2] - 2\mathrm{E}[X](\mu_X) + (\mu_X)^2 = \mathrm{E}[X^2] - 2(\mu_X)(\mu_X) + (\mu_X)^2 = \mathrm{E}[X^2] - (\mu_X)^2$.

→ 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?

POLL: If $Y \sim$ Uniform$[0,1]$, know $\mu_Y = 1/2$. What is $\mathrm{E}(Y^2)$?
(A) 1/4. **(B)** 1/3. **(C)** 1/2. **(D)** 2/3. **(E)** 3/4. **(F)** 1.

→ Here $\mathrm{E}(Y^2) = \int_{-\infty}^{\infty} y^2\, f_Y(y)\, dy = \int_0^1 y^2\,(1)\, dy = \frac{1}{3}y^3\Big|_{y=0}^{y=1} = \frac{1}{3}(1^3 - 0^3) = \frac{1}{3}$.

POLL: If $Y \sim$ Uniform$[0,1]$, know $\mu_Y = 1/2$ and $\mathrm{E}(Y^2) = 1/3$. What is $\mathrm{Var}(Y)$?
(A) 1/2. **(B)** 1/3. **(C)** 1/4. **(D)** 1/6. **(E)** 1/8. **(F)** 1/12.

→ Here $\mathrm{Var}(Y) = \mathrm{E}(Y^2) - (\mu_Y)^2 = (1/3) - (1/2)^2 = (1/3) - (1/4) = 1/12$.

→ So then $\mathrm{Sd}(Y) = \sqrt{\mathrm{Var}(Y)} = \sqrt{1/12} = 1/\sqrt{12}$.

- Suppose $Z \sim$ 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 $X_1 \sim$ Uniform$[4.9, 5.1]$, $X_2 \sim$ Uniform$[4,6]$, and $X_3 \sim$ Uniform$[0,10]$, then: $\mathrm{Var}(X_1) = (0.2)^2/12 \doteq 0.0033$, $\mathrm{Var}(X_2) = (2)^2/12 = 1/3 \doteq 0.033$, and $\mathrm{Var}(X_3) = (10)^2/12 = 100/12 \doteq 8.33$. So $\mathrm{Var}(X_3) \gg \mathrm{Var}(X_2) \gg \mathrm{Var}(X_1)$, which makes sense.

- In general, $(X-\mu_X)^2 \geq 0$, so always have $\mathrm{Var}(X) := \mathrm{E}[(X-\mu_X)^2] \geq 0$.

→ But $\mathrm{Var}(X) = \mathrm{E}[X^2] - (\mu_X)^2$, so $\mathrm{E}[X^2] - (\mu_X)^2 \geq 0$, i.e. $\mathrm{E}[X^2] \geq (\mu_X)^2$.

→ And, since $(\mu_X)^2 \geq 0$, always have $\mathrm{Var}(X) = \mathrm{E}[X^2] - (\mu_X)^2 \leq \mathrm{E}[X^2]$, too.

POLL: If $a, b \in \mathbf{R}$, then $\mathrm{Var}(aX+b)$ is equal to: **(A)** $\mathrm{Var}(X)$. **(B)** $a\,\mathrm{Var}(X)$. **(C)** $a\,\mathrm{Var}(X) + b$. **(D)** $a^2\,\mathrm{Var}(X)$. **(E)** $a^2\,\mathrm{Var}(X) + b$. **(F)** No idea.

- Using linearity, $\mathrm{Var}(aX+b) = \mathrm{E}[(aX+b-\mu_{aX+b})^2] = \mathrm{E}[(aX+b-a\mu_X-b)^2] = \mathrm{E}[(a(X-\mu_X))^2] = a^2\,\mathrm{E}[(X-\mu_X)^2] = a^2\,\mathrm{Var}(X)$. (Note: a^2, not a. And b irrelevant.)

→ Hence, $\mathrm{Sd}(aX+b) = \sqrt{\mathrm{Var}(aX+b)} = \sqrt{a^2\,\mathrm{Var}(X)} = |a|\,\mathrm{Sd}(X)$.

- e.g. If $Z \sim$ Uniform$[L,R]$, can write $Z = L + (R-L)U$ where $U \sim$ Uniform$[0,1]$.

→ Hence, $\mathrm{Var}(Z) = (R-L)^2\,\mathrm{Var}(U) = (R-L)^2/12$. Same as before. Easier!

- What about $\mathrm{Var}(X+Y)$ or $\mathrm{Var}(aX+bY)$? Later!

- e.g. $W \sim$ Exponential(λ). Know $\mu_W := \mathrm{E}(W) = 1/\lambda$. $\mathrm{Var}(W) = ??$

→ Well, $\mathrm{E}(W^2) = \int_{-\infty}^{\infty} w^2\, f_W(w)\, dw = \int_0^\infty w^2\, \lambda e^{-\lambda w}\, dw$.

→ 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, $\mathrm{E}(W^2) = 2\,\frac{1}{\lambda}\,\frac{1}{\lambda} = 2/\lambda^2$.

→ Then $\mathrm{Var}(W) = \mathrm{E}(W^2) - (\mu_W)^2 = (2/\lambda^2) - (1/\lambda)^2 = 1/\lambda^2$. Phew!

→ Hence, $\mathrm{Sd}(W) = 1/\lambda$.

- 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: $\text{E}(X^2) = \sum_{k=0}^{\infty} k^2 e^{-\lambda} \frac{\lambda^k}{k!} = \lambda e^{-\lambda} \sum_{k=1}^{\infty} \Big((k-1) + 1\Big) \frac{\lambda^{k-1}}{(k-1)!} = \lambda e^{-\lambda} \Big(\lambda \sum_{k=2}^{\infty} \frac{\lambda^{k-2}}{(k-2)!} + \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!}\Big) = \lambda e^{-\lambda} \left(\lambda e^{\lambda} + e^{\lambda}\right) = \lambda^2 + \lambda$.

  → Then $\text{Var}(X) = \text{E}(X^2) - (\text{E}(X))^2 = (\lambda^2 + \lambda) - (\lambda)^2 = \lambda$. Phew! Simple!

  → And then $\text{Sd}(X) = \sqrt{\lambda}$, so X is usually within about $\pm\sqrt{\lambda}$ of λ.

- What about variance of $\text{Geometric}(\theta)$?

  → Messy sum ... works out to $(1-\theta)/\theta^2$. [Problem 3.3.18; optional.]

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).

Covariance and Correlation (§3.3)

- We know that $\text{E}(X+Y) = \text{E}(X) + \text{E}(Y)$. What about $\text{Var}(X+Y)$?

- Well, $\text{Var}(X+Y) = \text{E}[(X+Y-\mu_{X+Y})^2] = \text{E}[(X+Y-\mu_X-\mu_Y)^2] = \text{E}[((X-\mu_X)+(Y-\mu_Y))^2] = \text{E}[(X-\mu_X)^2 + (Y-\mu_Y)^2 + 2(X-\mu_X)(Y-\mu_Y)]$.

- This equals $\text{Var}(X) + \text{Var}(Y) + 2\,\text{Cov}(X,Y)$, where $\text{Cov}(X,Y) := \text{E}[(X-\mu_X)(Y-\mu_Y)]$ is the covariance of X and Y.

  → We always have $\text{Cov}(X,Y) = \text{Cov}(Y,X)$.

  → If $\text{Cov}(X,Y) > 0$, then X and Y tend to increase together.

  → If $\text{Cov}(X,Y) < 0$, then X and Y tend to increase oppositely.

- Special case: If $Y = X$, then $\text{Cov}(X,Y) = \text{Cov}(X,X) = \text{E}[(X-\mu_X)(X-\mu_X)] = \text{E}[(X-\mu_X)^2] = \text{Var}(X)$. In particular, $\text{Cov}(X,X) \ge 0$.

  → Or, if $Y = -X$, then $\text{Cov}(X,Y) = \text{Cov}(X,-X) = \text{E}[(X-\mu_X)(-X-\mu_{-X})] = \text{E}[-(X-\mu_X)^2] = -\text{Var}(X)$. In particular, $\text{Cov}(X,-X) \le 0$.

POLL: If X and Y are independent, $\text{Cov}(X,Y)$ equals: **(A)** 0. **(B)** $\text{Var}(X+Y)$. **(C)** $\text{Var}(X) + \text{Var}(Y)$. **(D)** $\text{Var}(X)\,\text{Var}(Y)$. **(E)** $\sqrt{\text{Var}(X)\,\text{Var}(Y)}$. **(F)** No idea.

END WEDNESDAY #9

- If X and Y are independent, then so are $X - \mu_X$ and $Y - \mu_Y$.

  → So, $\mathrm{E}[(X-\mu_X)(Y-\mu_Y)] = \mathrm{E}[X-\mu_X]\,\mathrm{E}[Y-\mu_Y]$.

  → But by linearity, $\mathrm{E}[X-\mu_X] = \mathrm{E}(X) - \mu_X = \mu_X - \mu_X = 0$.

  → So, $\mathrm{Cov}(X,Y) = \mathrm{E}[(X-\mu_X)(Y-\mu_Y)] = \mathrm{E}[X-\mu_X]\,\mathrm{E}[Y-\mu_Y] = [\mu_X - \mu_X]\,[\mu_Y - \mu_Y] = 0 \cdot 0 = 0$. (A)

  → 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.

  → If X and Y are independent, then $\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$.

  → 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?)

POLL: If $Y \sim \mathrm{Binomial}(n,\theta)$, then what is $\mathrm{Var}(Y)$?
(A) θ. **(B)** $\theta(1-\theta)$. **(C)** $n\theta$. **(D)** $n\theta(1-\theta)$. **(E)** $n^2\theta(1-\theta)$. **(F)** No idea.

- Well, 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 $\mathrm{Cov}(aX+bY, Z)$?

POLL: $\mathrm{Cov}(aX+bY,Z)$ is equal to: **(A)** $\mathrm{Cov}(X+Y,Z)$. **(B)** $\mathrm{Cov}(X,Z) + \mathrm{Cov}(Y,Z)$. **(C)** $(a+b)[\mathrm{Cov}(X,Z)+\mathrm{Cov}(Y,Z)]$. **(D)** $(a+b)^2[\mathrm{Cov}(X,Z)+\mathrm{Cov}(Y,Z)]$. **(E)** $a\,\mathrm{Cov}(X,Z) + b\,\mathrm{Cov}(Y,Z)$. **(F)** $a^2\,\mathrm{Cov}(X,Z) + b^2\,\mathrm{Cov}(Y,Z)$.

  → Here $\mathrm{Cov}(aX+bY,Z) = \mathrm{E}[(aX+bY-\mu_{aX+bY})(Z-\mu_Z)]$
$= \mathrm{E}[(aX+bY-a\mu_X-b\mu_Y)(Z-\mu_Z)] = \mathrm{E}[(a(X-\mu_X)+b(Y-\mu_Y))(Z-\mu_Z)]$
$= a\,\mathrm{E}[(X-\mu_X)(Z-\mu_Z)] + b\,\mathrm{E}[(Y-\mu_Y))(Z-\mu_Z)] = a\,\mathrm{Cov}(X,Z) + b\,\mathrm{Cov}(Y,Z)$.

  → 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).

  → Hence, 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 $X + Y \sim \mathrm{Normal}(\mu_1 + \mu_2,\, \sigma_1^2 + \sigma_2^2)$.

→ So, also, $\mathrm{Sd}(X+Y) = \sqrt{\mathrm{Var}(X+Y)} = \sqrt{\sigma_1^2 + \sigma_2^2}$.

- 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 $\mathrm{Sd}(X)\,\mathrm{Sd}(Y) = |c|\,\mathrm{Var}(X)$.

POLL: If $Y = cX$ for some constant $c \in \mathbf{R}$, then $\mathrm{Cov}(X,Y)$ equals:
(A) 0. **(B)** $\mathrm{Sd}(X)$. **(C)** $\mathrm{Var}(X)$. **(D)** $c\,\mathrm{Var}(X)$. **(E)** $|c|\,\mathrm{Var}(X)$. **(F)** $c^2\,\mathrm{Var}(X)$.

→ Indeed, here $\mathrm{Cov}(X,Y) = \mathrm{Cov}(X,cX) = c\,\mathrm{Cov}(X,X) = c\,\mathrm{Var}(X)$. (D)

→ So, if $Y = cX$ where $c \geq 0$, then $\mathrm{Cov}(X,Y) = \mathrm{Sd}(X)\,\mathrm{Sd}(Y)$.

→ Or, if $Y = cX$ where $c < 0$, then $\mathrm{Cov}(X,Y) = -\mathrm{Sd}(X)\,\mathrm{Sd}(Y)$.

- FACT: These are the extremes: <u>always</u> $-\mathrm{Sd}(X)\,\mathrm{Sd}(Y) \leq \mathrm{Cov}(X,Y) \leq \mathrm{Sd}(X)\,\mathrm{Sd}(Y)$.

→ 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 $\mathrm{Corr}(X,Y) = \mathrm{Cov}(X,Y)/\sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)}$.

→ So, from the above, 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 $\mathrm{Corr}(X,Y) = \mathrm{Cov}(X,Y)\;/\;[\mathrm{Sd}(X)\,\mathrm{Sd}(Y)]$.

→ (Requires first computing μ_X, μ_Y, $\mathrm{Var}(X)$, $\mathrm{Var}(Y)$, $\mathrm{Cov}(X,Y)$, $\ldots$.)

- If X, Y <u>independent</u>, then $\mathrm{Cov}(X,Y) = 0$, so $\mathrm{Corr}(X,Y) = 0$. ("uncorrelated")

- Now suppose that Y is a <u>constant</u> 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 <u>only</u> defined for <u>non-constant</u> r.v.: $\mathrm{Var}(X) > 0$ and $\mathrm{Var}(Y) > 0$.

POLL: Suppose $Z = cY$ for some $c > 0$. Then $\mathrm{Corr}(X,Z)$ is equal to:
(A) $\mathrm{Corr}(X,Y)$. **(B)** $c\,\mathrm{Corr}(X,Y)$. **(C)** $c^2\,\mathrm{Corr}(X,Y)$. **(D)** $\mathrm{Corr}(X,Y)^2$.
(E) $c\,\mathrm{Corr}(X,Y)^2$. **(F)** $c^2\,\mathrm{Corr}(X,Y)^2$.

→ 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, $\mathrm{Corr}(X,Z) = \frac{\mathrm{Cov}(X,Z)}{\mathrm{Sd}(X)\,\mathrm{Sd}(Z)} = \frac{c\,\mathrm{Cov}(X,Y)}{\mathrm{Sd}(X)\,c\,\mathrm{Sd}(Y)} = \frac{\mathrm{Cov}(X,Y)}{\mathrm{Sd}(X)\,\mathrm{Sd}(Y)} = \mathrm{Corr}(X,Y)$. (A)

→ That is, $\mathrm{Corr}(X,cY) = \mathrm{Corr}(X,Y)$. <u>Unaffected</u> by the constant scale $c > 0$.

- If instead $Z = cY$ 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.

- EXAMPLE: 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 ⇒ uncorrelated, but uncorrelated $\not\Rightarrow$ independent.

Markov's Inequality (§3.6)

POLL: Suppose $X \geq 0$, and $\mathrm{E}(X) = 5$. What is the largest that $\mathrm{P}(X \geq 100)$ could be? **(A)** 1/5. **(B)** 1/10. **(C)** 1/20. **(D)** 1/100. **(E)** 1. **(F)** No idea.

- Can't be too large, or we would have $\mathrm{E}(X) \geq (100)\,\mathrm{P}(X \geq 100) \gg 5$.

  → $\mathrm{E}(X) = 5$ implies $(100)\,\mathrm{P}(X{\geq}100) \leq 5$, so $\mathrm{P}(X{\geq}100) \leq 5/100 = 1/20$. (C)

- 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 $\mathrm{E}(Z) = \mathrm{E}[a\, I_{X \ge a}] = a\, \mathrm{P}(X \ge a)$. Or, to write it differently, $\mathrm{E}(Z) = 0\, \mathrm{P}(Z = 0) + a\, \mathrm{P}(Z = a) = a\, \mathrm{P}(Z = a) = a\, \mathrm{P}(X \ge a)$.

→ So, $\mathrm{E}(X) \ge \mathrm{E}(Z) = a\ \mathrm{P}(X \ge a)$. Hence, $\mathrm{P}(X \ge a) \le \mathrm{E}(X) \,/\, a$. ∎

- e.g. If $X \ge 0$ and $\mathrm{E}(X) = 5$, then must have $\mathrm{P}(X \ge 100) \le 5/100 = 1/20$.

→ Also, $\mathrm{P}(X \ge 1000) \le 5/1000 = 1/200$. Small!

- But this is only for non-negative random variables. Better is …

Chebychev's Inequality (§3.6)

- Let Y be any random variable, with finite mean μ_Y.

→ If $\mathrm{Var}(Y)$ is small, then Y will "usually" be "pretty close" to μ_Y. More precise?

- Chebychev's Inequality: For any $a > 0$, $\mathrm{P}(|Y - \mu_Y| \ge a) \le \mathrm{Var}(Y) \,/\, a^2$.

- Proof: Let $X = (Y - \mu_Y)^2 \ge 0$. Then by Markov's Inequality, $\mathrm{P}(|Y - \mu_Y| \ge a) = \mathrm{P}((Y - \mu_Y)^2 \ge a^2) \le \mathrm{E}((Y - \mu_Y)^2) \,/\, a^2 = \mathrm{Var}(Y) \,/\, a^2$. ∎

- e.g. Suppose Z has mean 5 and variance 9. Then, $\mathrm{P}(Z \ge 17) = \mathrm{P}(Z - 5 \ge 12) \le \mathrm{P}(|Z - 5| \ge 12) \le 9/12^2 = 9/144 = 1/16 = 0.0625$. Unlikely!

→ And, this is true for any random variable with this mean and variance.

→ If we also knew that $Z \ge 0$, then we could use Markov's inequality directly to get that $\mathrm{P}(Z \ge 17) \le \mathrm{E}(Z)/17 = 5/17 \doteq 0.294$. (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.

- Note: We are omitting a few topics from Chapter 3, including:

→ Generating Functions (§3.4)

→ Conditional Expectation (§3.5)

→ Jensen's Inequality (§3.6.1)

→ General Expectations (neither discrete nor continuous) (§3.7)

→ All interesting! Check them out! Try the exercises! Ask me questions!

[END OF TEXTBOOK CHAPTER #3]

[Reminder: Midterm #2 on Wednesday in EX100 at normal class time..]

END MONDAY #9

(Midterm #2.)

END WEDNESDAY #10

Convergence of Random Variables (§4.2)

- 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 (§4.2)

- 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 $X_n \sim \text{Bernoulli}(\frac{1}{n})$, 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 $X_n \xrightarrow{P} Y$.

POLL: Let $U \sim \text{Uniform}[0,1]$, 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? **(A)** Yes. **(B)** No. **(C)** Not sure.

→ Well, for any $\epsilon > 0$, $\mathrm{P}(|X_n - Y| \geq \epsilon) \leq \mathrm{P}(X_n \neq Y) = \mathrm{P}(X_n = 1$ and $Y = 0) = \mathrm{P}[1/2 < U \leq (1/2) + (1/2^n)] = 1/2^n$.

→ And, this probability $\to 0$ as $n \to \infty$. So, yes!

POLL: Let $Y \sim \text{Uniform}[0,5]$, and $X_n = (1 + \frac{1}{n})Y$. Does $X_n \to Y$ in probability? **(A)** Yes. **(B)** No. **(C)** Not sure.

→ 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, $\lim_{n\to\infty} \mathrm{P}(|X_n - Y| \geq \epsilon) = 0$, i.e. $X_n \to Y$ in probability. Yes!

POLL: 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? **(A)** Yes. **(B)** No. **(C)** Not sure.

→ Here for $0 < \epsilon < 1/2$, we have $\text{P}(|X_n - (1/2)| \geq \epsilon) = 1$.

→ This does not $\to 0$. So, no!

→ 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.

Weak Law of Large Numbers (WLLN) (§4.2.1)

• 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. Mean? Variance?

→ First, by linearity, $\text{E}(M_n) = \frac{1}{n}[\text{E}(X_1) + \text{E}(X_2) + \ldots + \text{E}(X_n)] = \frac{1}{n}[n\,\mu] = \mu$.

→ Then, since the $\{X_n\}$ are independent, $\text{Var}(M_n) = (\frac{1}{n})^2[\text{Var}(X_1) + \text{Var}(X_2) + \ldots + \text{Var}(X_n)] \leq (\frac{1}{n})^2[v + v + \ldots + v] = (\frac{1}{n})^2[n\,v] = v/n$. (Not just v.)

→ Now, let $\epsilon > 0$, and consider $\text{P}(|M_n - \mu| \geq \epsilon)$.

→ Use Chebychev's Inequality! Since $\text{E}(M_n) = \mu$, therefore $\text{P}(|M_n - \mu| \geq \epsilon) \leq \text{Var}(M_n)/\epsilon^2 \leq v/n\epsilon^2$, 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, $\text{P}(a \leq X_n \leq b)$ 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 (later): 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 $\text{E}(X_n) = 1/2 =: \mu$, and $\text{Var}(X_n) = (1/2)(1 - (1/2)) = 1/4 =: v < \infty$.

→ 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, $\text{P}(|M_n - (1/2)| \geq \epsilon) \to 0$ for all $\epsilon > 0$.

→ e.g. $\epsilon = 0.003$: $\text{P}(|M_n - (1/2)| \geq 0.003) \to 0$.

→ So, for all sufficiently large n, $\text{P}(|M_n - (1/2)| \geq 0.003) < 0.01$ (say).

→ In particular, for those n, $\text{P}(M_n - (1/2) \geq 0.003) < 0.01$, i.e. $\text{P}(M_n \geq 0.503) < 0.01$, i.e. $\text{P}(M_n < 0.503) > 0.99$, 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 $1 \le X_n \le 6$, so $|X_n - 3.5| \le 2.5$, so $\mathrm{Var}(X_n) = \mathrm{E}(|X_n - 3.5|^2) \le (2.5)^2 =: v < \infty$, 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.

POLL: Suppose we repeatedly take free throws, with independent probability $\theta = 1/4$ of scoring each time. About how many attempts will it take to score 500 times? **(A)** 500. **(B)** 1000. **(C)** 1500. **(D)** 2000. **(E)** 2500. **(F)** No idea.

→ 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 $X_n \sim \text{Geometric}(1/4)$, 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" $= Z_1 + Z_2 + \ldots + Z_{500}$.

→ Recall (Problem 3.3.18): Geometric(θ) has finite variance $v = (1-\theta)/\theta^2 < \infty$.

→ (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) (§4.3)

- 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,...,5,5,7,5,5,......,5,5,7,5,5, ...,5,5,7,5,5,...........

POLL: Does this sequence $X_1, X_2, X_3, \ldots$ converge to 5 in probability?
(A) Yes. **(B)** No. **(C)** Not sure.

→ Well, for $1 \leq n \leq 9$, $\mathrm{P}(X_n = 7) = 1/9$ and $\mathrm{P}(X_n = 5) = 1 - [1/9]$.

→ And, for $10 \leq n \leq 99$, $\mathrm{P}(X_n = 7) = 1/90$ and $\mathrm{P}(X_n = 5) = 1 - [1/90]$.

→ And, for $100 \leq n \leq 999$, $\mathrm{P}(X_n = 7) = 1/900$ and $\mathrm{P}(X_n = 5) = 1 - [1/900]$.

→ And, for $1000 \leq n \leq 9999$, $\mathrm{P}(X_n = 7) = 1/9000$ 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 $\epsilon > 0$, $\lim_{n\to\infty} \mathrm{P}(|X_n - 5| \geq \epsilon) \leq \lim_{n\to\infty} \mathrm{P}(|X_n - 5| \neq 0) = \lim_{n\to\infty} \mathrm{P}(X_n = 7) = 0$.

→ 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 $n \geq N$, we have $|X_n - 5| \leq \epsilon$.

→ 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 $X_n \overset{a.s.}{\not\to} 5$. i.e. we do <u>not</u> 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 <u>some</u> $m \geq n$ with $|X_m - Y| \geq \epsilon$. That is, $A_n = \{\exists\ m \geq n \text{ with } |X_m - Y| \geq \epsilon\}$.

  → Or, as functions: $A_n = \{s \in S : \exists\ m \geq n \text{ with } |X_m(s) - Y(s)| \geq \epsilon\}$.

  → If $s \in \bigcap_{n=1}^{\infty} A_n$, this means we can always find <u>some</u> $m \geq n$ with $|X_m(s) - Y(s)| \geq \epsilon$, i.e. the sequence $\{X_n(s)\}$ does <u>not</u> converge as a sequence to $Y(s)$.

  → That is, $\bigcap_{n=1}^{\infty} A_n \subseteq \{s : X_n(s) \not\to Y(s)\}$. So, using monotonicity:

  → This shows: $\mathrm{P}\big(\{X_n\}$ does <u>not</u> converge as a sequence to $Y\big) \geq \mathrm{P}\big(\bigcap_{n=1}^{\infty} A_n\big)$.

  → We're assuming $X_n \xrightarrow{a.s.} Y$, so $\mathrm{P}\big(\{X_n\}$ <u>does</u> converge as a sequence to $Y\big) = 1$, so $\mathrm{P}\big(\{X_n\}$ does <u>not</u> converge as a sequence to $Y\big) = 0$. Hence, $\mathrm{P}\big(\bigcap_{n=1}^{\infty} A_n\big) = 0$.

  → So what? Well, here $A_n = \bigcup_{m=n}^{\infty} B_m$ where $B_m = \{|X_m - Y| \geq \epsilon\}$.

  → Now, $\bigcup_{m=n+1}^{\infty} B_m \subseteq \bigcup_{m=n}^{\infty} B_m$. Hence, $A_{n+1} \subseteq A_n$, i.e. the $\{A_n\}$ are <u>decreasing</u>.

  → So, by Continuity of Probabilities, $\lim_{n\to\infty} \mathrm{P}(A_n) = \mathrm{P}(\bigcap_{n=1}^{\infty} A_n) = 0$.

  → But $\{|X_n - Y| \geq \epsilon\} = B_n \subseteq \bigcup_{m=n}^{\infty} B_m = A_n$.

  → Hence, $\mathrm{P}(|X_n - Y| \geq \epsilon) = \mathrm{P}(B_n) \leq \mathrm{P}(A_n)$, 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 $X_n \xrightarrow{P} Y$, just need $\mathrm{P}(|X_n - Y| \geq \epsilon) \to 0$.

  → But for $X_n \xrightarrow{a.s.} Y$, need $\mathrm{P}(\exists\ m \geq n \text{ with } |X_m - Y| \geq \epsilon) \to 0$. (Stronger.)

- Confusing? Think about it this way:

  → X_n converges to Y <u>in probability</u> if for each $\epsilon > 0$, the sequence $a_n := \mathrm{P}(|X_n - Y| \geq \epsilon)$ converges to 0.

  → But, X_n converges to Y <u>almost surely</u> if for each $\epsilon > 0$, the sequence $b_n := \mathrm{P}(\exists\, m \geq n : |X_m - Y| \geq \epsilon)$ converges to 0.

  → Subtle! However, clearly $b_n \geq a_n$, so if $b_n \to 0$, then of course $a_n \to 0$.

  → That is, if $X_n \xrightarrow{a.s.} Y$, then $X_n \xrightarrow{P} Y$.

  → But the converse is not true! (e.g. the above "5 & 7" example)

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.

END MONDAY #10

POLL: Let $U \sim \text{Uniform}[0,1]$, and $X_n = I_{U \le (1/2)+(1/2^n)}$, and $Y = I_{U \le 1/2}$. Know $X_n \to Y$ in probability. Does $X_n \to Y$ a.s.? **(A)** Yes. **(B)** No. **(C)** Not sure.

→ Well, if $U \le 1/2$, then $Y = 1$, and also $X_n = 1$ for all n.

→ So, $(X_n) = 1, 1, 1, 1, \ldots$, so $X_n \to 1 = Y$ in this case.

→ Or, if $1/2 < U \le 1$, then $Y = 0$. What about X_n?

→ Well, in this case, $X_n = 1$ for small enough n that $1/2^n \ge U - (1/2)$.

→ But for large enough n that $1/2^n < U - (1/2)$, we have $X_n = 0$.

→ So, $(X_n) = 1, \ldots, 1, 0, 0, 0, \ldots$, so $X_n \to 0 = Y$ in this case.

→ So, in any case, $X_n \to Y$. So, $\text{P}(X_n \to Y) = 1$.

→ So, yes, $X_n \stackrel{a.s.}{\to} Y$ in this case. (A)

POLL: Suppose we change it slightly, so still $U \sim \text{Uniform}[0,1]$ and $X_n = I_{U \le (1/2)+(1/2^n)}$, but now $Y = I_{U < 1/2}$. Does $X_n \to Y$ a.s.? **(A)** Yes. **(B)** No. **(C)** Not sure.

→ Well, if $U < 1/2$, or $1/2 < U \le 1$, then $X_n \to Y$ just like before.

→ What if $U = 1/2$? Then $X_n = 1$ for all n, but $Y = 0$, so $X_n \not\to Y$.

→ Hence, the event $\{X_n \to Y\} = \{U \ne 1/2\}$.

→ But $\text{P}(U \ne 1/2) = 1$. So, $\text{P}(X_n \to Y) = 1$. So, $X_n \stackrel{a.s.}{\to} Y$. Yes! (A)

Strong Law of Large Numbers (SLLN) (§4.3.1)

- Theorem: For any sequence of random variables $X_1, X_2, X_3, \ldots$ which are i.i.d., each with finite 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., $M_n \stackrel{a.s.}{\to} \mu$).

  → Proof in more advanced books, e.g. http://probability.ca/grprob

  → Then, of course, also $M_n \stackrel{P}{\to} \mu$, 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 $\text{E}(X_n) = 1/2 =: \mu$.

  → 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 $\epsilon > 0$, $\text{P}(|M_n - (1/2)| \ge \epsilon) \to 0$.

  → So, for all sufficiently large n, i.e. $\text{P}(M_n < 0.503) > 0.99$, etc.

  → But the SLLN says more: $\text{P}(M_n \to 1/2) = 1$.

  → So, for all $\epsilon > 0$, $\text{P}(|M_n - 0.5| \le \epsilon$ for all sufficiently large $n) = 1$.

  → So e.g. $\text{P}(M_n < 0.503$ for all sufficiently large $n) = 1$.

  → In particular, $\text{P}(\exists\, n : M_n < 0.503) = 1$.

  → That is, $\text{P}(\exists\, n : X_1 + X_2 + \ldots + X_n < (0.503)n) = 1$. etc.

- Try it out in R! File http://probability.ca/Rslln (first choose theta).

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.

Central Limit Theorem (CLT) (§4.4.1)

• 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).

• How does the frequency distribution of $\frac{1}{n} S_n - \mu$ look?

→ Usually centered near 0 (makes sense).

→ Width is fairly small (how small?). And ...

POLL: The shape of the frequency distribution of $\frac{1}{n} S_n - \mu$ is approximately: **(A)** Uniform. **(B)** Binomial. **(C)** Poisson. **(D)** Exponential. **(E)** Normal. **(F)** Geometric.

→ The shape appears to be approximately ... normal! (E)

• 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!

• Now, let $Z_n = (\frac{1}{n} S_n - \mu) / \sqrt{\sigma^2/n} = \frac{S_n - n\mu}{\sqrt{n}\,\sigma}$. Then what are $\mathrm{E}(Z_n)$ and $\mathrm{Var}(Z_n)$?

→ Well, here $\mathrm{E}(S_n) = n\mu$, and $\mathrm{Var}(S_n) = n\sigma^2$.

→ So, $\mathrm{E}(Z_n) = \mathrm{E}\left(\frac{S_n - n\mu}{\sqrt{n}\,\sigma}\right) = \frac{\mathrm{E}(S_n) - n\mu}{\sqrt{n}\,\sigma} = \frac{n\mu - n\mu}{\sqrt{n}\,\sigma} = 0$.

→ And, $\mathrm{Var}(Z_n) = \mathrm{Var}\left(\frac{S_n - n\mu}{\sqrt{n}\,\sigma}\right) = \frac{\mathrm{Var}(S_n)}{(\sqrt{n}\,\sigma)^2} = \frac{n\sigma^2}{n\sigma^2} = 1$.

→ That is, $\mathrm{E}(Z_n) = 0$, and $\mathrm{Var}(Z_n) = 1$. ("standardised")

→ 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 (§4.4)]

→ Equivalently, $\lim_{n\to\infty} \mathrm{P}(\frac{S_n - n\mu}{\sqrt{n}\,\sigma} \le z) = \Phi(z)$.

→ 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. $z = 0$: $\lim = 1/2$)

→ 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 (text pp. 247–8): Use "moment-generating functions". (§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 $m_X(s) < \infty$ for all s (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. $m_X(0) = \mathrm{E}[e^{0\,X}] = \mathrm{E}[e^0] = \mathrm{E}[1] = 1$, and $m'_X(s) = \frac{d}{ds} m_X(s) = \frac{d}{ds}\mathrm{E}[e^{sX}] = \mathrm{E}[\frac{\partial}{\partial s} e^{sX}] = \mathrm{E}[X e^{sX}]$, so $m'_X(0) = \mathrm{E}[X]$. Similarly $m''_X(0) = \mathrm{E}[X^2]$. [In general for any $k \in \mathbf{N}$ we have $m_X^{(k)}(0) = \mathrm{E}[X^k]$: "moments"]

→ Also, if X and Y are independent, then $m_{X+Y}(s) = \mathrm{E}[e^{s(X+Y)}] = \mathrm{E}[e^{sX}\, e^{sY}] = \mathrm{E}[e^{sX}]\,\mathrm{E}[e^{sY}] = m_X(s)\, m_Y(s)$.

→ KEY FACT: For any r.v. $Y_1, Y_2, \ldots$, if $\lim_{n\to\infty} m_{Y_n}(s) = m_Y(s)$ for all s (at least in a neighbourhood of $s = 0$), then for all $y \in \mathbf{R}$, $\lim_{n\to\infty} \mathrm{P}(Y_n \le y) = \mathrm{P}(Y \le y)$, i.e. $\lim_{n\to\infty} F_{Y_n}(y) = F_Y(y)$, i.e. Y_n converges to Y in distribution.

- So, how can we prove the Central Limit Theorem (CLT)?

  → Need to 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 $m_Z(s) = \mathrm{E}[e^{sZ}] = \int_{-\infty}^{\infty} e^{sz}\, \frac{1}{\sqrt{2\pi}}\, e^{-z^2/2}\, dz = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}}\, e^{sz-(z^2/2)}\, dz = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}}\, e^{-(z-s)^2/2+(s^2/2)}\, dz = e^{s^2/2} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}}\, e^{-(z-s)^2/2}\, dz$.

  → But $\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}}\, e^{-(z-s)^2/2}\, dz = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}}\, e^{-w^2/2}\, dw = 1$. (normal density)

  → Hence, $m_Z(s) = e^{s^2/2}\,(1) = e^{s^2/2}$.

- 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 small 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 $n \to \infty$, $[1 + \frac{a}{n}]^n \to e^a$. (Here $a = s^2/2$.)

  → 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 (phew!). ∎

END WEDNESDAY #11

- Summary: If $X_1, X_2, \ldots$ are any i.i.d. random variables with finite mean μ and variance σ^2, and their partial sums or averages are "standardised" to have mean 0 and variance 1, then: Their probabilities always converge to Normal(0,1). Wow!

  → Other versions when $\{X_n\}$ not i.i.d., e.g. "Lindeberg CLT"; error bounds, etc.

  → This is why the Normal distribution is so important and common, e.g.
https://www.statology.org/example-of-normal-distribution/
https://www.mathsisfun.com/data/quincunx.html

Normal Approximations (§4.4.1)

- Okay, so we know that as $n \to \infty$, $P(\frac{S_n - n\mu}{\sqrt{n}\,\sigma} \le z) \to \Phi(z)$.

- Hence, for "reasonably large" n, we must have $P(\frac{S_n - n\mu}{\sqrt{n}\,\sigma} \le z) \approx \Phi(z)$.

  → How large? It depends on the distribution of the X_i, and error size.

  → Rough "rule of thumb": Pretty good approximation if $n \ge 30 \ldots$

  → Can use this to get approximate values for many probabilities!

- Example: Suppose $\{X_n\}$ are i.i.d. $\sim$ Exponential(4).

  → What is a good approximation to $P(X_1 + X_2 + \ldots + X_{100} \ge 30)$?

  → Here $\mu := E(X_i) = 1/\lambda = 1/4$, and $\sigma := Sd(X_i) = \sqrt{1/\lambda^2} = 1/\lambda = 1/4$.

  → So, if $S_{100} = X_1 + X_2 + \ldots + X_{100}$, then: $P(S_{100} \ge 30)$

$$= P\left(\frac{S_{100} - 100(1/4)}{\sqrt{100}\,(1/4)} \ge \frac{30 - 100(1/4)}{\sqrt{100}\,(1/4)}\right) = P\left(\frac{S_{100} - 100(1/4)}{\sqrt{100}\,(1/4)} \ge 2\right)$$

$$= P(Z_{100} \ge 2) \approx P(Z \ge 2) = P(Z \le -2) = \Phi(-2) \doteq 0.0228\,.$$

  → Here the value of $\Phi(-2)$ can found from software [e.g. "pnorm(-2)" in R], or from a table like textbook 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 $P(X_1 + X_2 + \ldots + X_{400} \le 1420)$?

  → Here $\mu := E(X_i) = (2+5)/2 = 3.5$, and $\sigma := Sd(X_i) = \sqrt{Var(X_i)} = \sqrt{(5-2)^2/12} \doteq 0.866$.

  → So, if $S_{400} = X_1 + X_2 + \ldots + X_{400}$, then: $P(S_{400} \le 1420)$

$$= P\left(\frac{S_{400} - 400(3.5)}{\sqrt{400}\,(0.866)} \le \frac{1420 - 400(3.5)}{\sqrt{400}\,(0.866)}\right) \doteq P\left(\frac{S_{400} - 400(3.5)}{\sqrt{400}\,(0.866)} \le 1.15\right)$$

$$\approx P(Z \le 1.15) = \Phi(1.15) = 1 - \Phi(-1.15) \doteq 1 - 0.1251 = 0.8749\,.$$

POLL: Suppose $\{X_n\}$ are i.i.d. $\sim$ Poisson(4). What is the normal approximation to $P(X_1 + X_2 + \ldots + X_{900} \ge 3700)$?
(A) $\Phi(1)$. **(B)** $\Phi(5/3)$. **(C)** $\Phi(-5/3)$. **(D)** $\Phi(5/4)$. **(E)** $\Phi(-5/4)$.

  → Here $\mu := E(X_i) = \lambda = 4$, and $\sigma := Sd(X_i) = \sqrt{Var(X_i)} = \sqrt{\lambda} = 2$.

→ Let $S_{900} = X_1 + X_2 + \ldots + X_{900}$.

→ Then $\mathrm{P}(X_1 + X_2 + \ldots + X_{900} \geq 3700) = \mathrm{P}(S_{900} \geq 3700)$

$$= \mathrm{P}\left(\frac{S_{900} - 900(4)}{\sqrt{900}\,(2)} \geq \frac{3700 - 900(4)}{\sqrt{900}\,(2)}\right) = \mathrm{P}\left(\frac{S_{900} - 900(4)}{\sqrt{900}\,(2)} \geq 5/3\right)$$

$$= \mathrm{P}\left(Z_{900} \geq 5/3\right) \approx \mathrm{P}\left(Z \geq 5/3\right) = \mathrm{P}\left(Z \leq -5/3\right) = \Phi(-5/3) \doteq 0.0478\,.$$

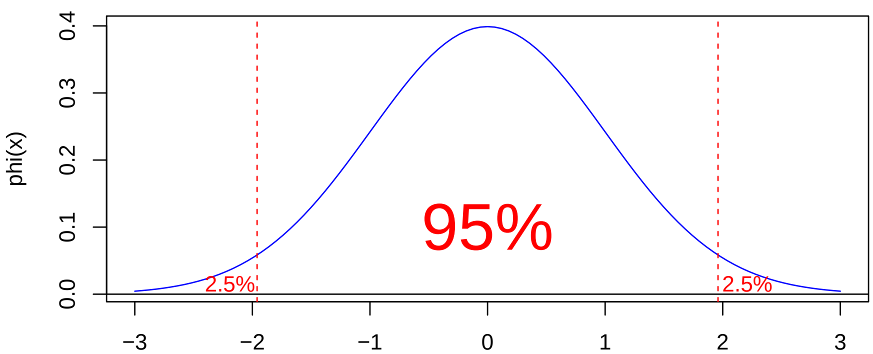
Suggested Homework: 4.4.5, 4.4.6, 4.4.7, 4.4.12, 4.4.13, [4.4.22, 4.4.23 given var].

Estimation and Confidence Intervals (§4.4.2)

- Fact: $\Phi(-1.96) \doteq 0.025$.

POLL: If $Z \sim \text{Normal}(0,1)$, then $\mathrm{P}(-1.96 \leq Z \leq +1.96)$ is approximately:
(A) 0.025. **(B)** 0.05. **(C)** 0.5. **(D)** 0.95. **(E)** 0.975.

→ Here $\mathrm{P}(Z \leq -1.96) = \Phi(-1.96) \doteq 0.025$, and $\mathrm{P}(Z \geq +1.96) = 1 - \Phi(+1.96) = \Phi(-1.96) \doteq 0.025$, 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 $\mathrm{P}(34.45 \leq X_1 + X_2 + \ldots + X_{200} \leq 45.54) \approx 0.95$.

  → 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: $\mathrm{P}(\frac{1}{n}S_n - 1.96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \frac{1}{n}S_n + 1.96\frac{\sigma}{\sqrt{n}}) \approx 0.95$.

  → 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 $\mathrm{P}(M_n - 1.96\frac{\sigma}{\sqrt{n}} \leq \mu \leq M_n + 1.96\frac{\sigma}{\sqrt{n}}) \approx 0.95$.

→ 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 $[0.007,\ 0.109]$. (Subtle.)

→ 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.

- Aside: The value 95% is "usual", but other values are also possible. (e.g. 99%)

→ 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 $\sigma_n := \sqrt{M_n(1-M_n)}$.

→ So, use the interval $[M_n - 1.96\sqrt{M_n(1-M_n)/n},\ M_n + 1.96\sqrt{M_n(1-M_n)/n}]$.

- 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.

POLL: Suppose you're shooting free throws, and score 86 out of 250 of them. Then a 95% confidence interval for the (unknown) true success rate θ is:

(A) $[(86) - 1.96\sqrt{(86)(1-(86))/250},\ (86) + 1.96\sqrt{(86)(1-(86))/250}]$.

(B) $[(86) - 1.96\sqrt{\frac{86}{250}(1-\frac{86}{250})/250},\ (86) + 1.96\sqrt{\frac{86}{250}(1-\frac{86}{250})/250}]$.

(C) $[\frac{86}{250} - 1.96\sqrt{(86)(1-(86))/250},\ \frac{86}{250} + 1.96\sqrt{(86)(1-(86))/250}]$.

(D) $[\frac{86}{250} - 1.96\sqrt{\frac{86}{250}(1-\frac{86}{250})},\ \frac{86}{250} + 1.96\sqrt{\frac{86}{250}(1-\frac{86}{250})}]$.

(E) $[\frac{86}{250} - 1.96\sqrt{\frac{86}{250}(1-\frac{86}{250})/250},\ \frac{86}{250} + 1.96\sqrt{\frac{86}{250}(1-\frac{86}{250})/250}]$.

(F) $[\frac{86}{250} - 3\sqrt{\frac{86}{250}(1-\frac{86}{250})/250},\ \frac{86}{250} + 3\sqrt{\frac{86}{250}(1-\frac{86}{250})/250}]$.

→ Here the number of scores is $S_{250} \sim \text{Binomial}(250, \theta)$, with θ unknown.

→ That is, $S_{250} = X_1 + X_2 + \ldots + X_{250}$, where the $\{X_i\}$ are i.i.d. $\sim$ Bernoulli(θ).

→ So, here $n = 250$, and $\mu = \theta$ (unknown).

→ 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, $\mathrm{P}(M_n - 1.96\,\sigma/\sqrt{n} \le \theta \le M_n + 1.96\,\sigma/\sqrt{n}) \approx 0.95$.

→ Here we observed that $n = 250$ and $M_{250} = \frac{86}{250} \doteq 0.344$, so:

→ 95% confidence interval: $[\frac{86}{250} - 1.96\,\sigma/\sqrt{250},\ \frac{86}{250} + 1.96\,\sigma/\sqrt{250}]$.

→ Here $\sigma = \sqrt{\theta(1-\theta)}$, which is unknown.

→ Estimate: $\theta \approx \frac{86}{250}$, so $\sigma^2 = \theta(1-\theta) \approx \frac{86}{250}(1-\frac{86}{250})$. This gives:

→ 95% C.I. $= [\frac{86}{250} - 1.96\sqrt{\frac{86}{250}(1-\frac{86}{250})/250},\ \frac{86}{250} + 1.96\sqrt{\frac{86}{250}(1-\frac{86}{250})/250}]$. (E)

→ This equals $[0.285,\ 0.403]$, which gives a 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. $\sim$ 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 are μ and σ in terms of λ?]

Q4. Suppose $\{X_n\}$ are i.i.d. $\sim$ Poisson(λ), where λ is unknown. Compute a 95% confidence interval for λ. [Hint: What are μ and σ in terms of 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 Algorithms (§4.5)

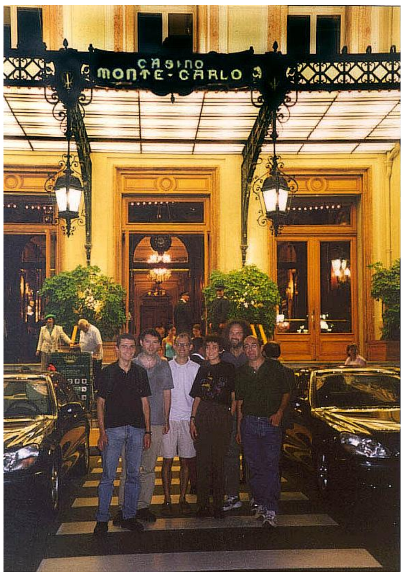
- e.g. Suppose $U \sim \text{Uniform}[0,1]$. What is $\mu := \mathrm{E}\Big(U^3\big[\sin(U^4) + \cos(U^5)\big]e^{-U^6}\Big)$?

→ 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 <u>computing by using randomness</u>.

→ e.g. To compute $\mu := \mathrm{E}\big(U^3[\sin(U^4)+\cos(U^5)]e^{-U^6}\big)$, first <u>generate</u> i.i.d. random values $U_1, U_2, \ldots, U_n \sim \text{Uniform}[0,1]$ 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 $\sigma^2 := \text{Var}(X_n) \le \text{E}[(X_n)^2] \le 4$, 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:

$$= [\tfrac{1}{50000}(11319.6) - 1.96\tfrac{2}{\sqrt{50000}},\ \tfrac{1}{50000}(11319.6) + 1.96\tfrac{2}{\sqrt{50000}}] \doteq [0.209, 0.244].$$

  - → So, 95% confident that $\mu := \text{E}[U^3(\sin(U^4) + \cos(U^5))e^{-U^6}] \in [0.209, 0.244]$.
  - → 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 $I = \text{E}[e^{\cos(U)}]$ where $U \sim \text{Uniform}[0,1]$.
  - → So, as before, first generate random i.i.d. values $U_1, U_2, \ldots, U_n \sim \text{Uniform}[0,1]$.
  - → Then set $X_i = e^{\cos(U_i)}$, so $\mu := \text{E}[X_i] = I$. And $\sigma \le \sqrt{\text{E}[(X_i)^2]} \le \sqrt{e^2} = e$.
  - → 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 95% confidence interval for I.

- Many other integrals can also be converted to expected values for Monte Carlo:
  - → e.g. $\int_5^8 \cos(x^7)\,dx = \int_5^8 [3\cos(x^7)]\,\frac{1}{3}\,dx = \text{E}[3\cos(X^7)]$ where $X \sim \text{Uniform}[5,8]$.

- And what about over infinite regions?

POLL: $\int_0^\infty \cos(x^7)\,e^{-5x}\,dx$ is equal to the expected value:
(A) $\text{E}[\cos(Y^7)]$ where $Y \sim \text{Uniform}[0,5]$. **(B)** $\text{E}[\cos(Y^7)]$ where $Y \sim \text{Exponential}(5)$.
(C) $\text{E}[5\cos(Y^7)]$ where $Y \sim \text{Exponential}(5)$. **(D)** $\text{E}[\frac{1}{5}\cos(Y^7)]$ where $Y \sim \text{Exponential}(1/5)$.
(E) $\text{E}[\frac{1}{5}\cos(Y^7)]$ where $Y \sim \text{Exponential}(5)$. **(F)** $\text{E}[5\cos(Y^7)]$ where $Y \sim \text{Exponential}(1/5)$.

→ Here $\int_0^\infty \cos(x^7)\,e^{-5x}\,dx = \int_0^\infty [\frac{1}{5}\cos(x^7)]\,5e^{-5x}\,dx = \text{E}[\frac{1}{5}\cos(Y^7)]$ where $Y \sim \text{Exponential}(5)$. (E)

POLL: $\int_{-\infty}^\infty \cos(x^7)\,e^{-x^2/2}\,dx$ is equal to the expected value:
(A) $\text{E}[\cos(Z^7)]$ where $Z \sim \text{Exponential}(2)$. **(B)** $\text{E}[\cos(Z^7)]$ where $Z \sim \text{Normal}(0,1)$.
(C) $\text{E}[\cos(Z^7)]$ where $Z \sim \text{Normal}(0,2)$. **(D)** $\text{E}[\frac{1}{\sqrt{2\pi}}\cos(Z^7)]$ where $Z \sim \text{Normal}(0,1)$.
(E) $\text{E}[\sqrt{2\pi}\cos(Z^7)]$ where $Z \sim \text{Normal}(0,1)$. **(F)** $\text{E}[\sqrt{2\pi}\cos(Z^7)]$ where $Z \sim \text{Normal}(0,2)$.

→ Here $\int_{-\infty}^\infty \cos(x^7)\,e^{-x^2/2}\,dx = \int_{-\infty}^\infty [\sqrt{2\pi}\cos(x^7)]\,\frac{1}{\sqrt{2\pi}}e^{-x^2/2}\,dx = \text{E}[\sqrt{2\pi}\cos(Z^7)]$ where $Z \sim \text{Normal}(0,1)$. (E)

END MONDAY #11

→ And can we do the same thing for sums, too?

POLL: The sum $\sum_{j=0}^{\infty} \cos(j^7)\,(2/3)^j$ is equal to the expected value:
(A) $E[\cos(X^7)]$ where $X \sim$ Poisson(1/3). **(B)** $E[\cos(X^7)]$ where $X \sim$ Poisson(3).
(C) $E[(1/3)\cos(X^7)]$ where $X \sim$ Poisson(3). **(D)** $E[\cos(X^7)]$ where $X \sim$ Geometric(1/3).
(E) $E[3\,\cos(X^7)]$ where $X \sim$ Geometric(1/3).
(F) $E[(1/3)\,\cos(X^7)]$ where $X \sim$ Geometric(1/3).

→ Here $\sum_{j=0}^{\infty} \cos(j^7)\,(2/3)^j = \sum_{j=0}^{\infty} [3\,\cos(j^7)]\,[1-(1/3)]^j (1/3) = E[3\,\cos(X^7)]$ where $X \sim$ Geometric(1/3). (E)

→ 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 introduced in 1733!
- 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()`

POLL: What is the probability that the needle will touch a line? [Best guess!]
(A) 1/3. **(B)** 1/2. **(C)** 2/3. **(D)** 3/4. **(E)** $2/\pi$.

→ 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 $\alpha \sim$ Uniform$[0°, 180°]$, i.e. $\alpha \sim$ Uniform$[0, \pi]$ in radians.

→ So, P(needle touches line) $= E[\sin(\alpha)] = \frac{1}{\pi}\int_0^{\pi} \sin(x)\,dx = \frac{1}{\pi}[-\cos(x)]\Big|_{x=0}^{x=\pi}$
$= \frac{1}{\pi}[-\cos(\pi) + \cos(0)] = \frac{1}{\pi}[-(-1) + (1)] = 2/\pi \doteq 0.637$. (Depends on π!) (E)

- Now, 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!

- 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, e.g.:

  → $\pi/4 = \arctan(1) = 1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \ldots$ [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 (§4.6)

- 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.)

  → But we will NOT cover them here. (Not on final exam.)

END WEDNESDAY #12

Final Announcements

- No class on Monday! (I will attend in case you have any questions.)
- **Please complete the online course evaluation!**
- During the coming days: TA tutorials and office hours (and Piazza).

  → Extra Instructor Office Hours: Fri Dec 6, 12:30–1:30, in SS 1073.

- ***** **Final Exam: Saturday December 7 at 2:00 PM**, in Benson Building (BN: 320 Huron St) room 322 for surnames A–TAN, or St Volodymyr Institute (VO: 620 Spadina Ave) Auditorium B for surnames TANE–ZZ. Three hours. Arrive early!

  → Closed book; bring TCard; can use a basic (IMPORTANT!) calculator only.

***** **Good luck on the exam, and with all of your future studies!** *****