Probability Review

CS 70, Summer 2019

8/13/19

Exam Tips

- Skim the exam before starting! Get a sense for what topic each question is testing
- Problems are usually clear about what parts depend on each other / whether you can use previous parts for later parts without doing the previous parts.
- Long answers sometimes award partial credit. Skim instructions!

Resources for Studying

- **Past exams**: skip questions not in coverage
- **Discussion**: try the questions again!
- **Lecture slides**: examples and exercises
- **Notes**: mainly for theorem statements and examples; less emphasis on proof
- **Homework**: lower priority; try to get key ideas and techniques from each problem

Counting I

- **First Rule of Counting**: multiply when choices *don't depend on each other*.
  - Lunch menu
- **Second Rule of Counting**: is there *overcounting* by the *same amount*? Divide by overcounts.
  - Anagrams with repeated letters
- **Stars and bars**: for *indistinguishable* items into *distinguishable* bins.
  - $x_1 + x_2 + x_3 = n$, splitting dollars

Counting II

- **Other counting techniques**:
  - Cases
  - Complement
  - Symmetry
  - Inclusion-exclusion
- **Combinatorial Proof**
  - Be explicit about the object you are counting:
    - Ex. all **binary** bitstrings
  - Count the LHS and RHS in two different ways.
    - Ex. for $\sum_{i=0}^{n} \binom{n}{i} = 2^n$. LHS is by cases, RHS is counted by the First Rule, 2 choices per bit.
  - Often times, summation = count by cases. Be explicit about what the cases are.

Discrete Probability I

- **Probability space**: if *uniform*, we reduce it to a counting problem. $\frac{\text{\# outcomes in event}}{\text{\# total outcomes}}$
- **Conditional probability**: a change in probability space: $\mathbb{P}[A|B] = \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[B]}$
- **Total probability rule**: "probability by casework." Cases are events B_i: $\mathbb{P}[A] = \mathbb{P}[A|B_1] \cdot \mathbb{P}[B_1] + \ldots + \mathbb{P}[A|B_n] \cdot \mathbb{P}[B_n]$
  - When in doubt, draw the tree

Discrete Probability II

- **Bayes**: "flip the conditioning"
  - Numerator comes from definition of conditional
  - Denominator comes from total probability rule
- **Event intersections**: chain rule!
  - $\mathbb{P}[A_1 \cap \ldots \cap A_n] = \mathbb{P}[A_1] \cdot \mathbb{P}[A_2|A_1] \cdot \mathbb{P}[A_3|A_1 \cap A_2] \cdots$
  - If events are **mutually independent**, can multiply their probabilities.
  - Pairwise **does not imply** mutual independence
- **Union bound**: sum of individual probabilities is an upper bound on probability of union

Discrete RVs I

- **Definition**: RVs are *functions* from outcomes to real numbers
  - $\{X = i\}$ should be treated as an event.
- **Special kinds of RVs**:
  - Bernoulli: models a single coin flip
    - Ex. any indicator variable
  - Binomial: sum of i.i.d. Bernoullis
    - Ex. number of heads in n coin flips
  - Geometric: # of i.i.d. trials until a "success"
    - Ex. number days until winning a lottery
  - Poisson: models rare events, given "rate"
    - Ex. typos on a page

Discrete RVs II

- **Functions of RVs**: are another RV!
  - Values will change using the function; probabilities don't change but may get *merged*
- **Expectation**: *weighted average* of values
  - $\mathbb{E}[X] = \sum_{a \in \text{values}(X)} a \cdot \mathbb{P}[X = a]$
- **Linearity of expectation**: for *any two* RVs X, Y (regardless of *independence*):

$$\mathbb{E}[aX + bY] = a\,\mathbb{E}[X] + b\,\mathbb{E}[Y]$$

Discrete RVs III

- **Tail sum**: another way of computing expectation for *integer valued* RVs.
  - $\mathbb{E}[X] = \sum_{i=1}^{\infty} \mathbb{P}[X \geq i]$
  - Used this for expectation of *geometric*
- **Coupon collector**:
  - Time after getting the $(i-1)$-th coupon until getting the i-th is Geometric$(\frac{n-i+1}{n})$
  - Use approximation $\sum_{i=1}^{n} \frac{1}{i} \approx \ln n$.
  - Tip: For questions, always draw the intervals and analyze them from scratch. Ex: RandomSort is **not** coupon collector but uses same strategy.

Discrete RVs IV

- **Variance**: expected distance to mean μ
  - $\text{Var}(X) = \mathbb{E}[(X-\mu)^2] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$
  - For computations, latter is more common.
  - Standard deviation $\sigma(X)$ is $\sqrt{\text{Var}(X)}$
  - If X, Y independent, then $\text{Var}(X+Y) = \text{Var}(X) + \text{Var}(Y)$.
  - If c is a constant, $\text{Var}(cX) = c^2\,\text{Var}(X)$.
  - If c is a constant, $\text{Var}(X+c) = \text{Var}(X)$.

Discrete RVs V

- **Covariance**: do RVs trend with each other?
  - $\text{Cov}(X,Y) = \mathbb{E}[(X-\mu_X)(Y-\mu_Y)] = \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y]$
  - If X, Y independent, then $\text{Cov}(X,Y) = 0$. The converse **is not true.**
  - $\text{Cov}(X,X) = \text{Var}(X)$
  - Covariance is **bilinear**
- **Correlation**: $\frac{\text{Cov}(X,Y)}{\sigma(X)\sigma(Y)}$
  - Always between -1 and 1 *inclusive*; equals -1 and 1 when $Y = \pm X$

Indicator Questions!

- Step 1: What is being counted here?
  - Usually: indicator for each thing that *could* contribute to the count
- Step 2: Distribution of each indicator?
- Step 3: Expectation
  - Usually: linearity of expectation
- Step 4: Variance: Expand $(X_1 + \ldots + X_n)^2$
  - What does $X_i X_j = 1$ mean in the problem?
  - Careful! Check if $X_i X_j$ is the same for all i, j.
  - Don't forget to subtract $\mathbb{E}[X]^2$

Break

What was your favorite topic from CS 70?
Least favorite?

Concentration I

- **Markov**: non-negative X, $\mathbb{P}[X \geq c] \leq \frac{\mathbb{E}[X]}{c}$
  - Check if RV has *non-negative* values
  - If not, can we shift it so that it's non-negative?
  - There are distributions where Markov is tight (i.e. achieve *equality*)
- **Chebyshev**: $\mathbb{P}[|X - \mathbb{E}[X]| \geq c] \leq \frac{\mathrm{Var}(X)}{c^2}$
  - Proven by applying Markov to $(X - \mathbb{E}[X])^2$.
  - Works for *all random variables*
  - Interpret $|X - \mathbb{E}[X]|$ as *distance to mean*

Concentration II

- **Law of Large Numbers**: sample mean approaches true mean
  - Formally, let $X_1, X_2, \ldots$ be i.i.d. RVs. Let $\mu = \mathbb{E}[X_i]$. Let $S_n = X_1 + X_2 + \ldots + X_n$. *Sample mean*: $\frac{1}{n} S_n$
  - The LLN says that with probability $\to 1$, the sample mean is in an ε-*interval* around μ.

Confidence Interval Problems!!

- We're aware we didn't cover these in depth; we won't ask for one from scratch, but would build it up step by step if we do test it.
- Step 1: What RV do we care about?
  - Compute its expectation and variance.
- Step 2: Mark important values (e.g. $\mathbb{E}[X]$) on a number line
  - Which interval is the event I care about?
- Step 3: Convert lower bounds to upper bounds; apply Chebyshev
  - E.g. change "at least 95% confidence" to "at most 5% on the complement"

Continuous Probability I

- **Events** are intervals
- **CDF** at x tells us $\mathbb{P}[X \leq x]$
  - CDF actually corresponds to an *event*
  - Usually when we compute PDF, *find the CDF first* and differentiate
- Area under **PDF** = probability of intervals
  - PDF should integrate to 1
  - Specifies "$\mathbb{P}[X = x]$" in the sense where $\mathbb{P}[X = x] \approx f_X(x)dx$

Continuous Probability II

- **Special kinds of RVs**:
  - Exponential: continuous version of geometric
    - Memoryless
  - Gaussian/ Normal:
    - Can shift and scale to a *standard normal*
    - Sum of two independent normals is also normal
- **CLT**: (normalized) averages $\sim \mathcal{N}(0,1)$.
  - Formally, let $X_1, X_2, \ldots$ be i.i.d. RVs. Let $\mu = \mathbb{E}[X_i]$. Let $A_n = \frac{X_1+X_2+\ldots+X_n}{n}$.
  - The CLT says as $n \to \infty$, $A_n' = \frac{A_n-\mu}{\sigma/\sqrt{n}}$ follows a $\mathcal{N}(0,1)$ distribution.

Tricks for Continuous!!

- **Joint distribution**: volumes under a surface
  - Translate your event (e.g. $t \leq X \leq Y$) into a region on the x-y plane. Integrate the joint PDF over this distribution.
  - For joint distributions that are *constants*, can work with shapes in the plane instead of integrals
- **Conditioning/total probability**
  - Try proceeding as you would with discrete.
  - Replace $\sum$ with $\int$, $\mathbb{P}[X = x]$ with $f_X(x)dx$

Markov Chains I

- **Markov property**: "memorylessness"
  - The next state only depends on current
- The **matrix and vector** view
  - Place transitions in a matrix. Rows sum to 1.
  - The distribution at time n is a *row vector*.
  - Given initial distribution $\mu^{(0)}$, the distribution at time n is $\mu^{(0)}P^n$.

Markov Chains II

- **First step** analysis
  - Use for expected hitting time, A before B
  - From each state, take a single step.
  - Where can we end up after that step? With what probability do we end up at each following state?
  - Does leaving that state contribute any "cost"? (See Three Tails HW)

Markov Chains III

- **Irreducible**: possible to go from any state to any other, in finite # of steps.
- **Period**: for state i, its period is the GCD of all *walks* from i to i
  - Aperiodic = all states have period 1
- **Stationary distribution**: $\pi = \pi P$
  - If irreducible, stationary *exists* and *is unique* and the *fraction of time* in state i approaches π_i
  - If irreducible, all states have the *same period*. If also aperiodic, the *probability of being in state i* ($\mu_i^{(n)}$) at a sufficiently large time $\to \pi_i$

What's Next?

More Probability: EE 126, Stat 134, Stat 140

AI / ML: CS 188, CS 189, Stat 154, Data 102

Data Science: Data 8, Data 100

Statistics: Stat 135, Stat 150

Grad: CS 271, Stat 205