Serik Sagitov, Chalmers Tekniska Högskola, September 8, 2003
http://www.math.chalmers.se/Stat/Grundutb/Chalmers/TMS100/

Basics of Probability and Statistics

Probability theory

mathematical study of uncertainty and random variation

1. Probability rules
2. Random variables
3. Joint distributions

Mathematical statistics

deals with variation in data using probability theory

4. Parameter estimation
5. Hypotheses testing
6. Simple linear regression
7. Chi-square tests
8. Decision theory and Bayesian inference

Lab assignment

data to be collected: sex, hair color, height, weight

Ex 1: aspirin treatment

Is heart attack risk reduced by taking aspirin?

11034 took placebo and 11037 took aspirin: of them
189 and 104 subsequently experienced heart attacks

1. Probability rules

1.1 Main concepts

random experiment $\to$ random event $\to$ probability

Def 1: sample space

Ω is the set of all possible outcomes in a random experiment (finite or infinite, discrete or continuous)

Def 2: random event A is a subset of Ω, $A \subset \Omega$

Def 3: probability $\mathrm{P}(A)$

number between 0 and 1 says how likely A is to occur

$\mathrm{P}(A) = 1$ means A is certain, $\mathrm{P}(A) = 0$ means impossible

$$\boxed{\text{probability} = \text{population proportion}}$$

1.2 Division rule

Division rule: if all outcomes are equally likely, then

$$\mathrm{P}(A) = \frac{\#(A)}{\#(\Omega)} = \frac{\text{number of favorable outcomes}}{\text{total number of outcomes}}$$

Ex 2: coin experiment toss a coin: $\#(\Omega) = 2$

Ex 3: die experiment roll a die: $\#(\Omega) = 6$

Ex 4: sibling sampling

Five families with two children:

three with boy and girl, two with boy and boy

Two sampling experiments

experiment 1: pick a family at random

experiment 2: pick a boy at random, consider his family

Find $\mathrm{P}(A)$, $A = \{\text{the chosen family has two boys}\}$

1.3 Basic combinatorics

How to count the numbers of outcomes $\#(\Omega)$
in an r-step experiment given
$N_i = \#$(outcomes in the i-th step), tree of outcomes

Multiplication principle: $\#(\Omega) = N_1 \times N_2 \times \ldots \times N_r$

Ex 5: two dice experiment

Two dice are rolled: $\#(\Omega) = 6 \times 6 = 36$
P(the sum of points on two dice equals 5) $= \frac{4}{36} = \frac{1}{9}$

Ex 6: sampling with replacement

Random experiment:
draw $n = 3$ balls with replacement from a box
containing $N = 4$ balls labelled $\{1, 2, 3, 4\}$
$\#(\Omega) = 4 \times 4 \times 4 = 64$

Def 4: permutation and combination

permutation = the ordered set of labels in the sample
combination = unordered set of labels in the sample

Number of permutations of N distinct objects taken n at a time: $N \times (N-1) \times \ldots \times (N-n+1) = \frac{N!}{(N-n)!}$

The number of combinations of N distinct objects taken n at a time equals $\binom{N}{n} = \frac{N!}{n!(N-n)!}$

Numbers $\binom{n}{k}$ form Pascal's triangle and are often called
binomial coefficients due to the expansion
$(a+b)^n = a^n + \binom{n}{1}a^{n-1}b + \ldots + \binom{n}{n-1}ab^{n-1} + b^n$

Ex 7: sampling without replacement

Four objects are taken 3 at a time

number of permutations $= 4 \times 3 \times 2 = 24$

number of combinations $= \frac{24}{3\times2\times1} = 4$

123 132 213 231 312 321	124 142 214 241 412 421
134 143 314 341 413 431	234 243 324 342 423 432

Def 5: multinomial coefficient

Number of possible allocations in the random experiment:

allocate n distinct objects into r distinct boxes

box sizes $n_1, \ldots, n_r$

total size of the boxes $n_1 + \ldots + n_r = n$

Multinomial coefficient $\binom{n}{n_1, n_2, \ldots, n_r} = \frac{n!}{n_1! n_2! \ldots n_r!}$

In particular binomial coefficient $\binom{n}{k} = \binom{n}{k, n-k}$

Ex 8: Wright-Fisher model

Population model: $N = 5$ of females per generation

girls choose mothers at random: $\#(\Omega) = 5^5 = 3125$

N daughters allocated among N mothers

Random events

$A = \{$daughter allocation $= (2, 0, 2, 0, 1)\}$

$B = \{$two mothers have two daughters each$\}$

$\#(A) = \binom{5}{2,0,2,0,1} = 30$, $\mathrm{P}(A) = \frac{30}{3125} = 0.01$

$\#(B) = \#(A) \times \binom{5}{2,2,1} = 900$, $\mathrm{P}(B) = 0.29$

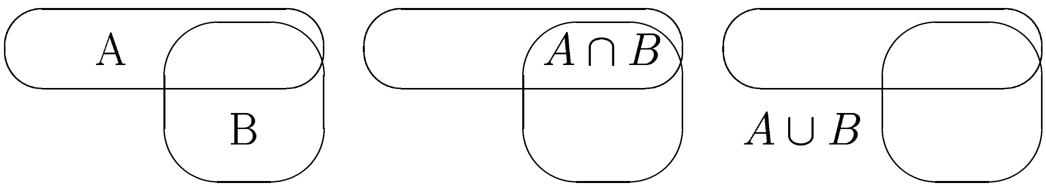
1.4 Addition rule of probability

$$\boxed{P(A \cup B) = P(A) + P(B) - P(A \cap B)}$$

Def 6: intersection and union of two events

$A \cap B = \{A \text{ and } B\}$, $A \cup B = \{A \text{ or } B \text{ or both}\}$

Venn diagrams



Def 7: mutually exclusive events

A and B are mutually exclusive, if $P(A \cap B) = 0$

If A and B are mutually exclusive, then

$P(A \cup B) = P(A) + P(B)$

Def 8: complementary event

$\bar{A} = \{A \text{ has not occurred}\}$ $\boxed{P(\bar{A}) = 1 - P(A)}$

Ex 9: molar absence

The absence of molars is an autosomal dominant trait

consider a son and a grandson of an affected male

$A = \{\text{son is affected}\}$ and $B = \{\text{grandson is affected}\}$

$A \cap B = \{\text{both the son and grandson are affected}\}$

$A \cup B = \{\text{either son or grandson or both are affected}\}$

Compute $P(A)$, $P(B)$, $P(A \cap B)$, and $P(A \cup B)$

hint: $B \subset A$, that is event B implies event A

1.5 Conditional probability

Def 9: joint probability of two events $\mathrm{P}(A \cap B)$

Def 10: conditional probability $\mathrm{P}(A|B) = \frac{\mathrm{P}(A \cap B)}{\mathrm{P}(B)}$

of a random event A given that event B has occurred

Multiplication rule of probability
$\mathrm{P}(A \cap B) = \mathrm{P}(A|B)\mathrm{P}(B)$

$\mathrm{P}(A \cap B \cap C) = \mathrm{P}(A|B \cap C)\mathrm{P}(B|C)P(C)$

The Law of Total Probability (LTP)
$\mathrm{P}(A) = \mathrm{P}(A|B)\mathrm{P}(B) + \mathrm{P}(A|\bar{B})\mathrm{P}(\bar{B})$

Given a partition $\{B_1, B_2, B_3\}$ of Ω

$\mathrm{P}(A) = \mathrm{P}(A|B_1)\mathrm{P}(B_1) + \mathrm{P}(A|B_2)\mathrm{P}(B_2) + \mathrm{P}(A|B_3)\mathrm{P}(B_3)$

Def 11: partition $\{B_1, B_2, B_3\}$ of Ω

pairwise mutually exclusive events, $B_1 \cup B_2 \cup B_3 = \Omega$

Ex 10: 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

Tree of outcomes: $6+36 = 42$ not equally likely outcomes

random event $A = \{\text{total die score} = 5\}$

Division rule:

$\mathrm{P}(A|H) = \frac{\#(A \cap H)}{\#(H)} = \frac{1}{6}$, $\mathrm{P}(A|T) = \frac{\#(A \cap T)}{\#(T)} = \frac{4}{36}$

Multiplication rule:

$\mathrm{P}(A \cap H) = \frac{1}{6} \cdot \frac{1}{2} = \frac{1}{12}$ and $\mathrm{P}(A \cap T) = \frac{1}{9} \cdot \frac{1}{2} = \frac{1}{18}$

LTP: $\mathrm{P}(A) = \frac{1}{12} + \frac{1}{18} = \frac{5}{36} = 0.139$

1.6 Bayes' formula

$$\boxed{\mathrm{P}(B|A) = \frac{\mathrm{P}(A|B)\mathrm{P}(B)}{\mathrm{P}(A)}}$$

Def 12: prior and posterior probabilities

$\mathrm{P}(B)$ the probability of B before a measurement

$\mathrm{P}(B|A)$ the probability of B after A is observed

Ex 11: a genetic test

I consider getting screened for a rare genetic disease

$B = \{\text{I have the disease}\}$

prior probabilities $\mathrm{P}(B) = 0.000001$, $\mathrm{P}(\bar{B}) = 0.999999$

The genetic test is 99% sensitive and 97% specific

$A = \{\text{positive test result}\}$

true and false positive $\mathrm{P}(A|B) = 0.99$, $\mathrm{P}(A|\bar{B}) = 0.03$

true and false negative $\mathrm{P}(\bar{A}|\bar{B}) = 0.97$, $\mathrm{P}(\bar{A}|B) = 0.01$

The total probability of a positive test result

LTP: $\mathrm{P}(A) = 0.99 \cdot 0.000001 + 0.03 \cdot 0.999999 = 0.03$

Posterior probabilities given a positive test result:

$\mathrm{P}(B|A) = \frac{0.99 \cdot 0.000001}{0.03} = 0.000033$, $\mathrm{P}(\bar{B}|A)$=0.999967

After the first positive result I will take the second test

updated prior probabilities

$\mathrm{P}(B|A) = 0.000033$, $\mathrm{P}(\bar{B}|A) = 0.999967$

$C = \{\text{second test result is positive}\}$

$\mathrm{P}(C|A) = 0.99 \cdot 0.000033 + 0.03 \cdot 0.999967 = 0.03$

$\mathrm{P}(B|A \cap C) = \frac{0.99 \cdot 0.000033}{0.03} = 0.0011$

$\mathrm{P}(\bar{B}|A \cap C) = 0.9989$

1.7 Independence

Def 13: independent events

Events A and B are called independent if knowing that one event has occured gives no information about the other event: $\mathrm{P}(A|B) = \mathrm{P}(A)$ and $\mathrm{P}(B|A) = \mathrm{P}(B)$

$$\boxed{A \text{ and } B \text{ are independent if } \mathrm{P}(A \cap B) = \mathrm{P}(A)\mathrm{P}(B)}$$

Def 14: mutually independent events

A, B, C are mutually independent if they are pairwise independent and $\mathrm{P}(A \cap B \cap C) = \mathrm{P}(A)\mathrm{P}(B)\mathrm{P}(C)$

Ex 12: Mendelian segregation

One gene with two alleles A (dominant) and a (recessive) offspring genotype of the cross $Aa \times Aa$:

$\mathrm{P}(AA) = \mathrm{P}(aa) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$, $\mathrm{P}(Aa) = 1 - \frac{1}{4} - \frac{1}{4} = \frac{1}{2}$

Phenotype 3:1 ratio: $\mathrm{P}(F_A) = \mathrm{P}(AA \cup Aa) = \frac{1}{4} + \frac{1}{2} = \frac{3}{4}$

Two genes with dominant A, B and recessive a, b alleles phenotype ratio for the offspring of the cross $\frac{AB}{ab} \times \frac{AB}{ab}$

$$\boxed{\begin{array}{l} \mathrm{P}(F_{Ab}|AA) = \mathrm{P}(\frac{Ab}{Ab}|AA) = p^2, \text{ where } p = \mathrm{P}(\text{crossover}) \\ \mathrm{P}(F_{Ab}|Aa) = \mathrm{P}(\frac{Ab}{ab}|Aa) = pq, \text{ where } q = 1 - p \end{array}}$$

LTP: $\mathrm{P}(F_{Ab}) = p^2 \cdot \frac{1}{4} + pq \cdot \frac{1}{2} = \frac{1-q^2}{4}$

$\mathrm{P}(F_{AB}) = \mathrm{P}(F_A) - \mathrm{P}(F_{Ab}) = \frac{3}{4} - \frac{1-q^2}{4} = \frac{1}{2} + \frac{q^2}{4}$

Unlinked genes $p = q = \frac{1}{2}$ give the phenotype ratio 9:3:3:1

$\mathrm{P}(F_{AB}) = \frac{9}{16}$, $\mathrm{P}(F_{Ab}) = \mathrm{P}(F_{aB}) = \frac{3}{16}$, $\mathrm{P}(F_{ab}) = \frac{1}{16}$

Ex 13: two tossings - one placing

Toss two fair coins, then for the third coin

choose H if two heads or two tails

choose T if one heads one tails

The three coin outcomes are pairwise independent

despite mutual dependence: $\mathrm{P}(T_1 \cap T_2 \cap T_3) = 0$

Ex 14: did Mendel cheat?

Let I be a dominant phenotype offspring of $Aa \times Aa$

$D = \{I\text{'s genotype is } AA\}$, $\bar{D} = \{I\text{'s genotype is } Aa\}$

$C = \{\text{all 10 offspring of } I \times I \text{ have dom. phenotype}\}$

Mendel's classification rule: if C, then accept D

misclassification probability $\mathrm{P}(C|\bar{D}) = (\frac{3}{4})^{10} = 0.056$

specificity $\mathrm{P}(\bar{C}|\bar{D}) = 0.944$, sensitivity $\mathrm{P}(C|D) = 1$

Fisher: Mendel's observed ratio $Aa : AA$

had to be closer to 0.63:0.37 rather than to 0.67:0.33

LTP: $\mathrm{P}(C) = \mathrm{P}(C|D) \cdot \frac{1}{3} + \mathrm{P}(C|\bar{D}) \cdot \frac{2}{3} = 0.37$

Ex 15: all-female disorder

Sex-biased condition: only girls in an affected family

$S = \{\text{the family is affected}\}$

population prevalence of the disorder $\mathrm{P}(S) = 0.01$

Consider a family with seven children

$A = \{\text{all seven children are girls}\}$, $\mathrm{P}(A|S) = 1$

$\mathrm{P}(A|\bar{S}) = 0.0078$, $\mathrm{P}(A) = 0.0177$, $\mathrm{P}(S|A) = 0.5643$