



Essential Discrete Distributions

Bernoulli to Poisson, with ML uses

Probability Theory • Lecture 4

Applied Data Analysis — Module I

October 2026

Roadmap & learning objectives

A “who’s who” of discrete distributions, each built from a *story* and treated with one derivation pattern:

story $\rightarrow$ PMF (check it sums to 1) $\rightarrow \mathbb{E} \rightarrow \text{Var} \rightarrow$ what it models.

- **Bernoulli** and **Binomial**: counting successes in independent trials; sums of binomials.
- **Geometric** and **Negative Binomial**: waiting for successes; the tail-sum formula; memorylessness (with proof).
- **Hypergeometric**: sampling *without* replacement; variance with the finite-population correction.
- **Poisson**: rare events; the Poisson limit theorem (full proof); sums and thinning; the Poisson process previewed.
- A map of how the families relate, and one table to remember.
- ML tie-in: Bernoulli likelihood $\Rightarrow$ log-loss; Binomial error bars; Poisson regression and overdispersion.



Where we stand after Lecture 3

$$\mathbb{E}[X] = \sum_x x p_X(x); \quad \mathbb{E}[aX+bY] = a\mathbb{E}X+b\mathbb{E}Y; \quad \text{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}X)^2; \quad X = \sum_i \mathbf{1}\{A_i\} \Rightarrow \mathbb{E}X = \sum_i \mathbb{P}(A_i).$$

- Independent X_i : $\text{Var}(\sum X_i) = \sum \text{Var}(X_i)$; in general covariances appear.
- The indicator method computes means of *counts* without touching dependence.
- Lecture 2 already gave us one PMF: $\mathbb{P}(k \text{ successes in } n \text{ trials}) = \binom{n}{k} p^k (1-p)^{n-k}$.

Today's question

A handful of stories — flip, count, wait, sample, rare event — generate almost every discrete distribution you will meet. Which story produces which PMF, and what are its mean and variance?





1.

Bernoulli trials and the Binomial

Bernoulli(p) — one trial

Definition

$X \sim \text{Bernoulli}(p)$ if $\mathbb{P}(X = 1) = p$ and $\mathbb{P}(X = 0) = 1 - p$, $0 \leq p \leq 1$.

Theorem

$\mathbb{E}[X] = p$, $\text{Var}(X) = p(1 - p)$, $\mathbb{E}[X^k] = p$ for every $k \geq 1$.

Proof.

$\mathbb{E}[X] = 1 \cdot p + 0 \cdot (1 - p) = p$. Since $X^k = X$ for $X \in \{0, 1\}$, $\mathbb{E}[X^k] = p$; hence $\text{Var}(X) = p - p^2$. $\square$

- A Bernoulli variable is an indicator: $X = \mathbf{1}\{A\}$ with $\mathbb{P}(A) = p$. Lecture 3's indicator facts are Bernoulli facts.
- The variance $p(1 - p)$ is maximal at $p = \frac{1}{2}$ (a fair coin is the most unpredictable) and vanishes at $p \in \{0, 1\}$.
- Models: a click / no click, defective / fine, a label $y \in \{0, 1\}$, one step of a random walk.



Binomial(n, p) — counting successes

Story: n independent trials, each a success with probability p ; X = number of successes.

Theorem

$X \sim \text{Bin}(n, p)$ has PMF $p_X(k) = \binom{n}{k} p^k (1-p)^{n-k}$, $k = 0, \dots, n$; $\mathbb{E}[X] = np$, $\text{Var}(X) = np(1-p)$.

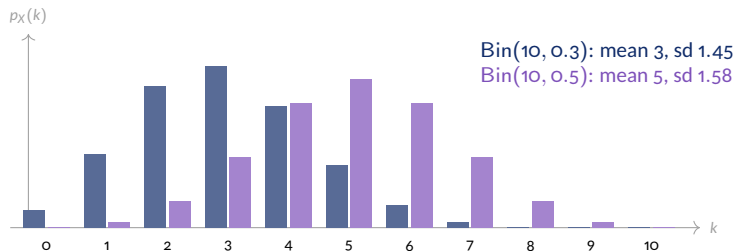
Proof.

PMF: Lecture 2 (each pattern with k successes has probability $p^k (1-p)^{n-k}$; there are $\binom{n}{k}$ patterns). Write $X = \sum_{i=1}^n \mathbf{1}\{A_i\}$, A_i = “trial i succeeds”. Indicator method: $\mathbb{E}[X] = np$. The indicators are independent, so variances add: $\text{Var}(X) = np(1-p)$. □

- Sums to 1 by the binomial theorem: $\sum_k \binom{n}{k} p^k (1-p)^{n-k} = (p + 1-p)^n = 1$.
- $\text{Bin}(1, p) = \text{Bernoulli}(p)$. Computing $\mathbb{E}$ from the PMF directly is painful; indicators make it one line.
- Relative frequency X/n has mean p and variance $p(1-p)/n$ — the sample-mean theorem in disguise.



The shape of the Binomial



- Unimodal; symmetric iff $p = \frac{1}{2}$; the mode is $\lfloor (n+1)p \rfloor$ (from $p_X(k)/p_X(k-1) = \frac{n-k+1}{k} \cdot \frac{p}{1-p}$).
- Mean np , spread $\sqrt{np(1-p)}$: as n grows the bulk concentrates around np on a scale $\sqrt{n}$ — the CLT picture (Lecture 14).



Sums of binomials

Theorem

If $X \sim \text{Bin}(n, p)$ and $Y \sim \text{Bin}(m, p)$ are independent, then $X + Y \sim \text{Bin}(n + m, p)$.

Proof 1 — the story.

X counts successes in n trials, Y in m further independent trials with the same p ; together they are $n + m$ independent trials. □

Proof 2 — convolution.

$$\mathbb{P}(X + Y = k) = \sum_j \mathbb{P}(X = j) \mathbb{P}(Y = k - j) = \sum_j \binom{n}{j} \binom{m}{k-j} p^k (1-p)^{n+m-k} = \binom{n+m}{k} p^k (1-p)^{n+m-k}$$
by Vandermonde's identity (Lecture 1). □

Two warnings

Different p 's: $X + Y$ is *not* binomial (it is “Poisson-binomial”). Dependent trials: not binomial either — e.g. $2X$ takes only even values.



Worked examples — Binomial

Defective items

20 items, each defective independently with $p = 0.02$. $\mathbb{P}(\geq 1 \text{ defect}) = 1 - 0.98^{20} \approx 0.332$;
expected number 0.4, sd $\sqrt{20 \cdot 0.02 \cdot 0.98} \approx 0.63$.

Guessing a multiple-choice exam

30 questions, 4 options each, pure guessing: $X \sim \text{Bin}(30, \frac{1}{4})$, $\mathbb{E}X = 7.5$, sd ≈ 2.37 .
 $\mathbb{P}(X \geq 15) = \sum_{k \geq 15} \binom{30}{k} (\frac{1}{4})^k (\frac{3}{4})^{30-k} \approx 0.002$ — a pass by luck is a 3σ event.

A/B test

Variant B shown to $n = 1000$ users; conversions $X \sim \text{Bin}(1000, p)$. If truly $p = 0.05$, then
 $\mathbb{E}X = 50$, sd ≈ 6.9 . Observing $X = 65$ is more than 2 sd above 50 — evidence (not proof) that B
beats a 5% baseline. Lecture 14 makes this a test.





2.

Waiting times: Geometric and Negative Binomial

Geometric(p) — waiting for the first success

Story: independent trials with success probability p ; X = the number of the trial on which the first success occurs.

Theorem

$$p_X(k) = (1-p)^{k-1}p \text{ for } k = 1, 2, \dots; \quad \mathbb{P}(X > k) = (1-p)^k; \quad \mathbb{E}[X] = \frac{1}{p}, \quad \text{Var}(X) = \frac{1-p}{p^2}.$$

Proof.

PMF: the first $k-1$ trials fail and the k -th succeeds; independence gives the product. Tail: $X > k$ iff the first k trials all fail. Sums to 1: $\sum_{k \geq 1} (1-p)^{k-1}p = p \cdot \frac{1}{1-(1-p)} = 1$ (geometric series — hence the name). Mean and variance: next slide. □

- Infinite support: no upper bound on the wait, but $\mathbb{P}(X > k) \rightarrow 0$ exponentially fast.
- Convention alert: some texts count *failures before* the first success, $Y = X - 1 \in \{0, 1, \dots\}$, with $\mathbb{E}Y = \frac{1-p}{p}$.



The tail-sum formula, and the geometric mean and variance

Theorem (Tail-sum formula)

If X takes values in $\{0, 1, 2, \dots\}$, then $\mathbb{E}[X] = \sum_{k \geq 1} \mathbb{P}(X \geq k) = \sum_{k \geq 0} \mathbb{P}(X > k)$.

Proof.

$X = \sum_{k \geq 1} \mathbf{1}\{X \geq k\}$ (for $X = x$, exactly x of the indicators equal 1). Take expectations — the indicator method with infinitely many nonnegative terms. $\square$

Geometric mean

$\mathbb{E}[X] = \sum_{k \geq 0} \mathbb{P}(X > k) = \sum_{k \geq 0} (1 - p)^k = \frac{1}{p}$. Expected number of rolls until the first six: 6.

Geometric variance

With $q = 1 - p$: $\mathbb{E}[X(X - 1)] = \sum_k k(k - 1)q^{k-1}p = pq \sum_k k(k - 1)q^{k-2} = pq \cdot \frac{2}{p^3} = \frac{2q}{p^2}$ (differentiate the geometric series twice). Hence $\mathbb{E}[X^2] = \frac{2q}{p^2} + \frac{1}{p}$ and $\text{Var}(X) = \frac{2q}{p^2} + \frac{1}{p} - \frac{1}{p^2} = \frac{q}{p^2}$.



Memorylessness

Theorem (Memorylessness)

If $X \sim \text{Geom}(p)$, then for all $m, n \geq 0$: $\mathbb{P}(X > m + n | X > m) = \mathbb{P}(X > n)$.

Proof.

$$\mathbb{P}(X > m + n | X > m) = \frac{\mathbb{P}(X > m + n)}{\mathbb{P}(X > m)} = \frac{(1 - p)^{m+n}}{(1 - p)^m} = (1 - p)^n = \mathbb{P}(X > n). \quad \square$$

- Having waited m trials without success tells you *nothing* about how much longer you will wait. The coin does not “owe” you a head.
- **Converse (exercise):** the geometric is the *only* such distribution on $\{1, 2, \dots\}$. With $G(n) = \mathbb{P}(X > n)$, memorylessness says $G(m + n) = G(m)G(n)$, whose only solutions are $G(n) = q^n$.
- The continuous analogue is the exponential distribution (Lecture 6), with the same proof.

The gambler's fallacy

“Red has come up five times, so black is due” claims that roulette remembers. Memorylessness says it does not.



Negative Binomial(r, p) — waiting for the r -th success

Story: independent trials; X = the trial on which the r -th success occurs.

Theorem

$$p_X(k) = \binom{k-1}{r-1} p^r (1-p)^{k-r}, \quad k = r, r+1, \dots; \quad \mathbb{E}[X] = \frac{r}{p}, \quad \text{Var}(X) = \frac{r(1-p)}{p^2}.$$

Proof.

PMF: trial k is a success and exactly $r - 1$ of the previous $k - 1$ trials are successes — $\binom{k-1}{r-1}$ arrangements, each of probability $p^r (1-p)^{k-r}$. Mean and variance: $X = G_1 + \dots + G_r$, where G_i is the waiting time from the $(i-1)$ -th to the i -th success. The G_i are i.i.d. $\text{Geom}(p)$ (fresh independent trials); add means, and add variances by independence. □

- $r = 1$ recovers the geometric. Same convention alert: many texts count only the failures, $X - r$.
- ML use: the standard **overdispersed** alternative to the Poisson for count data (Section 6).



Worked example — the coupon collector

There are m distinct coupons; each purchase yields a uniformly random one. T = purchases needed to collect all m .

Theorem

$$\mathbb{E}[T] = m \left(1 + \frac{1}{2} + \cdots + \frac{1}{m} \right) = m H_m \approx m \ln m + 0.577 m.$$

Proof.

$T = G_1 + \cdots + G_m$, where G_i is the number of purchases to get a *new* coupon when $i - 1$ are already owned. Each purchase is new with probability $\frac{m-i+1}{m}$, so $G_i \sim \text{Geom}\left(\frac{m-i+1}{m}\right)$ and $\mathbb{E}[G_i] = \frac{m}{m-i+1}$. Sum over i and reverse the order of summation. □

- $m = 50$ football stickers: $\mathbb{E}[T] = 50 \cdot H_{50} \approx 225$ purchases. The last few coupons cost most of the money.
- The G_i are independent, so $\text{Var}(T) = \sum_i \text{Var}(G_i) < \frac{\pi^2}{6} m^2$: sd of order m , much smaller than the mean $m \ln m$ for large m .
- Same structure: hash tables, random testing (“how many random inputs until every branch is hit”), Lecture 3’s distinct-values example.





3.

Sampling without replacement: Hypergeometric

Hypergeometric(N, K, n)

Story: an urn with N balls, K of them “special”. Draw n *without replacement*; X = number of special balls drawn.

Theorem

$$p_X(k) = \frac{\binom{K}{k} \binom{N-K}{n-k}}{\binom{N}{n}}, \quad \mathbb{E}[X] = n \frac{K}{N}, \quad \text{Var}(X) = n \frac{K}{N} \left(1 - \frac{K}{N}\right) \frac{N-n}{N-1}.$$

Proof of the PMF and the mean.

All $\binom{N}{n}$ subsets are equally likely; those with exactly k special balls number $\binom{K}{k} \binom{N-K}{n-k}$ (Lecture 1).

Mean: $X = \sum_{i=1}^n \mathbf{1}\{A_i\}$, A_i = “the i -th ball drawn is special”. By symmetry $\mathbb{P}(A_i) = K/N$ for every i (the i -th draw is equally likely to be any ball). Indicator method: $\mathbb{E}X = nK/N$ — dependence ignored. □

Sums to 1 by Vandermonde: $\sum_k \binom{K}{k} \binom{N-K}{n-k} = \binom{N}{n}$. Once more, Lecture 1’s counting pays off.



Hypergeometric variance — the finite-population correction

Proof of the variance.

Use the dependent-indicator formula. $\mathbb{P}(A_i) = \frac{K}{N}$; for $i \neq j$, $\mathbb{P}(A_i \cap A_j) = \frac{K}{N} \cdot \frac{K-1}{N-1}$ (chain rule).

$$\text{Var}(X) = n \frac{K}{N} \left(1 - \frac{K}{N}\right) + n(n-1) \left[\frac{K(K-1)}{N(N-1)} - \frac{K^2}{N^2} \right] = n \frac{K}{N} \left(1 - \frac{K}{N}\right) \left[1 - \frac{n-1}{N-1}\right],$$

and $1 - \frac{n-1}{N-1} = \frac{N-n}{N-1}$. (The bracketed covariance term is negative: drawing a special ball makes the next one less likely.) □

- Compare with $\text{Bin}(n, K/N)$: same mean, variance multiplied by the **finite-population correction** $\frac{N-n}{N-1} < 1$. Sampling without replacement is *less* variable — you cannot draw the same ball twice.
- If $n = N$ the variance is 0: you drew everything, $X = K$ for sure.
- If $N \gg n$, the correction $\rightarrow 1$ and $\text{Hyp}(N, K, n) \approx \text{Bin}(n, K/N)$: with a huge urn, replacement hardly matters.



Worked example — quality control, revisited

Batch of $N = 100$ items with $K = 5$ defective; inspect $n = 10$ without replacement.

Exact (hypergeometric) vs. approximate (binomial)

$$\mathbb{P}(X = 0) = \frac{\binom{95}{10}}{\binom{100}{10}} \approx 0.584, \quad \text{Bin}(10, 0.05) : 0.95^{10} \approx 0.599.$$

$\mathbb{E}X = 0.5$ in both models; $\text{Var } X = 0.475 \cdot \frac{20}{99} \approx 0.432$ (hyp.) vs. 0.475 (bin.).

Capture-recapture (Lincoln-Petersen)

Tag $K = 50$ fish, release, later catch $n = 40$ and find $k = 8$ tagged. Under the hypergeometric model $\mathbb{E}X = nK/N$; setting $8 = 40 \cdot 50/N$ gives the estimate $\hat{N} = 250$. Ecology, epidemiology, and estimating the number of bugs left in software all use this idea.

When to use which

Without replacement and n a non-negligible fraction of N (say $> 5\%$): hypergeometric. Otherwise the binomial is simpler and accurate.





4.

Rare events: the Poisson distribution

Poisson(λ)

Definition

$X \sim \text{Pois}(\lambda)$, $\lambda > 0$, if $p_X(k) = e^{-\lambda} \frac{\lambda^k}{k!}$, $k = 0, 1, 2, \dots$

Theorem

$$\sum_k p_X(k) = 1, \quad \mathbb{E}[X] = \lambda, \quad \text{Var}(X) = \lambda.$$

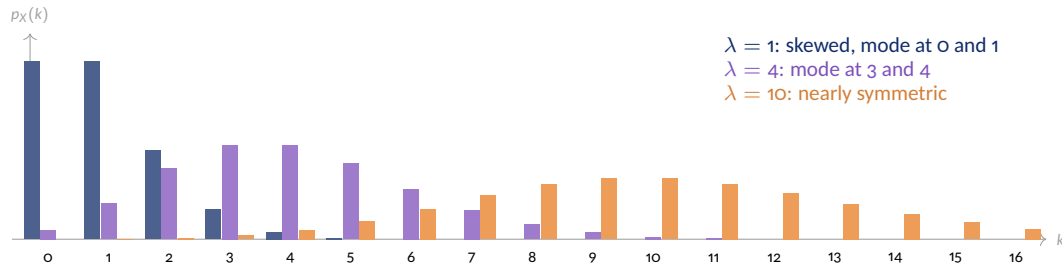
Proof.

$\sum_{k \geq 0} \frac{\lambda^k}{k!} = e^\lambda$ (exponential series). Mean: $\mathbb{E}X = \sum_{k \geq 1} k e^{-\lambda} \frac{\lambda^k}{k!} = \lambda e^{-\lambda} \sum_{k \geq 1} \frac{\lambda^{k-1}}{(k-1)!} = \lambda$. Factorial moment: $\mathbb{E}[X(X-1)] = \lambda^2 e^{-\lambda} \sum_{k \geq 2} \frac{\lambda^{k-2}}{(k-2)!} = \lambda^2$, so $\mathbb{E}[X^2] = \lambda^2 + \lambda$ and $\text{Var}(X) = \lambda$. □

- One parameter controls both location and spread: **mean = variance**. A testable prediction that real data often violate (Section 6).
- The trick — factorial moments $\mathbb{E}[X(X-1)]$ — is what generating functions systematize in Lecture 5.



The shape of the Poisson



- $p_X(k)/p_X(k-1) = \lambda/k$: probabilities rise while $k < \lambda$ and fall afterwards; the mode is $\lfloor \lambda \rfloor$ (two modes if λ is an integer).
- For large λ the shape approaches a bell with sd $\sqrt{\lambda}$ — the Normal approximation (Lecture 8).



The Poisson limit theorem

Theorem (Law of rare events)

Let $X_n \sim \text{Bin}(n, p_n)$ with $np_n \rightarrow \lambda > 0$ as $n \rightarrow \infty$. Then for every fixed k , $\mathbb{P}(X_n = k) \rightarrow e^{-\lambda} \frac{\lambda^k}{k!}$.

Proof.

Write $p_n = \lambda_n/n$ with $\lambda_n \rightarrow \lambda$. Then

$$\binom{n}{k} p_n^k (1 - p_n)^{n-k} = \underbrace{\frac{n(n-1) \cdots (n-k+1)}{n^k}}_{\rightarrow 1} \cdot \frac{\lambda_n^k}{k!} \cdot \underbrace{\left(1 - \frac{\lambda_n}{n}\right)^n}_{\rightarrow e^{-\lambda}} \cdot \underbrace{\left(1 - \frac{\lambda_n}{n}\right)^{-k}}_{\rightarrow 1}.$$

The first factor is a product of k terms $\frac{n-j}{n} \rightarrow 1$; the third uses $(1 - x_n/n)^n \rightarrow e^{-x}$ when $x_n \rightarrow x$. □

Many independent trials (n large), each rarely successful (p small), moderate expected count ($np \approx \lambda$)
 $\implies$ the number of successes is approximately $\text{Pois}(\lambda)$.

How good is the approximation?

$n = 20, p = 0.02$ (so $\lambda = 0.4$) and $n = 1000, p = 0.004$ (so $\lambda = 4$):

k	Bin(20, 0.02)	Pois(0.4)	Bin(1000, 0.004)	Pois(4)
0	0.668	0.670	0.0181	0.0183
1	0.272	0.268	0.0727	0.0733
2	0.053	0.054	0.1460	0.1465
3	0.006	0.007	0.1951	0.1954
4	0.0006	0.0007	0.1954	0.1954

- Already at $n = 20$ the error is in the third decimal; a standard bound is $\sum_k |\mathbb{P}(X_n = k) - \mathbb{P}(\text{Pois}(\lambda) = k)| \leq 2np^2$.
- Rule of thumb: np^2 small (e.g. $n \geq 20, p \leq 0.05$) — and the Poisson is a fine substitute, with one parameter instead of two.
- This is why Lecture 3's fixed points had mean 1 *and* variance 1: n cards, each fixed with probability $1/n$, nearly independent — $\text{Pois}(1)$, and indeed $\mathbb{P}(\text{no fixed point}) \rightarrow e^{-1} \approx 0.368$.



Sums and thinning of Poisson variables

Theorem (Superposition)

If $X \sim \text{Pois}(\lambda)$ and $Y \sim \text{Pois}(\mu)$ are independent, then $X + Y \sim \text{Pois}(\lambda + \mu)$.

Proof.

$\mathbb{P}(X + Y = k) = \sum_{j=0}^k e^{-\lambda} \frac{\lambda^j}{j!} e^{-\mu} \frac{\mu^{k-j}}{(k-j)!} = \frac{e^{-(\lambda+\mu)}}{k!} \sum_{j=0}^k \binom{k}{j} \lambda^j \mu^{k-j} = e^{-(\lambda+\mu)} \frac{(\lambda+\mu)^k}{k!}$ by the binomial theorem. □

Theorem (Thinning)

Let $N \sim \text{Pois}(\lambda)$ and mark each of the N events independently with probability q . The number M of marked events is $\text{Pois}(q\lambda)$, independent of the number $N - M$ of unmarked ones.

Proof sketch.

$\mathbb{P}(M = m, N - M = u) = \mathbb{P}(N = m + u) \binom{m+u}{m} q^m (1-q)^u = e^{-q\lambda} \frac{(q\lambda)^m}{m!} \cdot e^{-(1-q)\lambda} \frac{((1-q)\lambda)^u}{u!}$ — a product of two Poisson PMFs. □



Emails from two independent senders add up; a spam filter splits an inbox into two independent Poisson streams. Probability Theory • Lecture 4

Poisson as a model — and the Poisson process previewed

When counts are Poisson

Events occur **independently**, at a **constant rate**, **one at a time**. Then the number of events in any window of length t is $\text{Pois}(\lambda t)$, and counts in disjoint windows are independent — the **Poisson process** with rate λ (waiting times between events are exponential; Lecture 6).

Typically fits

Radioactive decays per second; calls per minute to a call centre (within a stationary hour); typos per page; goals per football match ($\lambda \approx 2.7$).

Typically fails

Clustered events (aftershocks, retweets), batches (a bus of tourists), time-varying rates (lunch-hour traffic) — all give $\text{Var} > \mathbb{E}$: **overdispersion**.

Worked example

A server gets $\lambda = 3$ requests per second on average. $\mathbb{P}(\text{none in a second}) = e^{-3} \approx 0.05$; $\mathbb{P}(\geq 7) \approx 0.034$; in a 10-second window the count is $\text{Pois}(30)$, $\text{sd} \approx 5.5$.

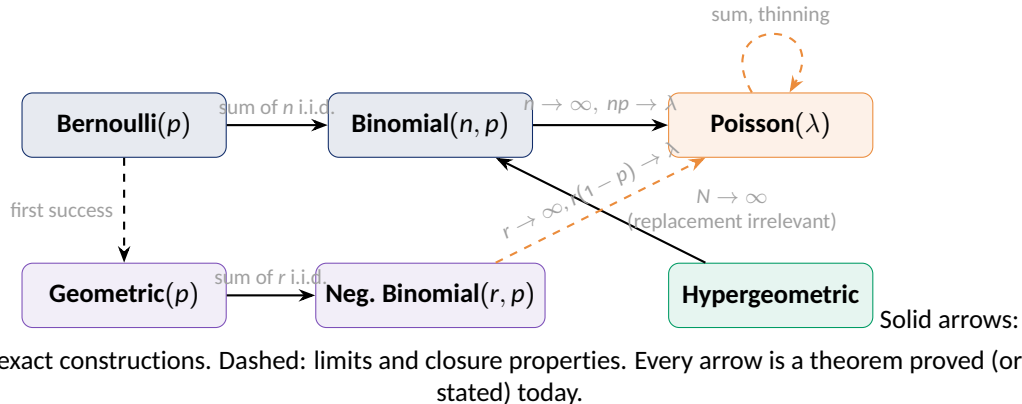




5.

The map and the table

How the families relate



One table to remember

Distribution	Story	Support	$\mathbb{E}$	Var
Bernoulli(p)	one trial	$\{0, 1\}$	p	$p(1 - p)$
Binomial(n, p)	successes in n trials	$\{0, \dots, n\}$	np	$np(1 - p)$
Geometric(p)	trials to first success	$\{1, 2, \dots\}$	$1/p$	$(1 - p)/p^2$
Neg. Binomial(r, p)	trials to r -th success	$\{r, r+1, \dots\}$	r/p	$r(1 - p)/p^2$
Hypergeometric(N, K, n)	special balls, no replacement	$\{0, \dots, n\}$	nK/N	$n \frac{K}{N} (1 - \frac{K}{N}) \frac{N-n}{N-1}$
Poisson(λ)	rare events, rate λ	$\{0, 1, \dots\}$	λ	λ

- Do not memorize the variances: re-derive them (indicators, tail sums, factorial moments) — that is what the exam rewards.
- Watch the geometric / negative binomial **conventions** (trials vs. failures).





6.

Machine learning connections

Bernoulli likelihood = log-loss

Labels $y_i \in \{0, 1\}$, model probabilities $\hat{p}_i = \mathbb{P}(Y_i = 1 \mid x_i)$. If the Y_i are independent Bernoulli($\hat{p}_i$), the probability of the observed labels is

$$L = \prod_{i=1}^n \hat{p}_i^{y_i} (1 - \hat{p}_i)^{1-y_i}, \quad -\frac{1}{n} \log L = -\frac{1}{n} \sum_{i=1}^n \left[y_i \log \hat{p}_i + (1 - y_i) \log(1 - \hat{p}_i) \right].$$

The right-hand side is the **binary cross-entropy** (log-loss). Maximizing the Bernoulli likelihood is minimizing log-loss.

- Logistic regression = Bernoulli likelihood + $\hat{p}_i = \sigma(\beta^\top x_i)$. The loss was never an arbitrary choice.
- For a single class-frequency parameter p : $\frac{d}{dp} \log L = \frac{\sum y_i}{p} - \frac{n - \sum y_i}{1-p} = 0 \Rightarrow \hat{p} = \bar{y}$. The maximum-likelihood estimate of a probability is the relative frequency (Module II develops this).



Binomial error bars on a test accuracy

#correct on a test set of size n is $\text{Bin}(n, a)$, a = true accuracy. The observed accuracy $\hat{a} = \text{\#correct}/n$ has

$$\mathbb{E}[\hat{a}] = a, \quad \text{Var}(\hat{a}) = \frac{a(1-a)}{n}, \quad \text{s.e.}(\hat{a}) = \sqrt{\frac{\hat{a}(1-\hat{a})}{n}} \text{ (plug-in).}$$

n	100	1,000	10,000	100,000
s.e. at $\hat{a} = 0.9$	0.030	0.0095	0.0030	0.00095

- Report $\hat{a} \pm 2$ s.e. (an approximate 95% interval; Lecture 14 justifies the “2”). On $n = 1000$, “0.90” means “somewhere in 0.88–0.92”.
- Comparing two models on the *same* test set: the errors are paired, and the relevant count is the number of examples on which exactly one model is right — again binomial (McNemar’s test, Module II).
- Rare positive class: the number of positives in the test set is approximately Poisson — a few dozen positives make recall estimates very noisy.



Poisson regression and overdispersion

Count targets $Y \in \{0, 1, 2, \dots\}$: visits, claims, purchases, defects.

Poisson regression (a GLM)

Model $Y | x \sim \text{Pois}(\lambda(x))$ with $\log \lambda(x) = \beta^\top x$ (log link keeps $\lambda > 0$ and makes effects multiplicative). Log-likelihood: $\sum_i [y_i \beta^\top x_i - e^{\beta^\top x_i} - \log y_i!]$ — concave in β , so easy to maximize. With no covariates, $\hat{\lambda} = \bar{y}$: the MLE of a Poisson rate is the sample mean.

Overdispersion

Poisson forces $\text{Var}(Y | x) = \mathbb{E}[Y | x]$. Real counts usually have $\text{Var} \gg \mathbb{E}$ (clustering, unobserved heterogeneity). Diagnostic: compare the sample variance of residual counts with the fitted mean. Fix: **Negative Binomial regression**, where $\text{Var}(Y | x) = \lambda + \lambda^2/r$ — an extra parameter r absorbs the excess; $r \rightarrow \infty$ recovers Poisson.

- Zero-inflation (too many exact zeros) is another common failure; mixture models handle it.
- Every link function, every loss in a GLM is the log-likelihood of one of today's distributions or their continuous cousins.





7.

Wrap-up

Pitfalls & summary

- **With vs. without replacement:** Binomial vs. Hypergeometric; same mean, variance shrinks by $\frac{N-n}{N-1}$.
- **Poisson assumes** independence, constant rate, no batches — and predicts $\text{Var} = \mathbb{E}$. Check it before using it.
- **Conventions:** geometric on $\{1, 2, \dots\}$ (trials) vs. $\{0, 1, \dots\}$ (failures); same for the negative binomial.
- **Sums:** binomials add only with equal p ; Poissons always add; geometrics add to a negative binomial.
- **Memorylessness** is a theorem about the geometric, and a fallacy about everything else.

Bernoulli $\xrightarrow{\sum}$ Binomial $\xrightarrow{n \rightarrow \infty, np \rightarrow \lambda}$ Poisson; Geometric $\xrightarrow{\sum}$ Negative Binomial;
Hypergeometric $\xrightarrow{N \rightarrow \infty}$ Binomial.

 Derivation toolkit used today: indicator method (means and variances of counts), tail-sum formula (waiting times), factorial moments (Poisson), Vandermonde and the binomial theorem (sums), and the limit $(1 - \lambda/n)^n \rightarrow e^{-\lambda}$

Seminar problems & next lecture

- 1 A fair die is rolled 12 times. Find the probability of exactly two sixes, and the mean and variance of the number of sixes.
- 2 Show that for $X \sim \text{Bin}(n, p)$, $\mathbb{P}(X \text{ is even}) = \frac{1+(1-2p)^n}{2}$. (Hint: expand $(q+p)^n$ and $(q-p)^n$.)
- 3 Prove that the geometric distribution is the only memoryless distribution on $\{1, 2, \dots\}$.
- 4 Two players alternate rolling a die; the first to roll a six wins. Find each player's probability of winning.
- 5 Deck of 52 cards, hand of 5. Find the expected number of aces and the variance; compare with the binomial approximation.
- 6 Typos occur at rate 0.2 per page. Find $\mathbb{P}(\text{a 10-page chapter has } \leq 1 \text{ typo})$ and $\mathbb{P}(\text{at least one page of 10 has } \geq 2 \text{ typos})$.
- 7 If $X \sim \text{Pois}(\lambda)$ and $Y \sim \text{Pois}(\mu)$ are independent, show that $X \mid X+Y=n$ is $\text{Bin}\left(n, \frac{\lambda}{\lambda+\mu}\right)$.
- 8 Show that the MLE of λ for an i.i.d. $\text{Pois}(\lambda)$ sample is $\bar{y}$, and that $\text{Var}(\hat{\lambda}) = \lambda/n$.



Next: generating functions — moments for free, and sums of independent variables in one line.