Lecture 4: Expectation

Ziyu Shao

School of Information Science and Technology
ShanghaiTech University

October 22, 2024

Outline

1. Expectation & Variance
2. Geometric and Negative Binomial
3. Indicator R.V.s and The Fundamental Bridge
4. Moments and Indicators
5. Poisson
6. Distance between Two Probability Distributions
7. Probability Generating Functions
8. Reading for Fun

Outline

1. Expectation & Variance
2. Geometric and Negative Binomial
3. Indicator R.V.s and The Fundamental Bridge
4. Moments and Indicators
5. Poisson
6. Distance between Two Probability Distributions
7. Probability Generating Functions
8. Reading for Fun

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 $x_1, x_2, \cdots$ 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

Theorem

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

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

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.

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

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

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

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.

Variance and Standard Deviation

Definition

The variance of an r.v. X is

$$\operatorname{Var}(X) = E(X - EX)^2 .$$

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

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

Properties of Variance

- For any r.v. X, $\mathrm{Var}(X) = E\left(X^2\right) - (EX)^2$.
- $\text{Var}(X + c) = Var(X)$ for any constant c.
- $\text{Var}(cX) = c^2 Var(X)$ for any constant c.
- If X and Y are independent, then $\text{Var}(X + Y) = Var(X) + Var(Y)$.
- $\text{Var}(X) \geq 0$ with equality if and only if $P(X = a) = 1$ for some constant a.

Properties of Variance

Outline

1 Expectation & Variance
2 Geometric and Negative Binomial
3 Indicator R.V.s and The Fundamental Bridge
4 Moments and Indicators
5 Poisson
6 Distance between Two Probability Distributions
7 Probability Generating Functions
8 Reading for Fun

Story: Geometric Distribution

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 $X \sim Geom(p)$.

Geometric PMF

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* $q = 1 - p$*.*

Memoryless Property

Theorem

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

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

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

Memoryless Property

Theorem

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

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

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

Memoryless Property

Theorem

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

First Success Distribution

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

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

Story: Negative Binomial Distribution

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 $X \sim NBin(r, p)$.

Negative Binomial PMF

Theorem

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

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

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

Geometric & Negative Binomial

Theorem

Let $X \sim \mathrm{NBin}(r, p)$, viewed as the number of failures before the rth 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. $\mathrm{Geom}(p)$.

Example: Expectation

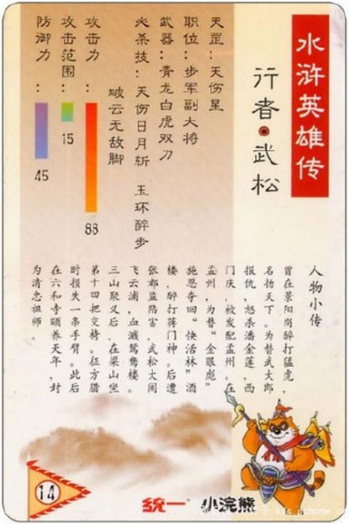
Let $X \sim NBin(r, p)$, find $E(X)$.

Example:


统一
President
羊肉串口味
小浣熊
干脆面
本包装已申请专利 伪冒必究
油炸面饼净含量：48g
总净含量：50g

Example:

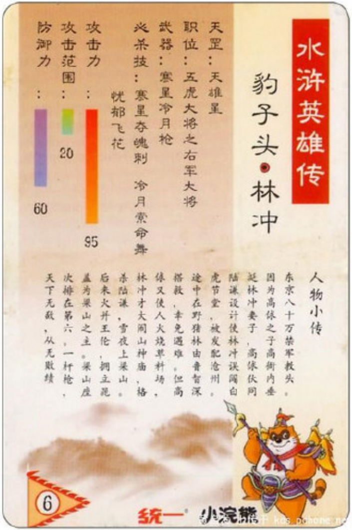
行者·武松
K



水浒英雄传
行者·武松
天罡：天伤星
职位：步军副大将
武器：青龙白虎双刀
必杀技：天伤日月斩 玉环醉步 破云无敌脚
攻击力：88
攻击范围：15
防御力：45
人物小传
曾在景阳岗醉打猛虎，名扬天下。为替武大郎报仇，怒杀潘金莲，西门庆，被发配孟州。在孟州，为替“金眼彪”施恩夺回“快活林”酒楼，醉打蒋门神。后遭张都监陷害，武松大闹飞云浦，血溅鸳鸯楼。三山聚义后，在梁山坐第十四把交椅。征方腊时损失一条手臂。此后在六和寺颐养天年，封为清忠祖师。
14
统一 小浣熊

Example:


豹子头·林冲
A ♠


水浒英雄传
豹子头·林冲
天罡：天雄星
职位：五虎大将之右军大将
武器：寒星冷月枪
必杀技：寒星夺魄刺 冷月索命舞 忧郁飞花
攻击力：95
攻击范围：20
防御力：60
人物小传
东京八十万禁军教头。因为高俅之子高衙内垂涎林冲妻子，高俅伙同陆谦设计使林冲误闯白虎节堂，被发配沧州。途中在野猪林由鲁智深搭救，幸免遇难。但高俅又使人火烧草料场，林冲才大闹山神庙，格杀陆谦，雪夜上梁山。后来火并王伦，拥立晁盖为梁山之主。梁山座次排在第六，一杆枪，天下无敌，从无败绩。
6
统一 小浣熊

Example:

108

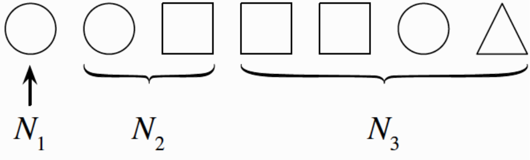
Example:

Example:

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 $\text{Var}(N)$.

Solution: Coupon Collector



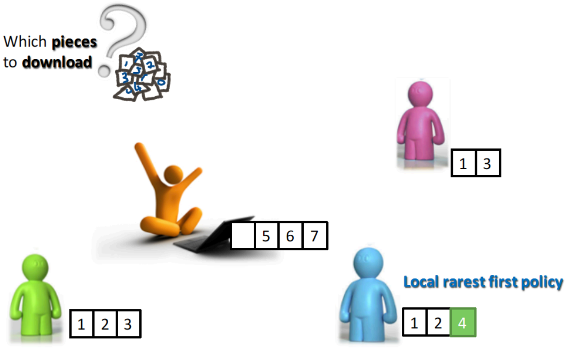
Solution: Coupon Collector

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

1 Expectation & Variance
2 Geometric and Negative Binomial
3 Indicator R.V.s and The Fundamental Bridge
4 Moments and Indicators
5 Poisson
6 Distance between Two Probability Distributions
7 Probability Generating Functions
8 Reading for Fun

Properties of Indicator R.V.

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

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

Fundamental Bridge Between Probability and Expectation

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:

$$P(A) = E(I_A).$$

Example: Booler's Inequality

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

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

Solution: Booler's Inequality

Example: Inclusion-Exclusion Formula

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

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

Solution: Inclusion-Exclusion Formula

Outline

1 Expectation & Variance

2 Geometric and Negative Binomial

3 Indicator R.V.s and The Fundamental Bridge

4 Moments and Indicators

5 Poisson

6 Distance between Two Probability Distributions

7 Probability Generating Functions

8 Reading for Fun

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
- $E(\binom{X}{2}) = \sum_{i<j} P(A_i \cap A_j)$
  - $E(X^2) = 2\sum_{i<j} P(A_i \cap A_j) + E(X)$.
  - $\mathrm{Var}(X) = 2\sum_{i<j} P(A_i \cap A_j) + E(X) - (E(X))^2$.

Moments of Binomial Random Variables

Outline

1 Expectation & Variance

2 Geometric and Negative Binomial

3 Indicator R.V.s and The Fundamental Bridge

4 Moments and Indicators

5 Poisson

6 Distance between Two Probability Distributions

7 Probability Generating Functions

8 Reading for Fun

Poisson Distribution

Definition

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

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

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

Example: Poisson Expectation & Variance

Poisson Approximation

Let $A_1, A_2, \cdots, A_n$ 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

Poisson & Binomial

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

Sum of Independent Poissons

Theorem

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

Poisson Given A Sum of Poissons

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* $\mathrm{Bin}(n, \lambda_1/(\lambda_1 + \lambda_2))$.

Poisson Approximation to Binomial

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.

Proof

Visitors to A Website

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.

Outline

1. Expectation & Variance
2. Geometric and Negative Binomial
3. Indicator R.V.s and The Fundamental Bridge
4. Moments and Indicators
5. Poisson
6. Distance between Two Probability Distributions
7. Probability Generating Functions
8. Reading for Fun

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

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 $d_{TV}(\mu, \nu) \leq p^2$.

The Law of Small Numbers

Theorem

Given independent random variables $Y_1, \cdots, Y_n$ *such that for any* $1 \leq m \leq n$, $\mathbb{P}(Y_m = 1) = p_m$ *and* $\mathbb{P}(Y_m = 0) = 1 - p_m$. *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

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

Gap of Poisson Approximation

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

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

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

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

Outline

1. Expectation & Variance
2. Geometric and Negative Binomial
3. Indicator R.V.s and The Fundamental Bridge
4. Moments and Indicators
5. Poisson
6. Distance between Two Probability Distributions
7. Probability Generating Functions
8. Reading for Fun

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

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

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

Example: Generating Dice Probabilities

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

Solution

PGF and Moments

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

- $E(X) = g'(t)|_{t=1}$
- $E(X(X-1)) = g''(t)|_{t=1}$

PGF and Moments

PGF and Moments

Binomial PMF

Binomial Moments

Example: Pattern Matching

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 $\text{Var}(N)$.

Example: Pattern Matching

Example: Pattern Matching

Example: Pattern Matching

Example: Pattern Matching

Outline

1. Expectation & Variance
2. Geometric and Negative Binomial
3. Indicator R.V.s and The Fundamental Bridge
4. Moments and Indicators
5. Poisson
6. Distance between Two Probability Distributions
7. Probability Generating Functions
8. Reading for Fun

Probability Method

- Paul Erdős initiated this method: Erdős 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.

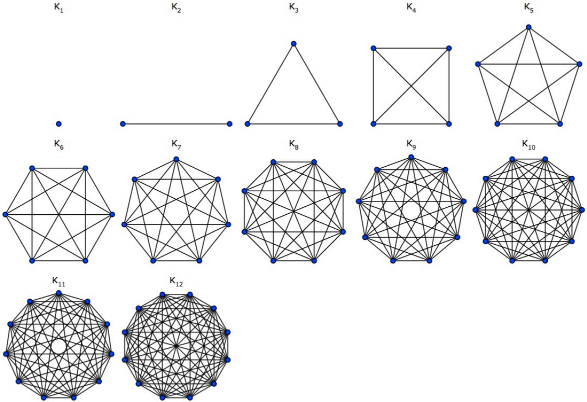
Principle II

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

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

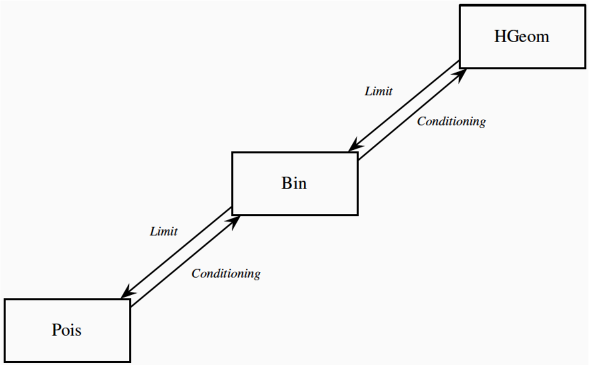
Remarks

- 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.

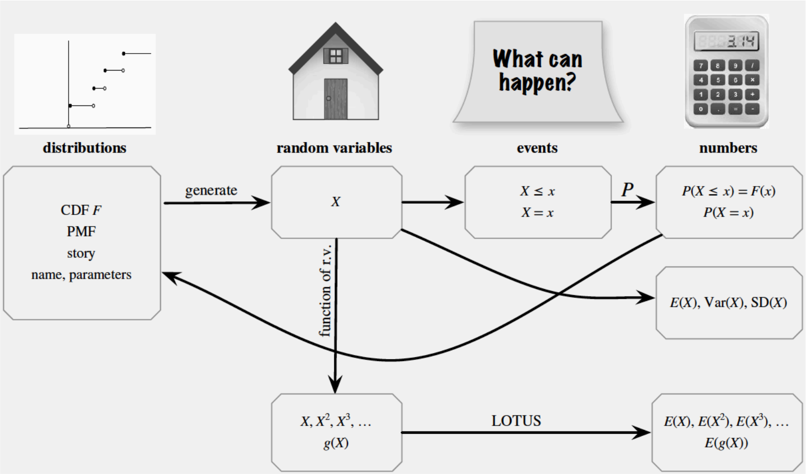
Summary 1

	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
X ≤ x
X = x
P
P(X ≤ x) = F(x)
P(X = x)
function of r.v.
E(X), Var(X), SD(X)
X, X², X³, …
g(X)
LOTUS
E(X), E(X²), E(X³), …
E(g(X))

References

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