Serik Sagitov, Chalmers Tekniska Högskola, August 16, 2004

2. Random variables

2.1 Probability distribution

Def 1: random variable

$X : \Omega \to (-\infty, \infty)$

is a number resulting from a random experiment

Ex 1: students' data

$X =$ sex, $Y =$ height, $\Omega =$ the set of students

$X : \Omega \to \{1, 2\}$ dichotomous random variable

$Y : \Omega \to (100, 250)$ continuous random variable

Def 2: probability distribution

Probability distribution records

all possible values of X and their probabilities

Discrete distribution $p(x) = \mathrm{P}(X = x)$, $\Sigma\, p(x) = 1$

probability mass function (pmf), a bar graph

Continuous distribution $f(x) = \frac{1}{\delta}\, \mathrm{P}(x < X < x + \delta)$

probability density function (pdf), a curve

Def 3: cumulative distribution function

$F(x) = \mathrm{P}(X \leq x) = \Sigma_{y \leq x}\, p(y)$ or $\int_{-\infty}^{x} f(y)dy$

increases from 0 to 1

$\mathrm{P}(X > a) = 1 - F(a)$, $\mathrm{P}(a < X \leq b) = F(b) - F(a)$

2.2 Mean value and standard deviatian

Def 4: mean, variance, st. deviatiation

Mean value $\mu = \mathrm{E}(X) = \Sigma\, xp(x)$ or $\int xf(x)dx$
gives the probability mass center of X

Variance $\sigma^2 = \mathrm{Var}(X) = \mathrm{E}((X-\mu)^2)$
is the mean squared deviation of X

Standard deviation $\sigma = \sqrt{\mathrm{Var}(X)}$
measures the distribution spread (in same units as X)

Properties of Expectation and Variance

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

$\mathrm{E}(c{\cdot}X) = c{\cdot}\mathrm{E}(X)$

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

$\mathrm{E}(X^k) = \Sigma\, x^k p(x)$ or $\int x^k f(x)dx$

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

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 Uniform distributions

Discrete uniform distr. $X \sim \mathrm{dU}(N)$: $x = 1, \ldots, N$
pmf $p(x) = \frac{1}{N}$, $\mu = \frac{N+1}{2}$, $\sigma^2 = \frac{N^2-1}{12}$

Uniform $X \sim \mathrm{U}(a, b)$: pdf $f(x) = \frac{1}{b-a}$, $a < x < b$
$\mu = \frac{a+b}{2}$, $\sigma^2 = \frac{(b-a)^2}{12}$, $\sigma = (b-a) \cdot 0.289$

Ex 3: systematic search

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

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

$\mu = 5000.5$ trials, $\sigma^2 = 8.3 \cdot 10^6$, $\sigma = 2886.8$ trials

Search time

$T = \frac{X}{1000}$ h, if you do 1000 combinations per hour

Continuous approximate distribution

$T \sim \mathrm{U}(0, 10)$, $\mu = 5$ h, $\sigma = 2$h 53 min

σ is not the mean devation which is 2h 30 min

Compare $\mathrm{P}(T > 3) = 0.7$ and

$\mathrm{P}(T > 5 | T > 2) = \frac{\mathrm{P}(T>5)}{\mathrm{P}(T>2)} = \frac{(10-5)/10}{(10-2)/10} = 0.625$

2.4 Binomial distribution

Def 5: Bernoulli trials

independently repeated experiment with

two possible outcomes: success or failure

Binomial $X \sim \mathrm{Bin}(n, p)$: $x = 0, 1, \ldots, n$
pmf $p(x) = \binom{n}{x} p^x q^{n-x}$, $\mu = np$, $\sigma = \sqrt{npq}$

Here X = number of successes in n Bernoulli trials

p = probability of success

$q = 1 - p$ = probability of failure

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 $X \sim \text{Bin}(5, \frac{1}{3})$

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

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

Ex 5: ascertainment bias

Cystic fibrosis is an autosomal recessive disease

consider 3 children to parents which both are carriers

Number of affected children

$X \sim \text{Bin}(3, 0.25)$, $\text{E}(\frac{X}{3}) = 0.25$

$p(0) = p(1) = \frac{27}{64}$, $p(2) = \frac{9}{64}$, $p(3) = \frac{1}{64}$

Observed number of affected children $Y = 1, 2, 3$

$\text{P}(Y = k) = \text{P}(X = k | X \geq 1) = \frac{\text{P}(X=k)}{\text{P}(X \geq 1)}$

$\text{P}(Y = 1) = \frac{27}{37}$, $\text{P}(Y = 2) = \frac{9}{37}$, $\text{P}(Y = 3) = \frac{1}{37}$

$\text{E}(\frac{Y}{3}) = 0.43$ is closer to the dominant proprotion 50%

2.5 Hypergeometric distribution

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

Hypergeom. $X \sim \text{Hg}(N, n, p)$, $0 \le x \le \min(n, Np)$
pmf $p(x) = \frac{\binom{Np}{x}\binom{Nq}{n-x}}{\binom{N}{n}}$, $\mu = np$, $\sigma = \sqrt{npq(1 - \frac{n-1}{N-1})}$

Reduced variance due to negative dependence

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:

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

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

X = number of Bernoulli trials until the first success

Geometric distribution $X \sim \mathrm{G}(p)$: $x = 1, 2, 3, \ldots$
$F(x) = 1 - q^x$, $p(x) = pq^{x-1}$, $\mu = \frac{1}{p}$, $\sigma^2 = \frac{q}{p^2}$

Skewed (non-symmetric) pmf shape

$p(x + 1) = p(x) \cdot q$

Lack of memory property for the geometric distribution

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

Ex 8: die experiment

$X = \#\{$die rolls untill the first "6"$\}$

$p(1) = \frac{1}{6} = 0.167$, $p(2) = \frac{5}{6} \cdot \frac{1}{6} = 0.139$

$p(3) = \frac{5}{6} \cdot \frac{5}{6} \cdot \frac{1}{6} = 0.116$, $p(4) = \frac{5}{6} \cdot \frac{5}{6} \cdot \frac{5}{6} \cdot \frac{1}{6} = 0.097$

$\mu = 6$, $\sigma = 5.48$

2.7 Exponential distribution

$$\boxed{\begin{array}{c}\text{Exponential distribution } X \sim \text{Exp}(\lambda)\text{: } 0 < x < \infty \\ F(x) = 1 - e^{-\lambda x},\ f(x) = \lambda e^{-\lambda x},\ \mu = \sigma = \frac{1}{\lambda}\end{array}}$$

$\lambda > 0$ is a scale parameter: $\text{Exp}(\lambda) = \frac{1}{\lambda}\,\text{Exp}(1)$

Exponential approximation of the geometric distribution

if success is rare: p is small and n is large so that $np = \lambda$ then $\frac{1}{n}\text{G}(p) \approx \text{Exp}(\lambda)$

Def 6: median value

such a value M that $\text{P}(X \geq M) = \text{P}(X \leq M)$

If distribution is symmetric, then median = mean

$$\boxed{\text{If } X \sim \text{Exp}(\lambda), \text{ then } M = \frac{\ln 2}{\lambda} = 0.693 \cdot \mu}$$

Ex 9: carbon-14 decay

$M = 5730$ years half-life of carbon-14, $\mu = 8267$ years

Ex 10: random search

Try the door codes at random

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

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

Search time

$T = \frac{X}{1000}$ hours, $T \sim \text{Exp}(0.1)$, $\mu = \sigma = 10$ hours

Continous memoryless distribution

$\text{P}(T > 5 | T > 2) = \frac{e^{-0.5}}{e^{-0.2}} = e^{-0.3} = 0.741 = \text{P}(T > 3)$

2.8 Poisson distribution

$$\boxed{X \sim \text{Pois}(\lambda)\text{: pmf } p(x) = \frac{\lambda^x}{x!}e^{-\lambda},\ x \geq 0,\ \mu = \sigma^2 = \lambda}$$

computational formula: $p(x+1) = p(x) \cdot \frac{\lambda}{x+1}$

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

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

Poisson process: random flow of events at the rate
λ events per time unit $\Rightarrow$ independent
interarrival times with common distribution $\text{Exp}(\lambda)$

Ex 11: cystic fybrosis

proportion of affected people $p = 1/3000$
$X = \#\{\text{affected in a random sample of size } n = 6000\}$
$X \sim \text{Bin}(n,p)$, $\mu = np = 2$, $\sigma^2 = npq = 1.9993$

Poisson approximation:

$$\text{P}(X = 3) = \binom{6000}{3}\left(\frac{1}{3000}\right)^3\left(\frac{2999}{3000}\right)^{5997} \approx \frac{2^3}{3!}e^{-2} = 0.180$$

$$\text{P}(X = 1) = 2e^{-2} = 0.271$$

$$\begin{aligned}\text{P}(X \leq 3) &= e^{-2} + 2e^{-2} + \frac{2^2}{2}e^{-2} + \frac{2^3}{6}e^{-2} \\ &= 0.135 + 0.271 + 0.271 + 0.180 = 0.857\end{aligned}$$

Ex 12: miscellaneous

radioactive disintegrations
detective items

centenarians of Switzerland

3 asteroids per MY hit the Earth, MY = million years

5 replacements per amino acid per 1000 MY

2.9 Normal distribution

Standard normal distribution $Z \sim \mathrm{N}(0,1)$
zero mean $\mu = 0$, unit spread $\sigma = 1$

Normal distribution table

z	1.00	1.28	1.64	1.96	2.00	2.33	2.58	3.00
$\Phi(z)$	.84	.90	.95	.975	.977	.99	.995	.9987
$2\Phi(z)$–1	.68	.80	.90	.95	.954	.98	.99	.9974

$\Phi(z) = \mathrm{P}(Z < z)$, $\mathrm{P}(Z > z) = 1 - \Phi(z)$

$\mathrm{P}(Z < -z) = 1 - \Phi(z)$, $\mathrm{P}(|Z| > z) = 2(1 - \Phi(z))$

$\mathrm{P}(-z < Z < z) = 2\Phi(z) - 1$

General normal distribution $X \sim \mathrm{N}(\mu, \sigma^2)$

location parameter $\mu = \mu_X$, scale parameter $\sigma = \sigma_X$

Standardized random variable $\frac{X-\mu}{\sigma} \sim \mathrm{N}(0,1)$

$\mathrm{P}(X < \mu + \sigma z) = \Phi(z)$

Normal pdf

Symmetric "bell curve" centered at μ

exact meaning of σ: two inflection points $\mu \pm \sigma$

$(\mu - 3\sigma)$ $\underline{2\%}$ $(\mu - 2\sigma)$ $\underline{14\%}$ $(\mu - \sigma)$ $\underline{34\%}$ (μ)

(μ) $\underline{34\%}$ $(\mu + \sigma)$ $\underline{14\%}$ $(\mu + 2\sigma)$ $\underline{2\%}$ $(\mu + 3\sigma)$

Three-sigma rule

99.74% of the N(μ, σ^2) values are within $\mu \pm 3\sigma$
it requires on average $\frac{1}{(1-0.9974)} = 385$ observations
to see a three-sigma outlier

Ex 13: Intellegence Quotient

Given IQ $\sim$ N(100,15^2) find
P(IQ < 85), P(IQ > 115), P(IQ > 130), P(IQ > 145)
P(|IQ – 100| > 45), P(IQ > 175) $= 3 \cdot 10^{-7}$

2.10 Central Limit Theorem

If $X_1, \ldots, X_n$ is a large number of
independent or weakly dependent values
and each of the values is relatively small
Then $(X_1 + \ldots + X_n)$ is approximately normal

Normal approximations

Bin$(n, p) \approx$ N(np, npq), $np \geq 5$, $nq \geq 5$
Pois$(\lambda) \approx$ N(λ, λ), $\lambda \geq 5$
Hg$(N, n, p) \approx$ N$(np, npq\frac{N-n}{N-1})$, $np \geq 5$, $nq \geq 5$

Sample mean

Random sample with large sample size n
independent repeated measurements $X_1, \ldots, X_n$
sample mean $\frac{X_1+\ldots+X_n}{n} \approx$ N$(\mu, \frac{\sigma^2}{n})$

Ex 14: diversification experiment

Three options of a special study support

a) take 4500 SEK

b) toss a coin and get 10000 SEK in case of heads

c) toss 10000 one-SEKs and collect all heads-up coins

Amount of money collected in the last case

$X \sim \text{Bin}(10000, 0.5)$, three-sigma rule: 5000 ± 150

Ex 6: aspirin teatment

$X = \#\{\text{heart attacks in the placebo group}\}$

Assuming no aspirin effect

$X \sim \text{Hg}(22071, 293, 0.4999) \approx \text{N}(146.48, 72.28)$

$\text{P}(X \geq 189) \approx 1 - \Phi(\frac{189-146.48}{8.50}) = 1 - \Phi(5)$

$= 0.0000003$ statistically significant aspirin effect

2.11 Probability distribution quiz

Suggest a probability distribution or pmf/pdf shape for

1. Number of children untill the first son
2. Waiting time for a bus
3. Your daily expenses
4. The number of matches when guessing 6 out of 49
5. The next digit in the number $\pi = 3.1415926535897$
6. 10 people throw out fingers, total number of fingers
7. Human lifelength
8. Number of W among 500 amino acids, $p_W = 1.3\%$