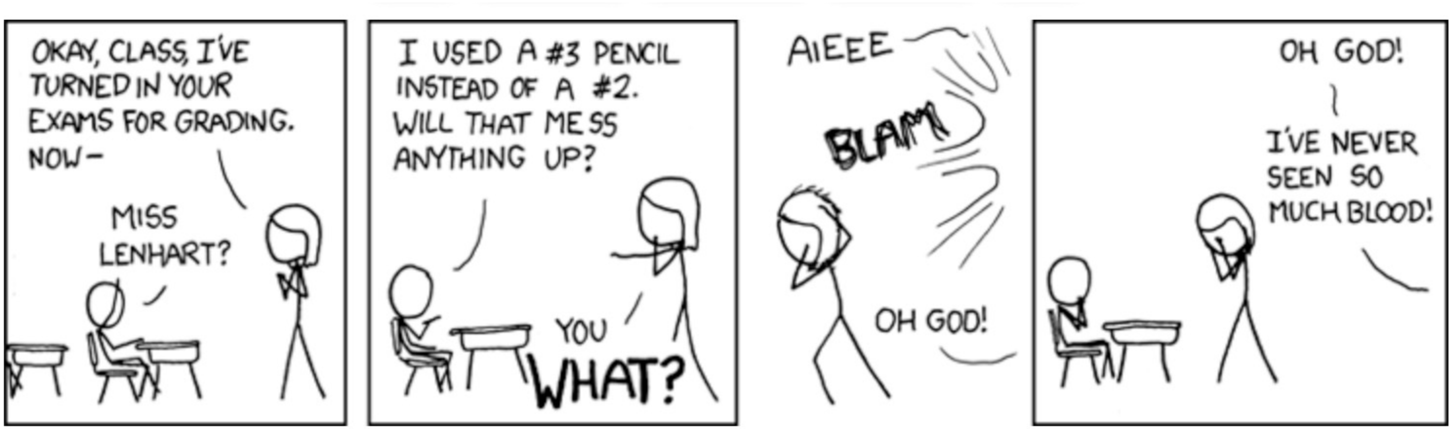
Stat 88: Probability & Math. Statistics in Data Science

SCANTRON


OKAY, CLASS, I'VE TURNED IN YOUR EXAMS FOR GRADING. NOW –
MISS LENHART?
I USED A #3 PENCIL INSTEAD OF A #2. WILL THAT MESS ANYTHING UP?
YOU WHAT?
AIEEE
BLAM
OH GOD!
OH GOD!
I'VE NEVER SEEN SO MUCH BLOOD!

https://xkcd.com/499/

Lecture 15: 3/8/2022

Talking about the midterm

Material on midterm: Chapters 1 – 5 (Lectures 1 – 14)

$P(A \cap B) = P(AB)$

$\cap \longleftrightarrow$ and/,

$\cup \longleftrightarrow$ or

- Probabilities definitions, bounds, fundamental rules
- Intersections (multiplication rules), conditional probability, independence

  $P(A \cap B) = P(A|B)P(B)$
- Symmetry in simple random sampling (drawing tickets from boxes, cards from decks)
- Addition rule (inclusion-exclusion)

  $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- Bayes' rule, partitions

  $P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(AB) - P(BC) - P(AC) + P(ABC)$
- Success, failure, random variables, binomial, hypergeometric, waiting times (geometric, negative binomial), poisson
- Pmf, cdf, how to go from one to the other

  $F(x) = P(X \leq x)$, $f(x) = P(X = x)$
- Exponential approximations, relation between binomial and poisson
- Expectation of rvs, rules of expectation, expectation of functions of random variables
- Joint distributions, conditional distributions
- Methods of indicators
- Conditional expectation and expectation by conditioning.

Make sure to put your name & your GSI 's name on the cheat sheet.

$$X = \begin{cases} 2 & \text{w.p } 1/5 \\ 1 & \text{w p } 3/10 \\ -1 & \text{w.p } 1/10 \\ 0 & \text{w.p } 2/5 \end{cases}$$

$$\mathbb{E}(g(X)) = \sum_x g(x) \cdot P(X=x)$$

$$\mathbb{E}(X^2) = \sum_x x^2 \cdot P(X=x)$$

$$= (2^2) \cdot \frac{1}{5} + 1^2 \cdot \frac{3}{10} + (-1)^2 \cdot \frac{1}{10} + 0^2 \cdot \frac{2}{5}$$

$$= \frac{4}{5} + \frac{3}{10} + \frac{1}{10} = \frac{6}{5}$$

Midterm review

Names and Parameters	Values	$P(X=k)$	$E(X)$
Bernoulli(p)/Indicator	$0, 1$	$P(X=1) = p$	p
Uniform on $\{1, 2, \ldots, N\}$	$1, 2, 3, \ldots, N$	$\frac{1}{N}$	$\frac{N+1}{2}$
Binomial(n, p)	$0, 1, 2, \ldots, n$	$\binom{n}{k} p^k (1-p)^{n-k}$	np
Hypergeometric (N, G, n)	$0, 1, 2, \ldots, n$	$\frac{\binom{G}{k}\binom{N-G}{n-k}}{\binom{N}{n}}$	$n\frac{G}{N}$
Poisson(μ)	$0, 1, 2, \ldots$	$e^{-\mu}\frac{\mu^k}{k!}$	μ
Geometric(p)	$1, 2, 3, \ldots$	$(1-p)^{k-1}p$	$\frac{1}{p}$

Waiting time when $r > 1$ (negative binomial)

Note that if X has the Poisson(μ) distribution, and Y has the Poisson(λ) distribution, then $X + Y$ has the Poisson($\mu + \lambda$) distribution.

Note that:

- Unless otherwise stated, dice have six sides and are fair, and coins have two distinct coins and are fair, cards are dealt without replacement.
- “at random” means equally likely (uniformly at random)

- i.i.d: independent and identically distributed
- pmf: probability mass function
- Cdf: cumulative distribution function

- When you study, verbalize. Try to describe the problem in words. Make sure you understand what the problem is asking for, and you understand what information you are given. For instance, the difference between waiting times and binomial random variables
- Draw pictures!! Draw pictures!!
- Do NOT assume independence!! Either the problem has to provide an assumption of independence, or the assumption has to follow from the conditions of the experiment, e.g. sampling with replacement, or events based on separate sets of tosses, etc.

During the exam

- Start with the easy questions. When you open the midterm, skim through it quickly and mark the questions that are easy to answer and do those first. Not only will this help you maximize your scoring potential but it will also bolster your confidence. Once you are done with answering the easy questions, it's time to tackle the ones you skipped. Don't over-think straightforward questions.

- Read each question carefully. As you know, assumptions matter. If you have misread those then your solution will be off. For example, confusing with and without replacement is a big problem. Also: "the fourth head is on the 20th toss" and "four heads in 20 tosses" are quite different.

- Forcing yourself to read slowly, underlining key assumptions as you read, is important for doing well. If you are done with the test, and you have time, check your work by reading each question afresh and thinking about how to solve it again instead of just reading over your answer.

During the exam

- Provide reasoning or a calculation in **all** questions. If you did a calculation in your head, **write out** the calculation you did in your head.

- Don't waste time simplifying any arithmetic or algebra unless you are explicitly asked to. We need to see your thought process, and are confident that you are able to use a calculator – we do not need a demonstration of this knowledge.

- If an answer is taking you numerous lines of calculation, or complicated algebra/calculus, you've probably missed something. Rethink, or move to a different problem.

- If you don't know how to do a problem, try not to leave it blank. Almost always, you will have an idea of what might be relevant. If you write that, and it is indeed important for the problem, you might get some partial credit. That said, you shouldn't expect partial credit for everything you write. We'll be looking for substantive ideas and progress towards a solution.

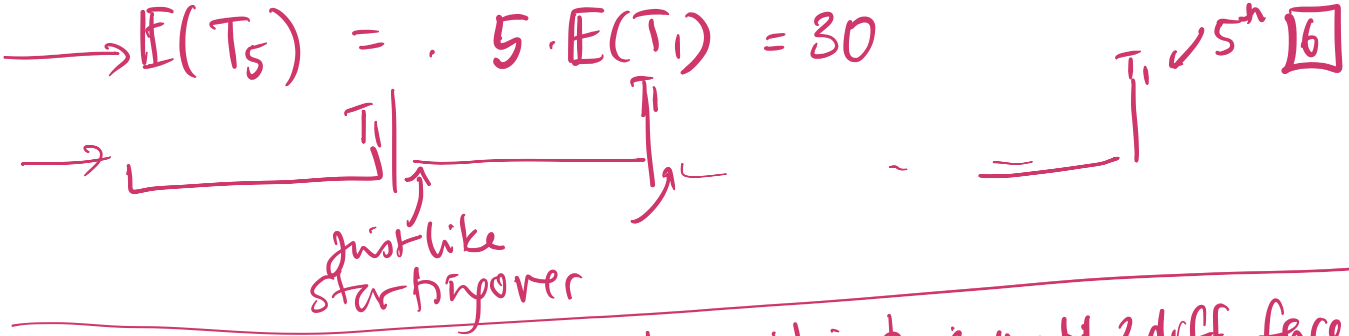
Exercise 5.7.14 from the text

1. A fair die is rolled repeatedly.
   a. Find the expected waiting time (number of rolls) till a total of 5 sixes appear
   b. A fair die is rolled repeatedly. Find the expected waiting time (number of rolls) until two *different* faces are rolled. → X

1a) T_1 : waiting time until 1 $\boxed{6}$ appears

$P(\boxed{6}) = \frac{1}{6}$ $\qquad E(T_1) = \frac{1}{p} = 6$

$\longrightarrow E(T_5) = 5 \cdot E(T_1) = 30$



(b) Let the expected waiting time until 2 diff faces be X

$X = X_1 + X_2$, $X_1 = 1$, X_2 = waiting time after first roll until a diff face is rolled

$\underbrace{\boxed{3}}\ \boxed{3}\ \boxed{2}$ $\quad X_1 = 1$ $\quad X = 1+2 = 3$

$X_2 = 2$

$X_2 \sim \text{Geom}\left(\frac{5}{6}\right)$

$$\mathbb{E}(X) = 1 + \mathbb{E}(X_2) = 1 + \frac{6}{5} = \frac{11}{5}$$

All who wander might be lost…

A lost tourist arrives at a point with 2 roads. Road A brings him back to the same point after 1 hour of walking. Road B leads to the city in 2 hours. Assuming the tourist randomly chooses a road at all times, what is the expected time until the tourist arrives to the city?

X = time until tourist arrives at city

S = road taken

$S = A$ w.p $\frac{1}{2}$ or B w.p $\frac{1}{2}$

Let $\mathbb{E}(X) = t$

$$\mathbb{E}(X) = \underset{S}{\underset{\uparrow}{\mathbb{E}}}\Big(\underset{X}{\underset{\uparrow}{\mathbb{E}}}(X|S)\Big)$$

t

$$\Rightarrow \quad t = \frac{1}{2}\underbrace{\mathbb{E}(X|S=A)} + \frac{1}{2}\mathbb{E}(X|S=B)$$

$$\frac{1}{2}(t+1) + \frac{1}{2}\cdot 2$$

$$t = \frac{t}{2} + \frac{1}{2} + 1 \Rightarrow t = 3 \text{ hrs}$$

Choosing cards

- Tamara chooses an integer *N* uniformly at random from *Pois(μ)*. She then picks *N* cards from a deck with replacement. Find the expected number of ace cards.

Let X = # of aces in N cards.

Suppose N=10, then X = # of aces in 10 cards w.p $\frac{4}{52}$ each draw

$$X \sim Bin\left(10, \tfrac{4}{52}\right) \qquad \mathbb{E}(X) = (10)\left(\tfrac{4}{52}\right)$$

Sp. N = n, $X \sim Bin\left(n, \frac{4}{52}\right)$, $\mathbb{E}(X) = \underline{n \cdot \frac{4}{52}}$

$$\mathbb{E}(N) = \mu$$

$$\mathbb{E}(X) = \mathbb{E}\Big(\overset{\substack{\text{function of } N \\ \downarrow}}{\mathbb{E}(X|N)}\Big) = \sum_{n=0}^{\infty} \underbrace{\mathbb{E}(X|N=n)} P(N=n)$$

$$= \sum_{n=0}^{\infty} n \cdot \frac{4}{52} \cdot P(N=n)$$

$$= \frac{4}{52} \underbrace{\sum_{n=0}^{\infty} n \cdot P(N=n)}_{=\mathbb{E}(N)=\mu}$$

$$\mathbb{E}(g(N)) = \sum_{n=0}^{\infty} g(n) P(N=n)$$

$= \boxed{\mu \cdot \frac{4}{52}}$

Problems from past midterms

Bella and her friend Kobie both live in households with four humans each. Each of the humans in Bella's household (and Kobie's) might take her for a walk that day. Let B be the number of walks that Bella gets on a randomly selected day, and K be the number of walks that Kobie goes on. Assume that B and K are iid random variables that have the uniform distribution on {1, 2, 3, 4}. Let M be the maximum of B and K.

(a) Write down the joint distribution of B and M.

(b) Are B and M independent?

(c) What is the expectation of M?

(a) Since B & K can each go from 1 to 4 there are 4·4 = 16 possible ordered pairs (b,k)

M	(b,k)
1	(1,1)
2	(1,2), (2,1), (2,2)
3	(1,3), (3,1), (2,3), (3,2), (3,3)
4	(1,4), (4,1), (2,4), (4,2), (3,4), (4,3), (4,4)

(eq. prob of 1/4)

$M \backslash B$	1	2	3	4	$f_M(m)$ ↓
1	$\frac{1}{16}$	0	0	0	$f_M(1) = \frac{1}{16}$
2	$\frac{1}{16}$	$\frac{2}{16}$	0	0	$f_M(2) = \frac{3}{16}$
3	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{3}{16}$	0	$f_M(3) = \frac{5}{16}$
4	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{4}{16}$	$f_M(4) = \frac{7}{16}$
$f_B(b)$ →	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	

$$f(4,2) = 0 \neq f_B(4) f_M(2)$$

$$\mathbb{E}(M) = \sum_{m=1}^{4} m \cdot f_M(m) = 1 \cdot \frac{1}{16} + 2 \cdot \frac{3}{16} + 3 \cdot \frac{5}{16} + 4 \cdot \frac{7}{16}$$

A fair four-sided die is rolled, and then a fair coin is tossed as many times as there are spots that show on the die. Let X be the number of spots rolled, and let Y be the total number of heads we get once we are done tossing the coin.

(a) Write down the joint distribution table for X and Y

(b) Write down the marginal distribution of Y and find its expectation.

(c) Given $Y = 1$, find the conditional distribution of X

$Y \backslash X$	1	2	3	4	
0	$P(X=1, Y=0)$ $\frac{1}{2}\cdot\frac{1}{4}$	$\frac{1}{4}\cdot\frac{1}{4}$	$\frac{1}{8}\cdot\frac{1}{4}$	$\frac{1}{16}\cdot\frac{1}{4}$	$\frac{1}{8}+\frac{1}{16}+\frac{1}{32}+\frac{1}{64}=\frac{8+4+2+1}{64}=\frac{15}{64}$
1	$\frac{1}{2}\cdot\frac{1}{4}$	$\frac{1}{2}\cdot\frac{1}{4}$	$\frac{3}{8}\cdot\frac{1}{4}$	$\frac{4}{16}\cdot\frac{1}{4}$	$\frac{1}{8}+\frac{1}{8}+\frac{3}{32}+\frac{1}{16}=\frac{4+4+3+2}{32}=\frac{13}{32}$
2	0	$\frac{1}{4}\cdot\frac{1}{4}$	$\frac{3}{8}\cdot\frac{1}{4}$	$\frac{3}{8}$ ~~$\frac{6}{16}$~~ $\cdot\frac{1}{4}$	$\frac{1}{16}+\frac{3}{32}+\frac{3}{32}=\frac{2+3+3}{32}=\frac{8}{32}$
3	0	0	$\frac{1}{8}\cdot\frac{1}{4}$	$\frac{4}{16}\cdot\frac{1}{4}$	$\frac{1}{32}+\frac{1}{16}=\frac{3}{32}$
4	0	0	0	$\frac{1}{16}\cdot\frac{1}{4}$	$\frac{1}{64}$
$f_X(x)$	1/4	1/4	1/4	1/4	1

$$P(X=1, Y=0) = P(Y=0 \mid X=1)\, P(X=1)$$

$$\frac{1}{2} \cdot \frac{1}{4}$$

(b)

y	0	1	2	3	4
$P(Y=y)$	$\frac{15}{64}$	$\frac{13}{32} = \frac{26}{64}$	$\frac{8}{32} = \frac{16}{64}$	$\frac{3}{32} = \frac{6}{64}$	$\frac{1}{64}$

$$\mathbb{E}(Y) = 0 \cdot \frac{15}{64} + 1 \cdot \frac{26}{64} + 2 \cdot \frac{16}{64} + 3 \cdot \frac{6}{64} + 4 \cdot \frac{1}{64}$$

$$= \frac{26 + 32 + 18 + 4}{64} = \frac{80}{64} = \frac{5}{4}$$

c)

	$X=1$	$X=2$	$X=3$	$X=4$
$Y=1$	$\frac{1}{2} \cdot \frac{1}{4}$	$\frac{1}{2} \cdot \frac{1}{4}$	$\frac{3}{8} \cdot \frac{1}{4}$	$\frac{4}{16} \cdot \frac{1}{4}$

$$P(X=1 \mid Y=1) = \frac{P(X=1, Y=1)}{P(Y=1)} = \frac{1/8}{13/32} = \frac{4}{13}$$

$$P(X=2 \mid Y=1) = \frac{1/8}{13/32} = \frac{4}{13}, \quad P(X=3 \mid Y=1) = \frac{3/32}{13/32} = \frac{3}{13}$$

$$P(X=4 \mid Y=1) = \frac{1/16}{13/32} = \frac{2}{13}$$

x	1	2	3	4
$P(X=x \mid Y=1)$	$4/13$	$4/13$	$3/13$	$2/13$

More about Bella

Bella has a wonderful veterinarian who has been in practice for many decades. He can usually examine a sick dog and narrow down the possible conditions to one of three, and then do a blood test. Suppose that in one instance, at first he assigns an equal probability to three conditions that a patient might have, call them $c_1, c_2, \& \ c_3$. He then performs a blood test that will be positive with probability 0.8 if the patient has c_1, 0.6 if the patient has condition c_2, and 0.4 if the patient has c_3. If the blood test turns out to be positive, now what are the probabilities of the three conditions given this new information?

Solution provided