Serik Sagitov, Chalmers Tekniska Högskola, February 9, 2005

2. Discrete random variables

$\mathbb{N}_0 = \{0, 1, 2, 3, \ldots\}$

$\mathbb{N} = \{1, 2, 3, \ldots\}$

2.1 Probability distribution

Def 1: discrete random variable

$X : \Omega \to \{x_1, x_2, x_3, \ldots\}$

is a number resulting from a random experiment

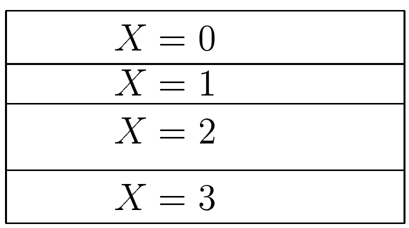
finitely or countably many possible values

$x_1 < x_2 < x_3 < \ldots$

Def 2: random count

$X : \Omega \to \mathbb{N}_0$

is a counting result in a random experiment



Partition of Ω caused by a r.v. X

Def 3: probability distribution

The probability distribution of a r.v. X

is the set of probabilities for all possible values of X

Probability mass function (pmf)

$p_k = \mathrm{P}(X = x_k)$, $k \in \mathbb{N}$, $p_1 + p_2 + p_3 + \ldots = 1$

Pmf for a random count

$p_k = \mathrm{P}(X = k)$, $k \in \mathbb{N}_0$, $p_0 + p_1 + p_2 + \ldots = 1$

Ex 1: coin-die experiment

first step: a fair coin is tossed: $\mathrm{P}(H) = \frac{1}{2}$, $\mathrm{P}(T) = \frac{1}{2}$

second step: a die is rolled once if H or twice if T

Discrete r.v. $D = \{\text{total die score}\}$

$p_0 = 0, p_1 = 6/72, p_2 = 7/72, \ldots, p_6 = 11/72$

$p_7 = 6/72, p_8 = 5/72, \ldots, p_{12} = 1/72$

$X = 1$	2	3	4	5	6	7
$X = 2$	3	4	5	6	7	8
$X = 3$	4	5	6	7	8	9
$X = 4$	5	6	7	8	9	10
$X = 5$	6	7	8	9	10	11
$X = 6$	7	8	9	10	11	12

Def 4: cumulative distribution function

$F(k) = \mathrm{P}(X \leq k) = p_0 + p_1 + \ldots + p_k$

increases from p_0 to 1

Properties of cdf

$\mathrm{P}(X > k) = 1 - F(k) = \mathrm{P}(X \geq k + 1)$

$\mathrm{P}(k_1 < X \leq k_2) = F(k_2) - F(k_1)$

$p_k = F(k) - F(k - 1)$

2.2 Mean value and standard deviation

Def 5: mean, variance, st. deviation

Mean value μ of X or expectation $\mathrm{E}(X)$

is the probability mass center of X

$\mu = \sum x_k p_k$

$\mu = p_1 + 2p_2 + 3p_3 + \ldots$ for a random count

Variance

$\sigma^2 = \mathrm{Var}(X) = \mathrm{E}((X-\mu)^2)$

is the mean squared deviation of X from its mean

Standard deviation

$\sigma = \sqrt{\mathrm{Var}(X)}$

measures the distribution spread in same units as X

Calculate the variance by $\sigma^2 = \mathrm{E}(X^2) - \mu^2$

Properties of Expectation and Variance

$\mathrm{E}(X+Y) = \mathrm{E}(X) + \mathrm{E}(Y)$

$\mathrm{E}(c \cdot X) = c \cdot \mathrm{E}(X)$, $\mathrm{Var}(c \cdot X) = c^2 \cdot \mathrm{Var}(X)$

$\mathrm{E}g(X) = \sum g(x_k) p_k$, $\mathrm{E}(X^i) = \sum (x_k)^i p_k$

$\mu = \sum_{k=0}^{\infty} (1 - F(k))$ for a random count

If X and Y are independent, then
$\mathrm{E}(XY)=\mathrm{E}(X)\mathrm{E}(Y)$, $\mathrm{Var}(X+Y)=\mathrm{Var}(X)+\mathrm{Var}(Y)$

Ex 2: students' grades

Compare three grade distributions:

Grade	2	3	4	5	Total
Student A	25	25	25	25	100%
Student B	40	10	10	40	100%
Student C	10	40	40	10	100%

X	$\mathrm{E}(X)$	$\mathrm{E}(X^2)$	$\mathrm{Var}(X)$	σ_X
Student A's grade	3.5	13.5	1.25	1.12
Student B's grade	3.5	14.1	1.85	1.36
Student C's grade	3.5	12.9	0.65	0.81

2.3 Discrete uniform distribution

Discrete uniform distr. with parameter N

$X \sim \mathrm{U}(N)$, $N \in \mathbb{N}$

$p_k = \frac{1}{N}$, $k = 1, \ldots, N$

$\mu = \frac{N+1}{2}$, $\sigma^2 = \frac{N^2-1}{12}$

Ex 3: systematic search

Open a lock by trying codes: $0000, 0001, 0002, \ldots$

number of trials required: $X \sim \mathrm{U}(10000)$

$\mu = 5000.5$ trials

$\sigma^2 = 8.3 \cdot 10^6$ squared trials

$\sigma = 2886.8$ trials

2.4 Binomial distribution

Binomial distribution with parameters n and p

$X \sim \mathrm{Bin}(n, p)$, $n \in \mathbb{N}$, $0 < p < 1$

$p_k = \binom{n}{k} p^k q^{n-k}$, $k = 0, 1, \ldots, n$, $q = 1 - p$

$\mu = np$

$\sigma^2 = npq$, $\sigma = \sqrt{npq}$

Def 6: Bernoulli trials

independently repeated experiment with

two possible outcomes: success or failure

Number of successes in n Bernoulli trials $X \sim \mathrm{Bin}(n, p)$

p is the probability of success

$q = 1 - p$ is the probability of failure

If $X \sim \text{Bin}(n, p)$, then $X = I_1 + \ldots + I_n$

where $I_1, \ldots, I_n$ are independent with $\text{P}(I_j = 1) = p$

$\text{P}(I_j = 0) = q$, $\text{E}(I_j) = p$, $\text{Var}(I_j) = pq$

Ex 4: sampling with replacement

Consider a box with white and black balls:

$N = 30$ the total number of balls

$p = \frac{1}{3}$ the proportion of black balls in the box

Randomly sample $n = 5$ balls with replacement

number of black balls in the sample $X \sim \text{Bin}(5, \frac{1}{3})$

$\text{P}(\text{BBBWW}) = p^3 q^2 = 0.0165$

$\text{P}(X = 3) = \binom{5}{3} \cdot p^3 q^2 = 0.165$

Ex 5: Ehrenfest model of diffusion

Suppose n molecules of a gas are in a container

divided into two equal parts by a permeable membrane

$X_t =$ number of molecules in the left part at time t

Transition probabilities

$\text{P}(X_{t+1} = k - 1 | X_t = k) = k/n$

$\text{P}(X_{t+1} = k + 1 | X_t = k) = (n - k)/n$

Equilibrium distribution $p_k = \text{P}(X_t = k)$

$p_k = p_{k-1}(n - k + 1)/n + p_{k+1}(k + 1)/n$

$p_k = \binom{n}{k} 2^{-n}$, $k = 0, 1, \ldots, n$

Equilibrium distribution is $\text{Bin}(n, 1/2)$

each molecule chooses one of two parts

independently at random

2.5 Hypergeometric distribution

Hypergeometric distribution with parameters N, n, p

$X \sim \mathrm{Hg}(N, n, p)$

$$n, N, (Np) \in \mathbb{N},\ n \leq N,\ 0 < p < 1$$

$p_k = \frac{\binom{Np}{k}\binom{Nq}{n-k}}{\binom{N}{n}}$, $\max(n - Nq, 0) \leq k \leq \min(n, Np)$

$\mu = np$

$\sigma^2 = npq(1 - \frac{n-1}{N-1})$

Sampling without replacement

N = the total number of balls in the box

p = initial proportion of black balls in the box

X = number of black balls in the sample of size n

$X = I_1 + \ldots + I_n$ with $\mathrm{P}(I_j = 1) = p$, $\mathrm{P}(I_j = 0) = q$

Reduced variance due to

negative dependence between $I_1, \ldots, I_n$

the more black balls are drawn

the less chances to see another black ball

The finite population correction $(1 - \frac{n-1}{N-1})$ is negligible when the sample fraction $\frac{n}{N}$ is small

Ex 6: sampling without replacement

5 balls sampled without replacement

from a box with 10 black and 20 white balls

$\binom{30}{5}$ unordered samples are equally likely

Division rule:

$\mathrm{P}(3 \text{ black} + 2 \text{ white}) = \frac{\binom{10}{3}\binom{20}{2}}{\binom{30}{5}} = \frac{120 \cdot 190}{142506} = 0.16$

Ex 7: aspirin teatment

placebo group: 11034 individuals, 189 heart attacks

aspirin group: 11037 individuals, 104 heart attacks

Statistical model

X = number of heart attacks in the placebo group

without aspirin effect $X \sim \mathrm{Hg}(N, n, p)$

$N = 22071$, $n = 293$, $p = \frac{11034}{22071} = 0.4999$

$\mathrm{P}(X = 189) = \frac{\binom{11034}{189}\binom{11037}{104}}{\binom{22071}{293}} = 0.00000015$

Even the maximal probability is small

$\mathrm{P}(X = 146) = \mathrm{P}(X = 147) = 0.0468$

A different proportion

$\mathrm{P}(X \geq 189)$ would be more informative

2.6 Geometric distribution

Geometric distribution with parameter p

$X \sim \mathrm{Geom}(p)$, $0 < p < 1$

$p_k = pq^{k-1}$, $k \in \mathbb{N}$, cdf $F(k) = 1 - q^k$

$\mu = \frac{1}{p}$, $\sigma^2 = \frac{q}{p^2}$

Bernoulli trials with probability of success p

X = number of trials until the first success

Skewed (non-symmetric) pmf shape

$p_{k+1} = p_k \cdot q$

Lack of memory property for the geometric distribution

$\mathrm{P}(X > t + k | X > t) = \frac{\mathrm{P}(X>t+k)}{\mathrm{P}(X>t)} = \frac{q^{t+k}}{q^t} = \mathrm{P}(X > k)$

Ex 8: birthday problem

Number of. people asked until the same birthday as yours

$X \sim \text{Geom}(1/365)$

$\text{P}(X > 253) = (\frac{364}{365})^{253} = 0.5$

mean of X is 365, median of X is 253

Ex 9: random search

Try the lock codes at random

number of trials required $X \sim \text{Geom}(10^{-4})$

$\mu = 10000$ trials, $\sigma \approx 10000$

$\text{P}(X > 10000) = (0.9999)^{10000} = 0.37 \approx e^{-1}$

2.7 Negative binomial distribution

Negative binomial distribution with parameters r, p

$X \sim \text{Nb}(r, p)$, $r \in \mathbb{N}$, $0 < p < 1$

$p_k = \binom{k-1}{r-1} p^r (1-p)^{k-r}$, $k = r, r+1, r+2, \ldots$

$\mu = \frac{r}{p}$, $\sigma^2 = \frac{rq}{p^2}$

Bernoulli trials with probability of success p

X = number of trials until the r-th success

$X = Y_1 + \ldots + Y_r$ with independent $Y_i \sim \text{Geom}(p)$

2.8 Poisson distribution

Poisson distribution with parameter λ

$X \sim \text{Pois}(\lambda)$, $\lambda > 0$

$p_k = \frac{\lambda^k}{k!} e^{-\lambda}$, $k \in \mathbb{N}_0$

computational formula: $p_{k+1} = p_k \cdot \frac{\lambda}{k+1}$

$\mu = \sigma^2 = \lambda$

Poisson approximation of the Binomial distr

Poisson distribution is a distribution law of rare events
small p and large n (jackpot wins, accidents)

$$\boxed{\text{Bin}(n,p) \approx \text{Pois}(np) \text{ if } n \geq 100,\ p \leq 0.01}$$

Exact meaning: for any fixed $k \in \mathbb{N}_0$
$\binom{n}{k} p^k (1-p)^{n-k} \sim \frac{n^k}{k!} p^k e^{-np} \to \frac{\lambda^k}{k!} e^{-\lambda}$
as $np \to \lambda$

Poisson process of radioactive disintegrations

Radioactivity as a flow of Bernoulli trials
p = probability of a disintegration per a millisecond
number of disintegrations during t hours
$X_t \sim \text{Pois}(\lambda t)$, where $\lambda = 1440000p$

Poisson process $\{X_t\}$ counts disintegrations
occuring at the rate λ disintegrations per hour

Other examples of rates
3 asteroids per MY hit the Earth, MY = million years
5 replacements per amino acid per 1000 MY

Ex 10: cystic fybrosis

proportion of affected people $p = 1/3000$
$X = \#\{\text{affected in a random sample of size } n = 6000\}$

Poisson approximation:
$\text{P}(X=3) = \binom{6000}{3} (\frac{1}{3000})^3 (\frac{2999}{3000})^{5997} \approx \frac{2^3}{3!} e^{-2} = 0.180$
$\text{P}(X=1) = 2e^{-2} = 0.271$
$\text{P}(X \leq 3) = e^{-2} + 2e^{-2} + \frac{2^2}{2} e^{-2} + \frac{2^3}{6} e^{-2} = 0.857$