

Overview

The ability to count the possible outcomes in an event is crucial to calculating probabilities. By a **permutation** of size r of n different items, we mean an **arrangement** of r of the items, where the order of the arrangement is important. If the order is not important, the arrangement is called a **combination**.

Example. There are 5×4 permutations and $5 \times 4 / (2 \times 1)$ combinations of size 2 of A, B, C, D, E

Permutations: AB, BA, AC, CA, AD, DA, AE, EA
BC, CB, BD, DB, BE, EB
CD, DC, CE, EC
DE, ED

Combinations: AB, AC, AD, AE, BC, BD, BE, CD, CE, DE

Standard reference books on probability theory give a comprehensive treatment of how these ideas are used to calculate the probability of occurrence of the outcomes of games of chance.

1

Examples: Recombinant Interference

$$\hat{r} = \frac{nr}{n} \quad \text{recombination fraction (gametes or...)}$$

Greater physical distance between loci \rightarrow greater chance to recombine - (homologous). Departure from additivity increases with distance - hence mapping. **Example:** 2 loci A, B, same chromosome, segregated for two alleles at each locus $\rightarrow A, a, B, b \rightarrow$ gametes AB, Ab, aB, ab. Parental types AB, ab gives Ab and aB recombinants. Simple ratio.

Example: For 3 linked loci, A, B, C, relationship based on simple prob. theory $r_{AC} = r_{AB} + r_{BC} - 2r_{AB}r_{BC}$ 3 possible RF, so Interference

$$r_{AC} = r_{AB} + r_{BC} - 2C^* r_{AB}r_{BC} \quad \text{more generally}$$

$$1 - C^* = \text{Interference}, \quad C^* = \text{Coeff. Coincidence}$$

$$C^* = \frac{r_{12}}{2r_{AB}r_{BC}} \quad \text{where } r_{12} \text{ true double recombinant frequency}$$

2

Conditional Probability: BAYES A move towards "Likelihood" Statistics

More formally **Theorem of Total Probability (Rule of Elimination)**

If the events B_1, B_2, \dots, B_k constitute a partition of the sample space S , such that $P\{B_i\} \neq 0$ for $i = 1, 2, \dots, k$, then for any event A of S

$$P\{A\} = \sum_{i=1}^k P\{B_i \cap A\} = \sum_{i=1}^k P\{B_i\}P\{A/B_i\}$$

So, if events B partition the space as above, then for any event A in S , where $P\{A\} \neq 0$

$$P\{B_r/A\} = \frac{P\{B_r \cap A\}}{\sum_{i=1}^k P\{B_i \cap A\}} = \frac{P\{B_r\}P\{A/B_r\}}{\sum_{i=1}^k P\{B_i\}P\{A/B_i\}} \quad \text{BAYES RULE}$$

3

Example - Bayes

40,000 people in a population of 2 million carry a particular virus. $P\{\text{Virus}\} = P\{V_1\} = 0.0002$

Tests to show presence/absence of virus, result in following:

$$P\{T/V_1\} = 0.99 \text{ and } P\{T/V_2\} = 0.01$$

$$P\{N/V_2\} = 0.98 \text{ and } P\{N/V_1\} = 0.02$$

where V_2 is the event virus absent, T , the event = positive test, N the event = negative test. (All *a priori* probabilities)

$$\text{So } P\{V_1/T\} = \frac{P\{V_1\}P\{T/V_1\}}{\sum_{i=1}^k P\{V_i\}P\{T/V_i\}} = 0.01 \quad \text{a posteriori}$$

← Total probability

where events V_i partition the sample space

4

Example - POPULATION GENETICS

Counts - Genotypic "frequencies"

GENE n alleles, $n(n+1)/2$ possible genotypes

Population Equilibrium HARDY-WEINBERG

Genes and "genotypic frequencies" constant from generation to generation (simple relationships genotypic and allelic frequencies)

e.g. 2 allele model p_A, p_a allelic freq. A, a respectively, so genotypic freq AA etc. p_{AA}, p_{Aa}, p_{aa} and have

$$p_{AA} = p_A p_A = p_A^2$$

$$p_{Aa} = p_A p_a + p_a p_A = 2 p_A p_a$$

$$p_{aa} = p_a^2$$

$$(p_A + p_a)^2 = p_A^2 + 2 p_A p_a + p_a^2$$

One generation of Random mating. **H-W single locus**

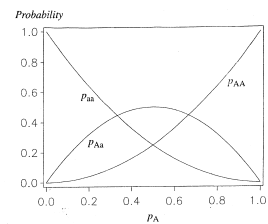
5

POPULATION PICTURE

NB: Frequency heterozygote maximum for 2 allelic frequencies = 0.5 (see Fig.)

Also

$$\frac{p_{Aa}}{p_{Aa} + p_{AA}} = \frac{2 p_A p_a}{2 p_A p_a + p_A^2} = \frac{2 p_a}{(1 + p_a)}$$



So, if rare allele, $p_a = 0.01$ say,

probability high that carried in heterozygous state

6

Example - Multiple Alleles Single Locus

- $p_1, p_2, \dots, p_i, \dots, p_n$ = "frequencies" alleles $A_1, A_2, \dots, A_i, \dots, A_n$, Possible genotypes = $A_{11}, A_{12}, \dots, A_{ij}, \dots, A_{nn}$
- Under H-W equilibrium, Expected genotype frequencies
 $(p_1 + p_2 + \dots + p_i + \dots + p_n)(p_1 + p_2 + \dots + p_j + \dots + p_n)$
 $= p_i^2 + 2p_i p_j + \dots + 2p_i p_j + \dots + 2p_{n-1} p_n + p_n^2$

So, e.g. 4 alleles, 10 genotypes

- Proportion of heterozygosity in population thus
 $P_H = 1 - \sum_i p_i^2$ used in screening of genetic markers

7

Example revisited: Expected genotypic frequencies for a 4-allele system; H-W $\equiv m$, proportion of heterozygosity in F2 progeny

Genotype	Expected frequency	P_i				
		$p_1=0.25$	$p_2=0.3$	$p_3=0.4$	$p_4=0.4$	
A_1A_1	p_1p_1	0.0625	0.09	0.16	0.16	0.49
A_1A_2	$2p_1p_2$	0.125	0.18	0.32	0.24	0.14
A_1A_3	$2p_1p_3$	0.125	0.12	0.08	0.16	0.14
A_1A_4	$2p_1p_4$	0.125	0.12	0.08	0.08	0.14
A_2A_2	p_2p_2	0.0625	0.09	0.16	0.09	0.01
A_2A_3	$2p_2p_3$	0.125	0.12	0.08	0.12	0.02
A_2A_4	$2p_2p_4$	0.125	0.12	0.08	0.06	0.02
A_3A_3	p_3p_3	0.0625	0.04	0.01	0.04	0.01
A_3A_4	$2p_3p_4$	0.125	0.08	0.02	0.04	0.02
A_4A_4	p_4p_4	0.0625	0.04	0.01	0.01	0.01
P_H		0.75	0.74	0.66	0.70	0.48

8

GENERALITY of PROBABILITY RULES and PROPERTIES - Examples in brief

- For ℓ loci, No. of genotypes, (n_i = No. alleles for locus i).

$$\frac{1}{2^\ell} \prod_{i=1}^{\ell} [n_i(n_i + 1)]$$

- Changes in gene frequency - from migration, mutation, selection
- Suppose native population has allelic freq. p_{n0} . Proportion m_i (relative to native population) migrates from i th of k populations to native population every generation; immigrants having allelic frequency p_i . So allelic frequency in mixed population

$$p_{n1} = \left[1 - \sum_{i=1}^k m_i \right] p_{n0} + \sum_{i=1}^k m_i p_i = p_{n0} + \sum_{i=1}^k m_i (p_i - p_{n0})$$

9

BAYES REVISITED- Example Accuracy of Assembled DNA sequences

- Want estimate of probability that i th letter of an assembled sequence is A, C, G, T or -
- Assume each fragment assembly correct, all portions equally reliable, sequencing errors indept. and uniform throughout sequence. Assume letters in sequence IID.
- Let $F^* = \{f_1, f_2, \dots, f_N\}$ set of fragments
- Fragments aligned into assembled sequence corresponding to cols. i in matrix, fragments corresponding to rows j
- Matrix elements x_{ij} are members of $B^* = \{A, C, G, T, -, 0\}$
- True sequence corresponding to n columns is $s = \{s_1, s_2, \dots, s_n\}$ where s contained in $\{A, C, G, T, -\} = A^*$

10

BAYES contd.

Track fragment orientatⁿ. $t_j = \begin{cases} 0 & \text{fragment } j \text{ is as} \\ 1 & \text{fragment } j \text{ is reverse complemented} \end{cases}$

Thus need estimation of

$P_i(M) = P\{s_i = M / x_{ij}, j = 1, \dots, N\}$ = probability i th letter from molecule "M", given matrix elements (of fragments).

Assuming knowledge of sequencing error rates:

$$P\{b / M\} = P\{x_{ij} = b / s_i = M\}, \quad M \in A^*, b \in B^*$$

so that Bayes gives

$$P_i(M) = \frac{P(M) \prod_{j=1}^N [(1-t_j)P(x_{ij} / M) + t_jP(\bar{x}_{ij} / a\bar{M})]}{\sum_{b \in A^*} P(b) \prod_{j=1}^N [(1-t_j)P(x_{ij} / b) + t_jP(\bar{x}_{ij} / \bar{b})]}$$

11

MEASURING PROBABILITIES – RANDOM VARIABLES

If a statistical experiment only gives rise to real numbers, the outcome of the experiment is called a **random variable**. If a random variable X takes values X_1, X_2, \dots, X_n with probabilities p_1, p_2, \dots, p_n

then the **expected** or **average value** of X is defined to be

$$E[X] = \sum_{j=1}^n p_j X_j$$

and its **variance** is

$$VAR[X] = E[X^2] - E[X]^2 = \sum_{j=1}^n p_j X_j^2 - E[X]^2$$

12

PROPERTIES

- Sums and Differences of Random Variables**

Define the **covariance** of two random variables to be
 $COVAR [X, Y] =$

$$E [(X - E[X]) (Y - E[Y])] = E[X Y] - E[X] E[Y]$$

If X and Y are **independent**, $COVAR [X, Y] = 0$.

Lemma $E [X \pm Y] = E[X] \pm E[Y]$
 $VAR [X \pm Y] = VAR [X] + VAR [Y]$

$$\pm 2COVAR [X, Y]$$

$$E [k \cdot X] = k \cdot E[X] \quad VAR [k \cdot X] = k^2 \cdot E[X]$$

for a constant k .

13

R.V. Properties Example. A lab. records the preparation time X and running time Y for given type of experiments. Over a period, records show:

	X=1	2	3	4	Totals
Y=1	7	5	4	4	20
2	2	6	8	3	19
3	1	2	5	3	11
Totals	10	13	17	10	50

$$E[X] = \{1(10)+2(13)+3(17)+4(10)\}/50 = 2.54$$

$$E[X^2] = \{1^2(10)+2^2(13)+3^2(17)+4^2(10)\}/50 = 7.5$$

$$VAR[X] = 7.5 - (2.54)^2 = 1.0484$$

$$E[Y] = \{1(20)+2(19)+3(11)\}/50 = 1.82$$

$$E[Y^2] = \{1^2(20)+2^2(19)+3^2(11)\}/50 = 3.9$$

$$VAR[Y] = 3.9 - (1.82)^2 = 0.5876$$

14

Example Contd.

$$E[X+Y] = \{ 2(7)+3(5)+4(4)+5(4)+3(2)+4(6)+5(8)+6(3)+4(1)+5(2)+6(5)+7(3) \} / 50 = 4.36$$

$$E[(X + Y)^2] = \{ 2^2(7)+3^2(5)+4^2(4)+5^2(4)+3^2(2)+4^2(6)+5^2(8)+6^2(3)+4^2(1)+5^2(2)+6^2(5)+7^2(3) \} / 50 = 21.04$$

$$VAR[(X+Y)] = 21.04 - (4.36)^2 = E[(X + Y)^2] - (E[X+Y])^2 = 2.0304 *$$

$$E[X Y] = \{ 1(7)+2(5)+3(4)+4(4)+2(2)+4(6)+6(8)+8(3)+3(1)+6(2)+9(5)+12(3) \} / 50 = 4.82$$

$$COVAR (X, Y) = 4.82 - (2.54)(1.82) = 0.1972$$

Alternative calculation to *

$$VAR[X] + VAR[Y] + 2 COVAR [X, Y] = 1.0484 + 0.5876 + 2 (0.1972) = 2.0304$$

15