Lecture 4: Expectation

Wen Dingzhu

School of Information Science and Technology (SIST)
ShanghaiTech University

wendzh@shanghaitech.edu.cn

March 2, 2023

上海科技大学
ShanghaiTech University

Overview

Outline

Expectation of A Discrete R.V.

Definition

The *expected value* (also called the *expectation* or *mean*) of a discrete r.v. X whose distinct possible values are $x_1, x_2, \ldots$ is defined by

$$E(X) = \sum_{j=1}^{\infty} x_j P(X = x_j).$$

If the support is finite, then this is replaced by a finite sum. We can also write

$$E(X) = \sum_{x} \underbrace{x}_{\text{value}} \underbrace{P(X = x)}_{\text{PMF at } x},$$

where the sum is over the support of X.

Distribution

$$\text{if } E(X) = E(Y) \not\Rightarrow X \sim Y.$$

$$X = \begin{cases} 100, & \text{w.p. } 1/2 \\ 0, & \text{w.p. } 1/2 \end{cases} \qquad Y = \begin{cases} 70, & \text{w.p. } 1/2, \\ 30, & \text{w.p. } 1/2. \end{cases}$$

Theorem

If X and Y are discrete r.v.s with the same distribution, then $E(X) = E(Y)$ (if either side exists).

$$E(X) = 50 \quad = \quad E(Y) = 50.$$

Linearity

The expected value of a sum of r.v.s is the sum of the individual expected values.

Theorem

For any r.v.s X, Y and any constant c,

$$E(X+Y) = E(X) + E(Y),$$

$$E(cX) = cE(X).$$

Monotonicity of Expectation

$Z = X - Y. \Rightarrow Z \geqslant 0$, w.p. 1.

$\Rightarrow E(Z) \geqslant 0. \Rightarrow E(X-Y) \geqslant 0$

$\Rightarrow E(X) - E(Y) \geqslant 0$

Theorem

Let X and Y be r.v.s such that $X \geq Y$ with probability 1. Then $E(X) \geq E(Y)$, with equality holding if and only if $X = Y$ with probability 1.

$\Rightarrow E(X) \geqslant E(Y).$

Expectation via Survival Function

Theorem

Let X be a nonnegative integer-valued r.v. Let F be the CDF of X, and $G(x) = 1 - F(x) = P(X > x)$. The function G is called the survival function of X. Then

Tail distribution

$$E(X) = \sum_{n=0}^{\infty} G(n).$$

$= \sum_{n=0}^{\infty} P(X>n).$

$= \sum_{n=1}^{\infty} P(X\geq n)$

That is, we can obtain the expectation of X by summing up the survival function (or, stated otherwise, summing up tail probabilities of the distribution).

Proof

$m \geq n \geq 1.$

$$\sum_{n=0}^{\infty} G(n) = \sum_{n=1}^{\infty} P(X \geq n)$$

$$= \sum_{n=1}^{\infty} \left[\sum_{m=n}^{\infty} P(X=m) \right]$$

Fubini Theorem:

$$= \sum_{m=1}^{\infty} \sum_{n=1}^{m} P(X=m).$$

$$= \sum_{m=1}^{\infty} m \cdot P(X=m)$$

$$+ 0 \cdot P(X=0)$$

$$= E(X).$$

$$P(X \geq 1) = P(X=1) + P(X=2) + P(X=3) + \cdots$$

$+$

$$P(X \geq 2) = P(X=2) + P(X=3) + \cdots$$

$+$

$$P(X \geq 3) = P(X=3) + \cdots$$

$+$

$\vdots \qquad \vdots$

$$1 \cdot P(X=1) + 2 \cdot P(X=2) + 3 \cdot P(X=3) + \cdots$$

$$= \sum_{n=1}^{\infty} n \cdot P(X=n). + 0 \cdot P(X=0)$$

$$= \sum_{n=0}^{\infty} n \cdot P(X=n). = E(X).$$

Law Of The Unconscious Statistician (LOTUS)

Theorem

If X is a discrete r.v. and g is a function from $\mathbb{R}$ to $\mathbb{R}$, then

$$E(g(X)) = \sum_x g(x)P(X = x),$$

where the sum is taken over all possible values of X.

$$E(e^{tX}) = \sum_x e^{tx} P(X=x).$$

Variance and Standard Deviation

X: $\frac{1}{2}$ 49, $\frac{1}{2}$ 51

Y: 0, 100

$E(X) = 50 = E(Y)$.

$Var(X) = 1$ $Var(Y) = 2500$

Definition

The variance of an r.v. X is

$(x - EX)^2$ distance.

$$Var(X) = E(X - EX)^2.$$

The square root of the variance is called the *standard deviation (SD)*:

$$SD(X) = \sqrt{Var(X)}.$$

Properties of Variance

① Let $\mu = E(X)$. $Var(X) = E[(X-\mu)^2] = E[X^2 - 2\mu X + \mu^2] = E(X^2) - \mu^2$.

② $Y = X + c$ $E(Y) = E(X) + c$. $E[(Y - E(X) - c)^2] = E[(X - E(X))^2]$

③ $Y = cX$, $E(Y) = cE(X)$.

- ① For any r.v. X, $Var(X) = E(X^2) - (EX)^2$.
- ② $Var(X + c) = Var(X)$ for any constant c.
- ③ $Var(cX) = c^2 Var(X)$ for any constant c.
- ④ If X and Y are independent, then $Var(X + Y) = Var(X) + Var(Y)$.

$E(XY) = E(X)E(Y)$.

- ⑤ $Var(X) \geq 0$ with equality if and only if $P(X = a) = 1$ for some constant a.

④ $Var(X+Y) = E[(X+Y-(EX+EY))^2] = E\{[X-E(X)+Y-E(Y)]^2\}$

$= E\{[X-E(X)]^2 + 2[X-E(X)][Y-E(Y)] + [Y-E(Y)]^2\}$

$= Var(X) + 2E[X-E(X)]E[Y-E(Y)] + Var(Y)$

Properties of Variance

⑤ $Var(X) \geq 0 \iff E[(X-EX)^2] \geq 0$.

$\iff E(X^2) - (EX)^2 \geq 0$

$E(X^2) \geq (EX)^2$.

Outline

Story: Geometric Distribution

$k \geq 0$.

$P(X=k) = (1-p)^k \cdot p$.

$k+1$ Bernoulli trails.

first k trails failure,

the $(k+1)$-th trail success.

Consider a sequence of independent Bernoulli trials, each with the same success probability $p \in (0, 1)$, with trials performed until a success occurs. Let X be the number of **failures** before the first successful trial. Then X has the Geometric distribution with parameter p; we denote this by $X \sim \mathrm{Geom}(p)$.

Geometric PMF

① $E(X)=\sum_{k=0}^{\infty} k\cdot P(X=k)=\sum_{k=0}^{\infty} k\cdot q^k\cdot p=p\cdot\sum_{k=0}^{\infty} k\cdot q^k=\frac{q}{p}.=\frac{1-p}{p}.$

② $P(X\geq 0)=1.\quad k\geq 1.$

$P(X\geq k)=1-P(X<k)=1-P(X\leq k-1)=1-\sum_{j=0}^{k-1} P(X=j)=1-\sum_{j=0}^{k-1} q^j p.$

$=1-p\cdot\frac{1-q^k}{1-q}$

$=q^k.$

Theorem

If $X \sim \mathrm{Geom}(p)$, then the PMF of X is

$$P(X=k)=q^k p$$

for $k=0,1,2,\ldots$, where $q=1-p$.

$E(X)=\sum_{k=0}^{\infty} P(X>k)=\sum_{k=1}^{\infty} P(X\geq k)=\frac{1-p}{p}.$

$=\sum_{k=1}^{\infty} q^k.$

Memoryless Property

1° $k=0$, $P(X\geq n \mid X\geq 0) = P(X\geq n)$.

2° $k\geq 1$, $P(X\geq n+k \mid X\geq k) = \dfrac{P(X\geq n+k,\ X\geq k)}{P(X\geq k)} = \dfrac{P(X\geq n+k)}{P(X\geq k)} = \dfrac{q^{n+k}}{q^k}$

Theorem

If $X \sim \mathrm{Geom}(p)$, then for any positive integer n, $= q^n$.

$$P(X \geq n+k \mid X \geq k) = P(X \geq n)$$

for $k = 0, 1, 2, \ldots$

Memoryless Property

1° $P(X\geq n+k|X\geq k)=\frac{P(X\geq n+k)}{P(X\geq k)}=P(X\geq n) \Rightarrow P(X\geq n+k)=P(X\geq n)P(X\geq k)$.

2° $k=0$; $P(X\geq n)=P(X\geq n)P(X\geq 0) \Rightarrow P(X\geq 0)=1$.

3° $G(n)=P(X\geq n)$; $G(0)=1$.

Theorem

Suppose for any positive integer n, discrete random variable X satisfies

$$P(X \geq n+k | X \geq k) = P(X \geq n)$$

non-negative integer,

for $k = 0, 1, 2, \ldots,$ *then* $X \sim \mathrm{Geom}(p)$.

$G(1)=P(X\geq 1)\leq 1$.

Let $0\leq q=G(1)\leq 1$

4° $G(n+k)=G(n)\cdot G(k)$, $n=k=1$. $G(2)=G^2(1)=q^2$.

$G(n)=q^n \Rightarrow P(X\geq n)=q^n$.

$P(X=k)=P(X\geq k-1)-P(X\geq k)$

$=q^{k-1}-q^k=q^k(1-q)$

$\Rightarrow P(X=0)=1-P(X\geq 1)=1-q$.

Memoryless Property

Theorem

Geometric distribution is the one and the only one discrete distribution that is memoryless.

First Success Distribution

$X \sim Geom(p)$ $\quad P(X=k) = q^k \cdot p$.

$Y \sim FS(p)$

$Y = X+1$. $\quad P(Y=k) = q^{k-1} \cdot p$

Definition

In a sequence of independent Bernoulli trials with success probability p, let Y be the number of trials until the first successful trial, including the success. Then Y has the First Success distribution with parameter p; we denote this by $Y \sim \mathrm{FS}(p)$.

Example: Geometric & First Success Expectation

1° $E(X) = \frac{1-p}{p}$

2° $E(Y) = E(X+1) = 1 + E(X) = 1/p$.

Let $X \sim \mathrm{Geom}(p)$ and $Y \sim \mathrm{FS}(p)$, find $E(X)$ and $E(Y)$.

Story: Negative Binomial Distribution

$P(X=n)$: n failures before the r-th success

$\binom{n+r-1}{n}(1-p)^n \cdot p^r$

$(n+r)$ trails.

the $(n+r)$-th trail is successful.

In a sequence of independent Bernoulli trials with success probability p, if X is the number of failures before the r^{th} success, then X is said to have the Negative Binomial distribution with parameters r and p, denoted $X \sim \mathrm{NBin}(r, p)$.

$(n+r-1)$ trails, $(r-1)$ success

n failures.

$\binom{n+r-1}{n}(1-p)^n \cdot p^{r-1}$

Negative Binomial PMF

① $\binom{-r}{n} = \frac{(-r)!}{n!(-r-n)!} = \frac{(-r-n+1)(-r-n+2)\cdots(-r)}{n!} = \frac{(-r)(-r-1)\cdots(-r-n+1)}{n!}$

② $\binom{n+r-1}{r-1} = \frac{(n+r-1)!}{n!(r-1)!} = \frac{r(r+1)\cdots(n+r-1)}{n!} = (-1)^n \cdot \binom{-r}{n}$

Theorem

If $X \sim \mathrm{NBin}(r,p)$, then the PMF of X is

$$P(X=n) = \binom{n+r-1}{r-1} p^r q^n$$

Valid PMF

for $n = 0, 1, 2, \ldots$, where $q = 1-p$.

$\sum_{n=0}^{\infty} P(X=n) = 1$

$p^{-r} = (1-q)^{-r} = \sum_{n=0}^{\infty} \binom{-r}{n} (-q)^n \, 1^{-r-n} = \sum_{n=0}^{\infty} \binom{-r}{n} (-q)^n$

$= \sum_{n=0}^{\infty} \binom{n+r-1}{r-1} q^n$

$1 = \sum_{n=0}^{\infty} \binom{n+r-1}{r-1} q^n p^r = \sum_{n=0}^{\infty} P(X=n)$.

Geometric & Negative Binomial

$X_1 \sim Geom(p)$ X_1: # of failures before the 1st success. memoryless

$X_2 \sim Geom(p)$ X_2: # . . . . . between the 1st and 2nd success.

$X_3 \sim Geom(p)$ X_3 = # . . . . 2nd and 3rd . . .

Theorem

Let $X \sim \mathrm{NBin}(r, p)$, viewed as the number of failures before the r^{th} success in a sequence of independent Bernoulli trials with success probability p. Then we can write $X = X_1 + \cdots + X_r$ where the X_i are i.i.d. $\mathrm{Geom}(p)$.

$X_i \sim Geom(p)$ X_i: . . . . . $(i-1)$th . . . ith . . . .

Example: Expectation

Method 1: $E(X) = \sum_{n=0}^{\infty} n \cdot P(X=n) = \sum_{n=0}^{\infty} n \cdot \binom{n+r-1}{r-1} \cdot p^r \cdot q^n$.

Method 2: $X = X_1 + X_2 + \cdots + X_r$. $X_i \sim \text{Geom}(p)$.

Let $X \sim \mathrm{NBin}(r, p)$, find $E(X)$.

$$E(X) = \underbrace{E(X_1)}_{\frac{1-p}{p}} + E(X_2) + \cdots + E(X_r) = r \cdot \frac{1-p}{p}.$$

Example: 小浣熊干脆面与水浒英雄卡

Example: 小浣熊干脆面与水浒英雄卡

Example: 小浣熊干脆面与水浒英雄卡

Example: 小浣熊干脆面与水浒英雄卡

为了收集齐108张水浒英雄卡，平均而言你需要购买多少包小浣熊方便面？

Example: 盲盒收集

Example: 盲盒收集

Model: Coupon Collector

Suppose there are n types of toys, which you are collecting one by one, with the goal of getting a complete set. When collecting toys, the toy types are random (as is sometimes the case, for example, with toys included in cereal boxes or included with kids' meals from a fast food restaurant). Assume that each time you collect a toy, it is equally likely to be any of the n types. Let N denote the number of toys needed until you have a complete set. Find $E(N)$ and $Var(N)$.

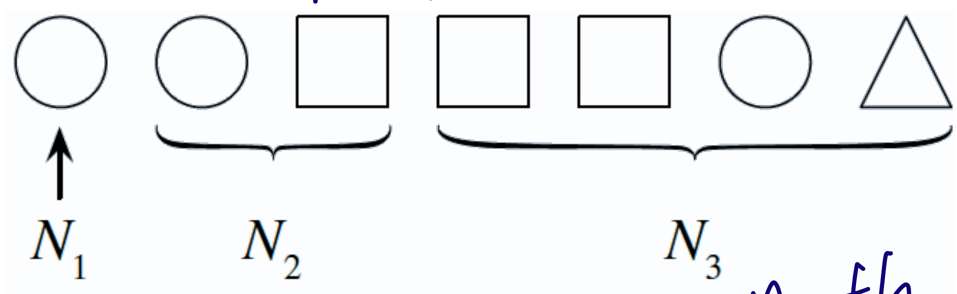
Solution: Coupon Collector

1° N: # of toys needed to obtain all types of toys.

$N = \underline{N_1 + N_2 + \cdots} + N_n$. Memoryless.

N_1: # of toys until the first new toy type appeared. $N_1 = 1$.

N_2: additional # of toys until the second new toy type appeared.



N_n: . . . . n-th . . . . .

2° $N_2 \sim FS(1-\frac{1}{n})$. $N_3 \sim FS(1-\frac{2}{n})$, $N_j \sim \underline{FS(1-\frac{j-1}{n})}$.

3° $E(N) = E(N_1) + E(N_2) + \cdots + E(N_n)$. $FS(\frac{n-j+1}{n})$

Solution: Coupon Collector

$$E(N) = E(N_1) + E(N_2) + \cdots + E(N_n)$$

$$= 1 + \frac{n}{n-1} + \cdots + n.$$

$$= n\left(\frac{1}{n} + \frac{1}{n-1} + \cdots + 1\right). = n \cdot \sum_{j=1}^{n} \frac{1}{j}. \qquad n \gg 1$$

$$\approx n \cdot [\ln n + 0.577).$$

$$n = 108. \qquad \approx 567.98.$$

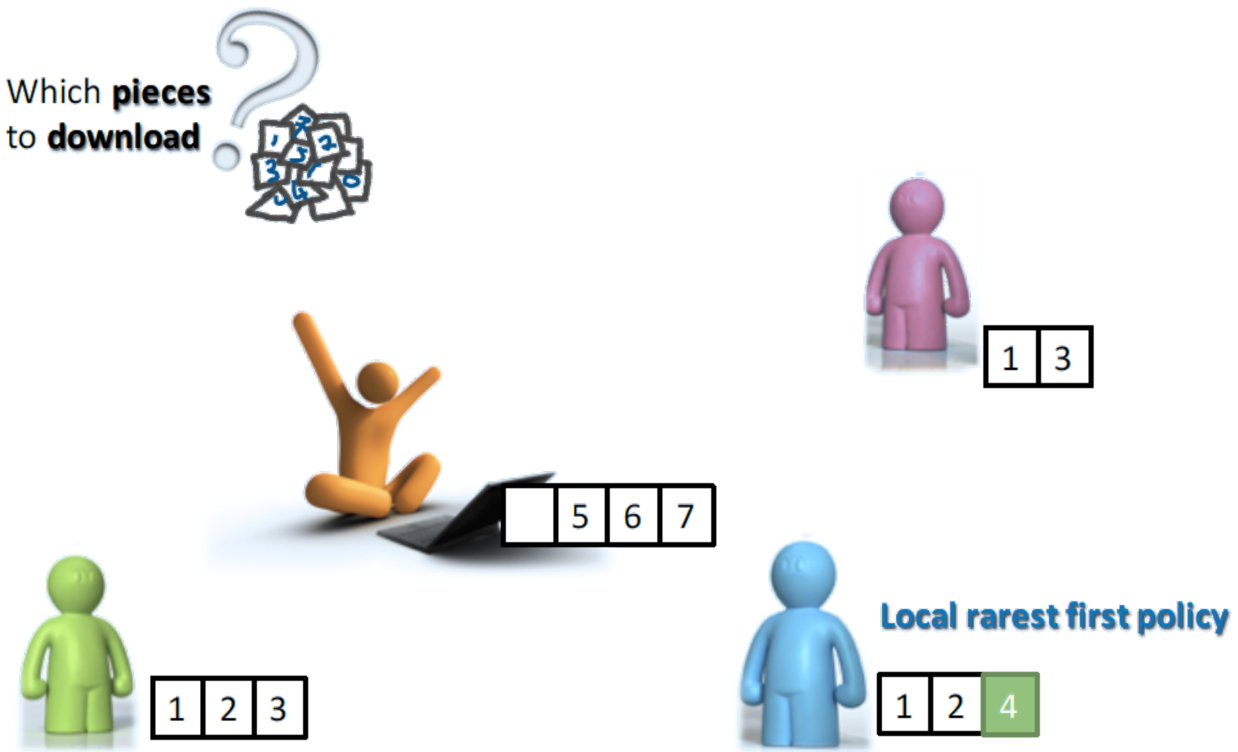
4° $Var(N) = Var(N_1) + Var(N_2) + \cdots + Var(N_n).$

Application: Peer-to-Peer System

- Target file is decomposed into n pieces.
- Each peer randomly downloads pieces and uploads pieces from its neighbors.
- $\Theta(n \ln n)$ downloads to complete the downloading file.
- The last block problem: missing the last piece (stop at 99% downloading progress).

Application: Peer-to-Peer System

- Solution adopted by BitTorrent:
  - tries to download a block that is least replicated among its neighbors.
  - maximize the diversity of content in the system, i.e., make the number of replicas of each block as equal as possible.



Outline

Properties of Indicator R.V.

$A^c = \bar{A}$.

$I_A = \begin{cases} 1, & \text{if event } A \text{ occurs.} \\ 0, & \text{otherwise.} \end{cases}$

Let A and B be events. Then the following properties hold.

1. $(I_A)^k = I_A$ for any positive integer k. $1^k = 1$, $0^k = 0$.
2. $I_{A^c} = 1 - I_A$.
3. $I_{A \cap B} = I_A I_B$.
4. $I_{A \cup B} = I_A + I_B - I_A I_B$.

③ $I_{A \cap B} = 1 \Rightarrow A \cap B$ occurs. Both A and B occurs.

$I_A = 1$, $I_B = 1$

$I_{A \cap B} = 0$, $A \cap B$ not occurs

$I_A = 0$, or $I_B = 0$.

$$I_{A \cup B} \overset{②}{=} 1 - I_{\overline{A \cup B}} = 1 - I_{A^c \cap B^c} \overset{③}{=} 1 - I_{A^c} I_{B^c} \overset{②}{=} 1 - (1 - I_A)(1 - I_B) = I_A + I_B - I_A I_B$$

Fundamental Bridge Between Probability and Expectation

$$I_A = \begin{cases} 1, & \text{if event } A \text{ occurs}, \quad P(A). \\ 0, & \text{otherwise}, \quad 1-P(A). \end{cases}$$

$$E(I_A) = 1 \cdot P(A) + 0 \cdot (1 - P(A)) = P(A).$$

Theorem

There is a one-to-one correspondence between events and indicator r.v.s, and the probability of an event A is the expected value of its indicator r.v. I_A:

$$P(\underline{A}) = \underline{E(I_A)}.$$

union

set operation

low complexity

Example: Booler's Inequality

<1> $I(A_1 \cup A_2 \cup \cdots \cup A_n) \le I(A_1) + I(A_2) + \cdots + I(A_n)$. ✓

1° if LHS $= 0$; RHS ≥ 0. ✓

2° if LHS $= 1$, $\Rightarrow$ at least one A_j occurs. $\exists j$, $I(A_j) = 1$.

$\Rightarrow$ RHS ≥ 1.

For any n events $A_1, A_2, \ldots, A_n$,

$$P(\bigcup_{i=1}^{n} A_i) \le \sum_{i=1}^{n} P(A_i).$$

<2> Take expectation for both sides.

Solution: Booler's Inequality

Example: Inclusion-Exclusion Formula

1° $I(A_1 \cup A_2 \cup \cdots \cup A_n) = 1 - I(\overline{A_1 \cup A_2 \cup \cdots \cup A_n})$

$I(\overline{A_1 \cup A_2 \cup \cdots \cup A_n}) = I(\bar{A}_1 \cap \bar{A}_2 \cap \cdots \cap \bar{A}_n) = I(A_1^c) I(A_2^c) \cdots I(A_n^c)$

$= [1 - I(A_1)][1 - I(A_2)] \cdots [1 - I(A_n)]$

$= 1 - \sum_i I(A_i) + \sum_{i<j} I(A_i) I(A_j) + \cdots + (-1)^n I(A_1) \cdots I(A_n)$

For any events $A_1, \ldots, A_n$,

$$P(\bigcup_{i=1}^{n} A_i) = \sum_i P(A_i) - \sum_{i<j} P(A_i \cap A_j) + \sum_{i<j<k} P(A_i \cap A_j \cap A_k)$$
$$- \cdots + (-1)^{n+1} P(A_1 \cap \cdots \cap A_n).$$

2° $I(A_1 \cup A_2 \cup \cdots \cup A_n) = \sum_i I(A_i) - \sum_{i<j} I(A_i) I(A_j) + \cdots + (-1)^{n+1} I(A_1) \cdots I(A_n)$

3° Take expectations for both sides.

Solution: Inclusion-Exclusion Formula

Outline

Moments of Indicator Methods

- Given n events $A_1, \ldots, A_n$ and indicators I_j, $j = 1, \ldots, n$.
- $X = \sum_{j=1}^{n} I_j$: the number of events that occur.
- $\binom{X}{2} = \sum_{i<j} I_i I_j$: the number of pairs of distinct events that occur.

$\frac{X(X-1)}{2}$

$I_i I_j = I(A_i) I(A_j) = I(A_i \cap A_j)$.

- $E(\binom{X}{2}) = \sum_{i<j} P(A_i \cap A_j)$. $\Rightarrow E\left[\frac{X(X-1)}{2}\right] = \sum_{i<j} P(A_i \cap A_j)$.
  - $E(X^2) = 2\sum_{i<j} P(A_i \cap A_j) + E(X)$.
  - $Var(X) = 2\sum_{i<j} P(A_i \cap A_j) + E(X) - (E(X))^2$.

$Var(X) = E(X^2) - (EX)^2$

Moments of Binomial Random Variables

$X \sim Bin(n,p)$;

1° consider n independent Bernoulli trails, each with success prop. p.

event A_i: the i-th trail is a success. $P(A_i) = p$. $\forall 1 \le i \le n$.

$$I_j = I(A_j) \sim Bern(p). \quad \forall 1 \le j \le n.$$

2° # of successful trails. $X = \sum_{j=1}^{n} I_j$.

① $E(X) = E\left(\sum_{j=1}^{n} I_j\right) = np$.

② $E\left[\binom{X}{2}\right] = \sum_{i<j} P(A_i \cap A_j) = \sum_{i<j} P(A_i)\, P(A_j) = \binom{n}{2} \cdot p^2 = \frac{n(n-1)}{2} p^2$

$E(X^2) = 2\sum_{i<j} P(A_i \cap A_j) + E(X) = n(n-1)p^2 + np$.

$Var(X) = E(X^2) - (EX)^2 = np(1-p)$

③ $E\left[\binom{X}{k}\right] = \binom{n}{k} p^k$.

Outline

Poisson Distribution

e^{λ} Taylor expansion

① Valid PMF $\sum_{k=0}^{\infty} P(X=k) = \sum_{k=0}^{\infty} \frac{e^{-\lambda}\lambda^k}{k!} = e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k}{k!}$

$= 1.$

Definition

An r.v. X has the *Poisson distribution* with parameter λ if the PMF of X is

$$P(X = k) = \frac{e^{-\lambda}\lambda^k}{k!}, k = 0, 1, 2, \ldots$$

We write this as $X \sim \mathrm{Pois}(\lambda)$.

$E(X) = Var(X) = \lambda.$

Example: Poisson Expectation & Variance

① $E(X) = \sum_{k=0}^{\infty} k \cdot P(X=k) = \sum_{k=0}^{\infty} k \cdot \frac{e^{-\lambda} \cdot \lambda^k}{k!} = \sum_{k=1}^{\infty} \frac{e^{-\lambda} \cdot \lambda^{k-1} \cdot \lambda}{(k-1)!} = \lambda \cdot \underbrace{\sum_{k=1}^{\infty} \frac{e^{-\lambda} \cdot \lambda^{k-1}}{(k-1)!}}_{1}$

$= \lambda.$

$Var(X) = E(X^2) - (EX)^2.$

② $E(X^2) = \sum_{k=0}^{\infty} k^2 \cdot P(X=k) = \sum_{k=1}^{\infty} \frac{k^2 \cdot e^{-\lambda} \lambda^k}{k!} = e^{-\lambda} \cdot \lambda + \sum_{k=2}^{\infty} \frac{k e^{-\lambda} \cdot \lambda^k}{(k-1)!}$

$= e^{-\lambda} \cdot \lambda + \underline{\sum_{k=2}^{\infty} \frac{e^{-\lambda} \cdot \lambda^k}{(k-2)!}} + \underline{\sum_{k=2}^{\infty} \frac{e^{-\lambda} \lambda^{k-1}}{(k-1)!} \cdot \lambda}$

$= e^{-\lambda} \cdot \lambda + \lambda^2 + (1 - e^{-\lambda}) \cdot \lambda = \lambda^2 + \lambda.$

③ $Var(X) = E(X^2) - (EX)^2 = \lambda.$

Poisson Approximation

Cramer - Lundberg model.

Let $A_1, A_2, \ldots, A_n$ be events with $p_j = P(A_j)$, where n is large, the p_j are small, and the A_j are independent or ~~weakly dependent~~. Let

$$X = \sum_{j=1}^{n} I(A_j)$$

count how many of the A_j occur. Then X is approximately $\mathrm{Pois}(\lambda)$, with $\lambda = \sum_{j=1}^{n} p_j$.

Example: Birthday Problem Revisited

1° m people; $\binom{m}{2}$ pairs, $j = 1, 2, \cdots, \binom{m}{2}$.

Prob (each pair of people has the same birthday) $= \frac{1}{365}$.

2° A_j: "j-th pair of people has the same birthday". $P(A_j) = \frac{1}{365} = p$.

$I_j = I(A_j)$; $n = \binom{m}{2}$; $X = \sum_j I_j$. # of birthday-match pairs.

3° Poisson approximation: $X \sim \text{Pois}(\lambda)$. $\lambda = np = \binom{m}{2} p$.

4° Prob. (at least one birthday match) $= P(X \geq 1) = 1 - P(X = 0) = 1 - e^{-\lambda}$

$m = 23$, $\lambda = \binom{23}{2} \cdot \frac{1}{365} = \frac{253}{365}$ $\quad P(X \geq 1) = 1 - e^{-\lambda} \approx 0.5002$

Poisson & Binomial

- Poisson $\Longrightarrow$ Binomial: **conditioning**
- Binomial $\Longrightarrow$ Poisson: **taking a limit**

Sum of Independent Poissons

$$P(X+Y=k) \overset{\text{LOTP}}{=} \sum_{j=0}^{k} P(X+Y=k \mid X=j)\, P(X=j)$$

$$= \sum_{j=0}^{k} \underline{P(Y=k-j \mid X=j)}\, P(X=j)$$

$$= \sum_{j=0}^{k} P(Y=k-j)\, P(X=j) = \sum_{j=0}^{k} \frac{e^{-\lambda_2} \cdot \lambda_2^{k-j}}{(k-j)!} \cdot \frac{e^{-\lambda_1} \cdot \lambda_1^{j}}{j!}$$

Theorem

If $X \sim \mathrm{Pois}(\lambda_1)$, $Y \sim \mathrm{Pois}(\lambda_2)$, *and* X *is independent of* Y, *then* $X + Y \sim \mathrm{Pois}(\lambda_1 + \lambda_2)$.

$$= \frac{e^{-(\lambda_1+\lambda_2)}}{k!} \cdot \sum_{j=0}^{k} \underline{\frac{k!}{(k-j)!\cdot j!} \cdot \lambda_1^{j} \cdot \lambda_2^{k-j}}$$

$$\binom{k}{j} \cdot \lambda_1^{j} \lambda_2^{k-j}$$

$$= \frac{e^{-(\lambda_1+\lambda_2)} \cdot (\lambda_1+\lambda_2)^{k}}{k!}$$

Poisson Given A Sum of Poissons

$$P(X=k \mid X+Y=n) = \frac{P(X=k, X+Y=n)}{P(X+Y=n)} = \frac{P(X=k, Y=n-k)}{P(X+Y=n)}$$

$$= \frac{P(X=k)P(Y=n-k)}{P(X+Y=n)}.$$

Theorem

If $X \sim \mathrm{Pois}(\lambda_1)$, $Y \sim \mathrm{Pois}(\lambda_2)$, and X is independent of Y, then the conditional distribution of X given $X+Y=n$ is $\mathrm{Bin}(n, \lambda_1/(\lambda_1+\lambda_2))$.

$$p = \frac{\lambda_1}{\lambda_1+\lambda_2}$$

$$= \frac{\frac{\cancel{e^{-\lambda_1}} \cdot \lambda_1^k}{k!} \cdot \frac{\cancel{e^{-\lambda_2}} \cdot \lambda_2^{n-k}}{(n-k)!}}{\frac{\cancel{e^{-(\lambda_1+\lambda_2)}} \cdot (\lambda_1+\lambda_2)^n}{n!}} = \frac{n!}{(n-k)!k!} \cdot \left(\frac{\lambda_1}{\lambda_1+\lambda_2}\right)^k \left(\frac{\lambda_2}{\lambda_1+\lambda_2}\right)^{n-k}$$

$$\binom{n}{k} \cdot p^k (1-p)^{n-k}$$

Poisson Approximation to Binomial

1° $\lambda = np$. Given k, $(0 \le k \le n)$. $X \sim \mathrm{Bin}(n, p)$

$$P(X=k) = \binom{n}{k} \cdot p^k (1-p)^{n-k} = \frac{n!}{k!(n-k)!} p^k (1-p)^k, \quad p = \lambda/n.$$

Theorem

If $X \sim \mathrm{Bin}(n, p)$ and we let $n \to \infty$ and $p \to 0$ such that $\lambda = np$ remains fixed, then the PMF of X converges to the $\mathrm{Pois}(\lambda)$ PMF. More generally, the same conclusion holds if $n \to \infty$ and $p \to 0$ in such a way that np converges to a constant λ. λ^k/n^k

$$= \frac{n(n-1)\cdots(n-k+1)}{k!} \cdot \left(\frac{\lambda}{n}\right)^k \cdot \left(1-\frac{\lambda}{n}\right)^{n-k} = 1 \cdot \left(1-\frac{1}{n}\right) \cdots \left(1-\frac{k-1}{n}\right) \frac{\lambda^k}{k!} \left(1-\frac{\lambda}{n}\right)^{n-k}$$

$$= \frac{\lambda^k}{k!} \cdot \underbrace{\left[1 \cdot \left(1-\frac{1}{n}\right) \cdots \left(1-\frac{k-1}{n}\right)\right]}_{1} \cdot \underbrace{\left(1-\frac{\lambda}{n}\right)^n}_{e^{-\lambda}} \cdot \underbrace{\left(1-\frac{\lambda}{n}\right)^{-k}}_{1} \xrightarrow{n \to \infty} \frac{e^{-\lambda} \cdot \lambda^k}{k!}$$

$n \to \infty$ $\frac{\lambda^k}{k!}$ $\sim \mathrm{Pois}(\lambda)$

Proof

Visitors to A Website

$n=10^6$, $p=2\times10^{-6}$, $\lambda=np=2$.

$P(Y=k)=\frac{e^{-\lambda}\cdot\lambda^k}{k!}$

Y: # of visitors. $Y\sim Pois(\lambda)$

The owner of a certain website is studying the distribution of the number of visitors to the site. Every day, a million people independently decide whether to visit the site, with probability $p = 2 \times 10^{-6}$ of visiting. Give a good approximation for the probability of getting *at least three* visitors on a particular day.

$P(Y\geq3) = 1-P(Y=0)-P(Y=1)-P(Y=2)$

$= 1-5e^{-2} \doteq 0.3233$

Outline

Typical Distance Measures

- Total Variation Distance
- Kullback–Leibler Divergence
- Jensen–Shannon Divergence
- Bhattacharyya Distance
- Wasserstein Distance (or called "Kantorovich–Rubinstein")

Total Variation Distance

- Distance measure between two probability distributions.
- Apply such measure to characterize the accuracy of Poisson approximation.

Definition

The total variation distance between two distributions μ and ν on a countable set Ω is

$$\begin{aligned} d_{TV}(\mu,\nu) &= \|\mu-\nu\|_{TV} \\ &= \frac{1}{2}\sum_{x\in\Omega}|\mu(x)-\nu(x)|. \end{aligned}$$

Example

① $\mu(1) = p,\ \mu(0) = 1-p.\quad \mu(n) = 0.\ \forall n \geq 2.$ $\quad = 1 - \nu(0) - \nu(1)$

$\nu(n) = \frac{e^{-p} \cdot p^n}{n!},\quad n \geq 0.$ $\quad \sum_{n=2}^{\infty} \nu(n)$

② $2 d_{TV}(\mu, \nu) = \sum_{n \in \Omega} |\mu(n) - \nu(n)| = |\mu(0) - \nu(0)| + |\mu(1) - \nu(1)| + \sum_{n=2}^{\infty} |\mu(n)|$

Let μ be the distribution with $\mu(1) = p$ and $\mu(0) = 1 - p$. Let ν be a Poisson distribution with mean p. Then we have $d_{TV}(\mu, \nu) \leq p^2$.

$$= |1-p-e^{-p}| + |p - e^{-p}p| + (1 - e^{-p} - e^{-p} \cdot p)$$

$$= e^{-p} + p - 1 + p(1-e^{-p}) + 1 - e^{-p} - e^{-p} \cdot p$$

$$= 2p(1-e^{-p}) \leq 2p^2$$

③ $d_{TV}(\mu, \nu) \leq p^2$

The Law of Small Numbers

The Law of Rare Events

Theorem

Given independent random variables $Y_1, \ldots, Y_n$ such that for any $1 \leq m \leq n$, $P(Y_m = 1) = p_m$ and $P(Y_m = 0) = 1 - p_m$. Let $S_n = Y_1 + \cdots + Y_n$. Suppose

$$\sum_{m=1}^{n} p_m \to \lambda \in (0, \infty) \quad \text{as } n \to \infty,$$

and

$$\max_{1 \leq m \leq n} p_m \to 0 \quad \text{as } n \to \infty,$$

then

$$d_{TV}(S_n, \mathrm{Pois}(\lambda)) \to 0 \quad \text{as } n \to \infty.$$

Gap of Poisson Approximation

- A bound on the gap due to Hodges and Le Cam (1960):

$$d_{TV}(S_n, \mathrm{Pois}(\lambda)) \leq \sum_{m=1}^{n} p_m^2,$$

- by Stein-Chen method (C.Stein 1987) we can have a tighter bound on the gap:

$$d_{TV}(S_n, \mathrm{Pois}(\lambda)) \leq \min(1, \frac{1}{\lambda}) \sum_{m=1}^{n} p_m^2.$$

Outline

Probability Generating Function

Definition

The *probability generating function* (PGF) of a nonnegative integer-valued r.v. X with PMF $p_k = P(X = k)$ is the generating function of the PMF. By LOTUS, this is

$$= \sum_{k=0}^{\infty} P(X=k) \cdot t^k$$

$$f(t) = E(t^X) = \sum_{k=0}^{\infty} p_k t^k. \quad = \sum_{k=0}^{\infty} p_k t^k$$

The PGF converges to a value in $[-1, 1]$ for all t in $[-1, 1]$ since $\sum_{k=0}^{\infty} p_k = 1$ and $|p_k t^k| \leq p_k$ for $|t| \leq 1$.

Example: Generating Dice Probabilities

① $E[t^X] = \sum_{k=0}^{\infty} P(X=k)t^k = \sum_{k=0}^{\infty} p_k t^k$. is polynomial function of t,

the coefficient of t^{18}: $P(X=18)$.

t^{19}: $P(X=19)$

t^{20}: $P(X=20)$

② $E[XY] = E[X]E[Y]$
if X, Y independent.

$E[XY] = \sum_x \sum_y xy\, P(X=x)P(Y=y) = \sum_x x P(X=x) \cdot \sum_y y \cdot P(Y=y)$
$= E[X]E[Y]$

Let X be the total from rolling 6 fair dice, and let $X_1, \ldots, X_6$ be the individual rolls. What is $P(X = 18)$?

$X = X_1 + X_2 + X_3 + X_4 + X_5 + X_6$.

$6 \le X \le 36$.

③ $E[t^X] = E[t^{X_1}] \cdot E[t^{X_2}] \cdots E[t^{X_6}]$.
$= (E[t^{X_1}])^6$

$E[t^{X_1}] = \sum_{j=1}^{6} P(X_1 = j) \cdot t^j = \frac{1}{6}(t + t^2 + \cdots + t^6)$.

④ $E[t^X] = (E[t^{X_1}])^6 = \frac{t^6}{6^6}(1 + t + \cdots + t^5)^6$

$P(X=18)$
$= \frac{3431}{6^6}$

Solution

PGF and Moments

① $g(t) = \sum_{k=0}^{\infty} p_k t^k = p_0 + \sum_{k=1}^{\infty} p_k t^k$; $g'(t) = p_1 + \sum_{k=2}^{\infty} p_k t^{k-1} \cdot k$

$g'(t)|_{t=1} = \sum_{k=1}^{\infty} p_k \cdot k + 0 \cdot p_0 = E(X)$.

Let X be a nonnegative integer-valued r.v. with PMF $p_k = P(X = k)$, and the PGF of X is $g(t) = \sum_{k=0}^{\infty} p_k t^k$, we have

- $E(X) = g'(t)|_{t=1}$, ✓
- $E(X(X-1)) = g''(t)|_{t=1}$. $E[X^2] - E[X]$. $Var(X) = E[X^2] - (EX)^2$

② $g''(t) = \sum_{k=2}^{\infty} k(k-1) \cdot p_k \cdot t^{k-2}$. $g''(t)|_{t=1} = \sum_{k=2}^{\infty} k(k-1) \cdot p_k + 0 \cdot p_1 + 0 \cdot p_0 = E[X(X-1)]$.

③ $g'''(t)|_{t=1} =$

PGF and Moments

1° $g(t)=\sum_{k=0}^{\infty} p_k t^k = p_0 + \sum_{k=1}^{\infty} p_k t^k$; $P(X=0)=g(0)=p_0$,

2° $g'(t)= p_1 + \sum_{k=2}^{\infty} p_k t^{k-1}\cdot k$; $P(X=1)=g'(0)=p_1$,

3° $g''(t) = 2p_2 + \sum_{k=3}^{\infty} k\cdot(k-1)\cdot p_k t^{k-2}$; $P(X=2)=\dfrac{g''(0)}{2}$.

4° $g^{(m)}(t) = m!\cdot p_m + \sum_{k=m+1}^{\infty} \dfrac{k!}{(k-m)!}\cdot t^{k-m}\cdot p_k$.

$$P(X=m)=\frac{g^{(m)}(0)}{m!}.$$

PGF and Moments

Binomial PMF

i.i.d.

① $X \sim Bin(n,p)$. $X = X_1 + \cdots + X_n$. $X_i \sim Bern(p)$. $1 \le i \le n$.

② $g_X(t) = E[t^X] = E[t^{X_1+X_2+\cdots+X_n}] = E[t^{X_1}] \cdots E[t^{X_n}] = (E[t^{X_1}])^n$

③ $E[t^{X_i}] = t^0 \cdot (1-p) + t^1 \cdot p = pt + q$, $q = (1-p)$.

④ $g_X(t) = (pt+q)^n = \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} t^k$. | $P(X=k) = \binom{n}{k} p^k q^{n-k}$

⑤ $P(X=0) = g_X(0) = q^n$. $P(X=1) = g_X'(0) = npq^{n-1}$

$$P(X=k) = \frac{g_X^{(k)}(0)}{k!} = \binom{n}{k} \cdot p^k q^{n-k}$$

Binomial Moments

$$g_X(t) = (pt+q)^n. \qquad p+q=1.$$

$$g'_X(t) = np(pt+q)^{n-1}, \qquad g'_X(t)\big|_{t=1} = np = E[X].$$

$$g''_X(t) = n(n-1)\cdot p^2(pt+q)^{n-2}, \qquad g''_X(t)\big|_{t=1} = n(n-1)p^2 = E[X(X-1)]$$

$$Var(X) = g''_X(t)\big|_{t=1} + g'_X(t)\big|_{t=1} - \left(g'_X(t)\big|_{t=1}\right)^2$$

$$E\left[\binom{X}{k}\right] = \binom{n}{k} p^k, \quad k \geq 2$$

Example: Pattern Matching

$H = p$. T: $q = 1-p$.

① $p_k = P(N=k)$; $p_0 = 0$; $p_1 = 0$; $p_2 = p^2$; $p_3 = (1-p)p^2 = qp^2$

$$N=4 \begin{cases} \overset{1}{H} \ \overset{2}{T} \ \overset{3}{H} \ \overset{4}{H} \\ T \ T \ H \ H \end{cases}$$

② First-step method.

Suppose a coin with probability p for heads is tossed repeatedly, and we obtain a sequence of H and T (H denotes Head and T denotes Tail). Let N denote the number of toss to observe the first occurrence of the pattern "HH". Find $E(N)$ and $Var(N)$.

$k \geq 3$, S_1: result of first toss, $S_1 = H$ or T.

$$\text{LOTP}: p_k = P(N=k) = P(N=k \mid S_1=H)\,P(S_1=H) + P(N=k \mid S_1=T)\,P(S_1=T)$$
$$= P(N=k, S_1=H) + P(N=k, S_1=T).$$

Example: Pattern Matching

S_2: result of the 2nd toss

③ $P(N=k, S_1=H) = P(S_1=H)\,P(S_2=T)\cdot P(N=k-2) = p\cdot q\,p_{k-2}$

$P(N=k, S_1=T) = P(S_1=T)\,P(N=k-1) = q\,p_{k-1}$.

④ $p_k = P(N=k, S_1=H) + P(N=k, S_1=T)$

$p_k = pq\,p_{k-2} + q\,p_{k-1}, \quad k \geq 3$

$p_0 = 0, \quad p_1 = 0, \quad p_2 = p^2,$

Method 1:

sequence: find p_k.

$E[N] = \sum_{k=0}^{\infty} k\cdot p_k$.

Method 2: PGF $\to E[N]$

Example: Pattern Matching

$$\begin{cases} p_k = pq\,p_{k-2} + q\,p_{k-1}, & k \geq 3. \\ p_0 = 0,\ p_1 = 0,\ p_2 = p^2. \end{cases}$$

⑤ PGF of N. $g(t) = E[t^N] = \sum_{k=0}^{\infty} p_k t^k = \sum_{k=1}^{\infty} p_k t^k = \sum_{k=2}^{\infty} p_k t^k$

$$= p_2 t^2 + \sum_{k=3}^{\infty} p_k t^k = p^2 t^2 + \sum_{k=3}^{\infty} p_k t^k = g(t).$$

on the other hand, $p_k = pq\,p_{k-2} + q\,p_{k-1}$, $k \geq 3$.

$$\sum_{k=3}^{\infty} p_k t^k = \sum_{k=3}^{\infty} (q\,p_{k-1} + pq\,p_{k-2}) t^k = \sum_{k=3}^{\infty} q \cdot p_{k-1} \cdot t^k + \sum_{k=3}^{\infty} pq\,p_{k-2} t^k$$

$$= qt \sum_{k=3}^{\infty} p_{k-1} t^{k-1} + pq t^2 \sum_{k=3}^{\infty} p_{k-2} t^{k-2}$$

$$= qt \sum_{k=2}^{\infty} p_k t^k + pq t^2 \sum_{k=1}^{\infty} p_k t^k$$

$$= qt\,g(t) + pqt^2 g(t) = (qt + pqt^2)\,g(t)$$

Example: Pattern Matching

$$\Rightarrow g(t) = \frac{p^2 t^2}{1 - qt - pqt^2} \qquad \text{PGF of } N.$$

$$E[N] = g'(t)\big|_{t=1} = g'(1) = \frac{1}{p} + \frac{1}{p^2}$$

$$Var[N] = g''(1) + g'(1) - [g'(1)]^2 = \frac{1 - p^5 - 5pq^2}{q^2 p^4}$$

"HHH"

Example: Pattern Matching

Outline

Probability Method

- Paul Erdös initiated this method: Erdös Method.
- Widely used in information theory & combinatorics & theoretical computer science.
- Main idea: to prove the existence of a structure with certain properties using probability or expectation.

Principle I

- First we construct an appropriate probability space of structures.
- Then we show that a randomly chosen element in this space has the desired properties with positive probability.

Theorem (The Possibility Principle)

Let A be the event that a randomly chosen object in a collection has a certain property. If $P(A) > 0$, then there exists an object with such property.

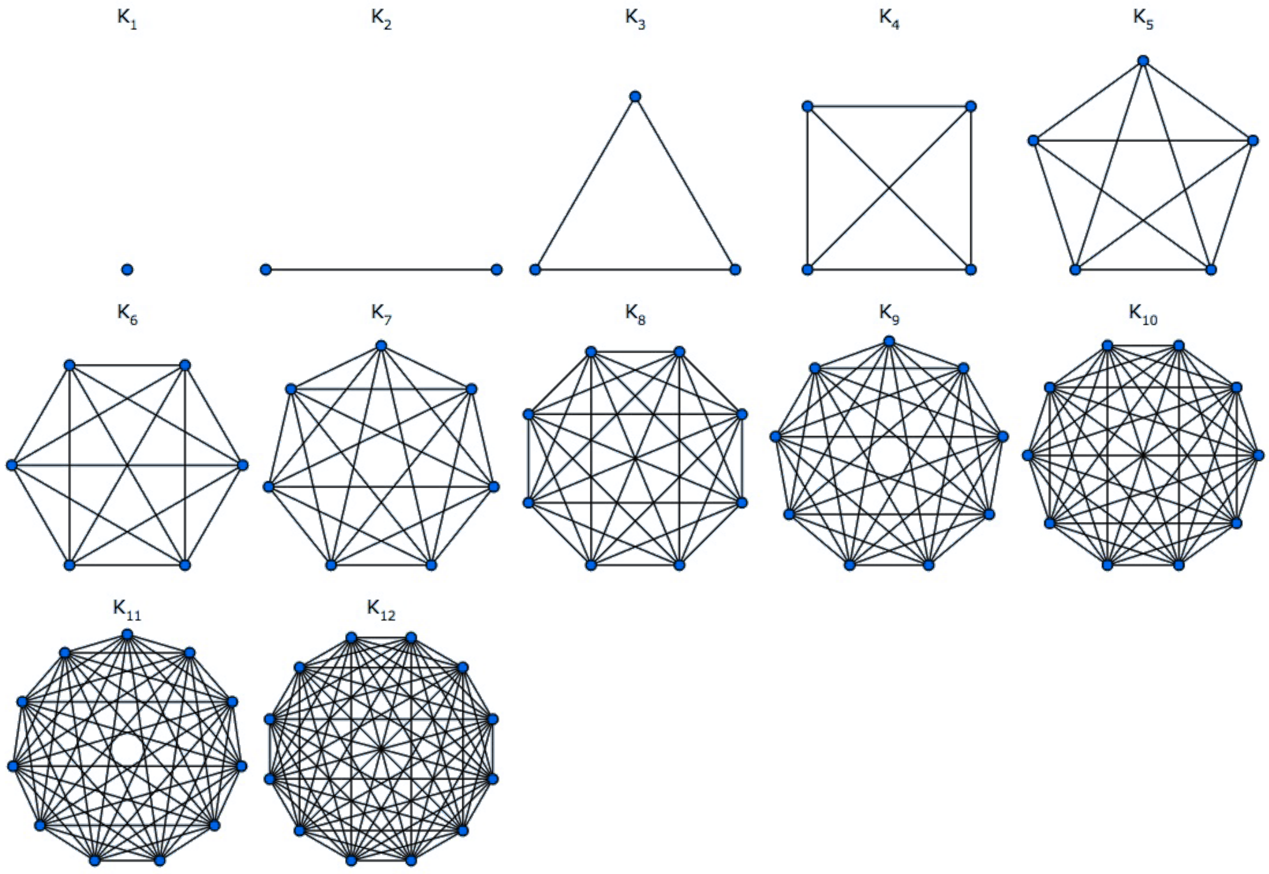
Principle II

Theorem (The Good Score Principle)

Let X be the score of a randomly chosen object. If $E(X) \geq c$, then there exists an object with a score of at least c.

Example: Graph Coloring

- Complete graph (clique): a simple undirected graph in which every pair of distinct vertices is connected by a unique edge.
- Complete graph K_n: a graph with n nodes and $\binom{n}{2}$ edges.



Example: Graph Coloring

Theorem

Given a complete graph $K_n (n \geq 3)$, if $\binom{n}{m} 2^{-\binom{m}{2}+1} < 1$, then it is possible to color the edges of K_n with two colors so that it has no monochromatic K_m subgraph $(1 < m < n)$.

Testing Polynomial Identities

- Randomized algorithms can be dramatically more efficient than their best known deterministic counterparts.
- Input two polynomials Q and R over n variables, with coefficients in some field, and decides whether $Q \equiv R$.
- Example: $Q(x_1, x_2) = (1 + x_1)(1 + x_2)$, $R(x_1, x_2) = 1 + x_1 + x_2 + x_1 x_2$.
- n-variable polynomial $\prod_{i=1}^{n}(x_i + x_{i+1})$ expands into $O(2^n)$ monomials.

The Schwartz-Zippel Algorithm

- A Monte Carlo algorithm with a bounded probability of false positive and no false negative.
- Input polynomial $M(x_1, \ldots, x_n)$ and test whether $M \equiv 0 (M = Q - R)$.
- Assign values $r_1, \ldots, r_n$ chosen independently and uniformly at random from a finite set S to $x_1, \ldots, x_n$.
- Test if $M(r_1, \ldots, r_n) = 0$, outputting "Yes" if so and "No" otherwise.
- If "No", then $M \not\equiv 0$.
- if "Yes", it is possible that $M \not\equiv 0$ but $r_1, \ldots, r_n$ happens to be a zero of M.

Schwartz-Zippel Lemma

Lemma

Let $M \in F(x_1, x_2, \ldots, x_n)$ be a non-zero polynomial of total degree $d \geq 0$ over a field F. Let S be a finite subset of F and let $r_1, r_2, \ldots, r_n$ be selected at random independently and uniformly from S. Then

$$P[M(r_1, r_2, \ldots, r_n) = 0] \leq \frac{d}{|S|}.$$

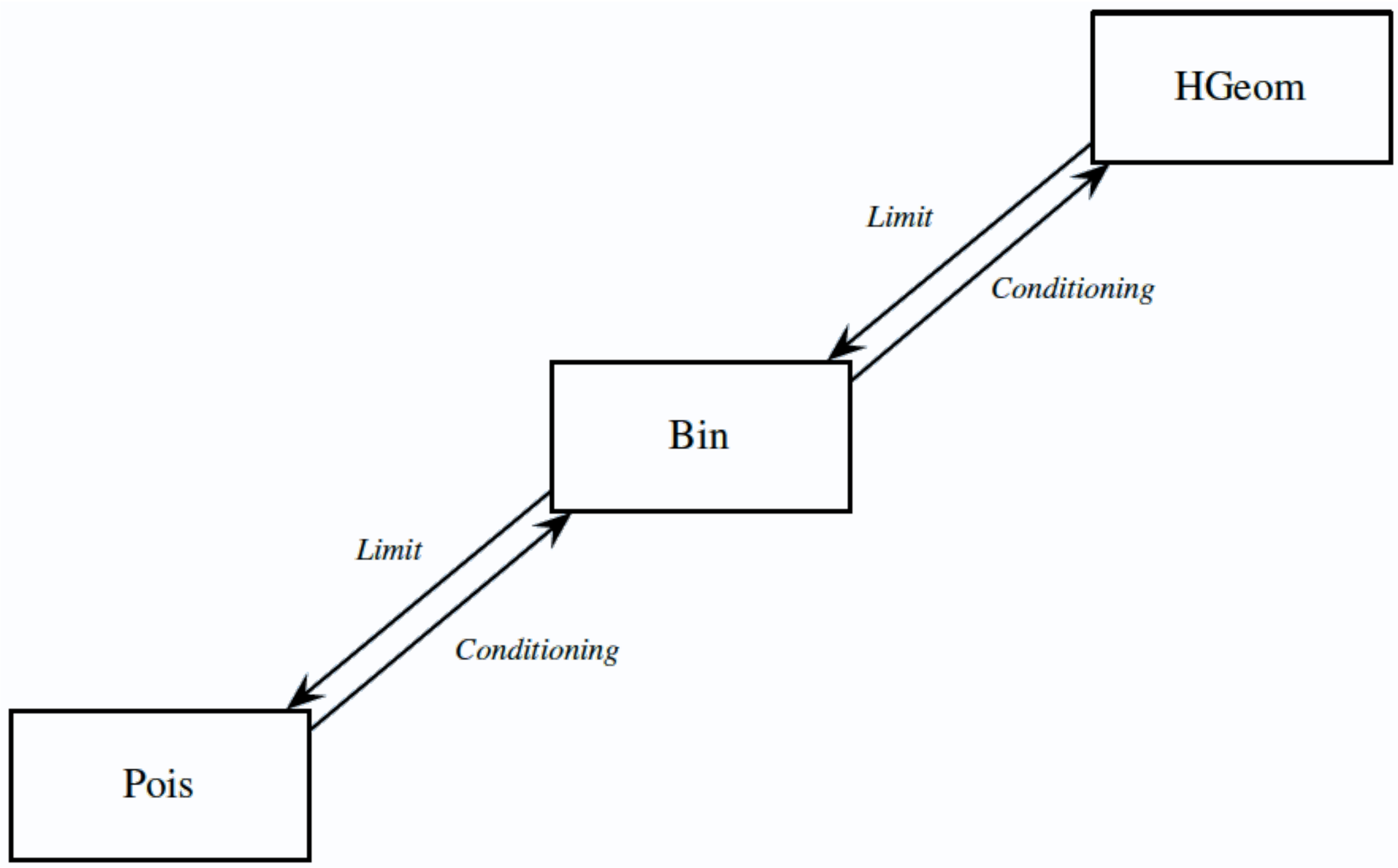
Remarks

- If we take the set S to have cardinality at least twice the degree of our polynomial ($|S| \geq 2d$), we can bound the probability of error (false positive) by $1/2$.
- This can be reduced to any desired small number by repeated trials.

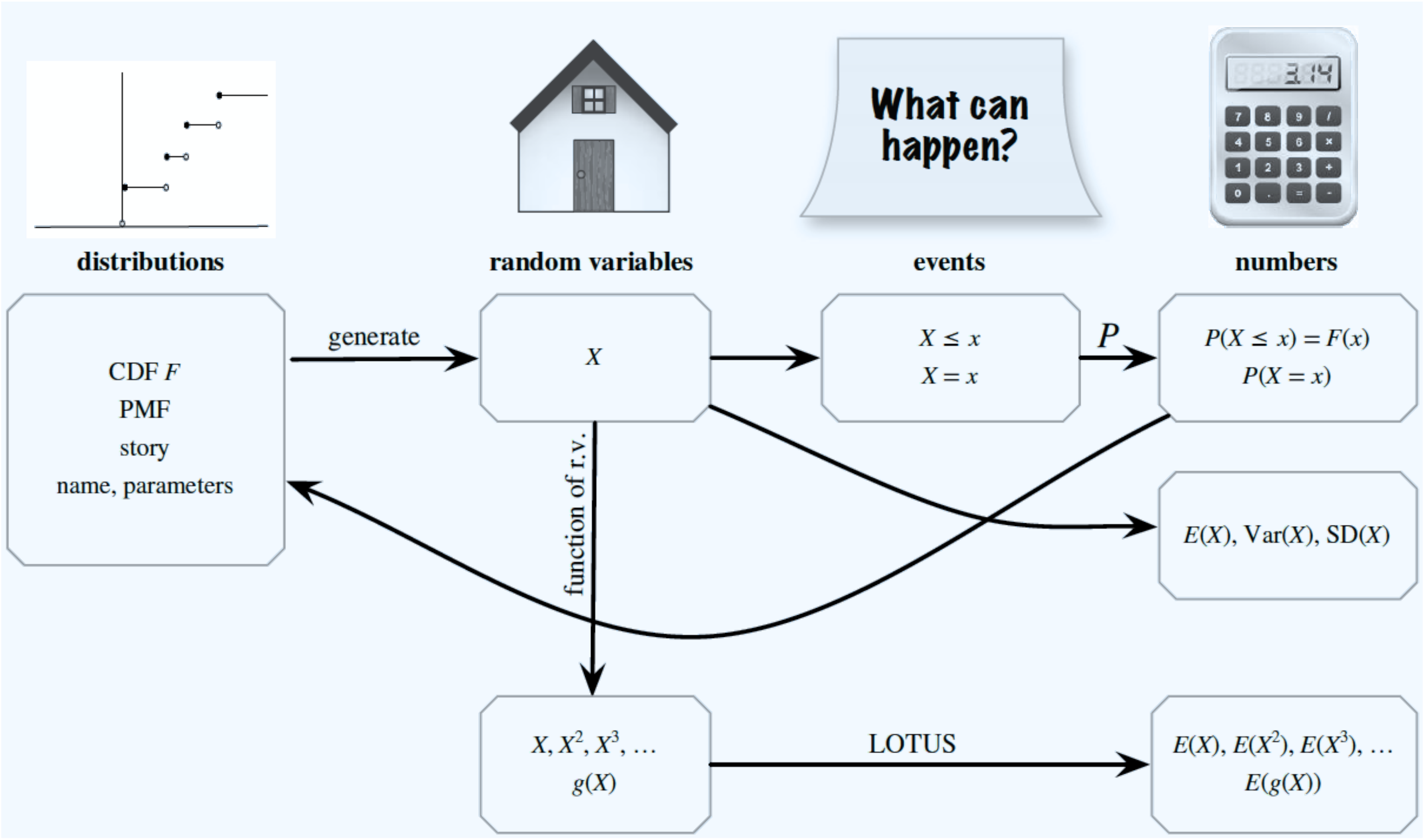
Summary 1

	With replacement	**Without replacement**
Fixed number of trials	Binomial	Hypergeometric
Fixed number of successes	Negative Binomial	Negative Hypergeometric

Summary 2



Summary 3


What can happen?
distributions
random variables
events
numbers
CDF F
PMF
story
name, parameters
generate
X
function of r.v.
$X \leq x$
$X = x$
P
$P(X \leq x) = F(x)$
$P(X = x)$
$E(X)$, Var(X), SD(X)
$X, X^2, X^3, \ldots$
$g(X)$
LOTUS
$E(X), E(X^2), E(X^3), \ldots$
$E(g(X))$

References

- Chapters 4 & 6 of **BH**
- Chapter 2 of **BT**