Overview

- Hypothesis Testing
  - Null Hypothesis (H_0)
  - Statistic
  - Null Distribution
  - P-value

- Parametric Tests
  - Null distribution has a specified form

- Non-Parametric (Distribution-Free) Tests
  - Exact Tests
  - Rank Tests
  - Randomization Tests
    - Permutation Tests
    - Bootstrap Tests

Example: 4 observations sampled from 2 different populations

	Group1	Group2
	11	2
	14	9
	7	0
	8	5
mean	10	4

- H_0: Population means are equal
- Statistic: $S = \overline{X_1} - \overline{X_2}$
  - $S_{observed} = 6$

- **Null (Permutation) Distribution:**
  - **There are 8!/(4!4!) = 70 ways to rearrange the data**
  - **2 of these have a larger value for S than observed**

	Group1	Group2	Group1	Group2
	11	2	11	2
	14	7	14	8
	9	0	9	0
	8	5	7	5
mean	10.5	3.5	10.25	3.75

- **P-value:**
  - **(1-sided) p-value = 3/70 = .0429**
  - **(2-sided) p-value = 6/70 = .0857**

- **Problem: with 15 observations in each of 2 populations there are 155,117,520 arrangements to evaluate!**

Example: Correlation Coefficient between 2 sets of measurements

X	Y
4	3
8	5
2	4
5	7
9	8

- H_0: corr(X,Y) = 0
- Statistic: Pearson's correlation coefficient
  - $r_{observed} = 0.6612$
- Permutation Distribution
  - What gets permuted?

Approximating the Permutation Distribution

- **How many permutations?**

  Let p be the true p-value

$$\hat{p} \pm 1.96\sqrt{p(1-p)}$$ **is a 95% CI for p**

➔ **The number permutations (N) needed to be within +/- .01 of the true p-value is**

$$N \geq (1.96/.01)^2\, p(1-p)$$

p=.5 ➔ N $\simeq$ 10,000

Hardy-Weinberg Proportions (HWP)

Under random mating, with respect to a given locus in a large population, genotype frequencies are expected to be the product of allele frequencies.

Let p_A and p_a be the allele frequencies for *A* and *a*

p_{AA} , p_{Aa} , and p_{aa} be genotype frequencies for *AA, Aa*, and *aa*

HWP ➔ $p_{AA} = p_A^2,\ p_{Aa} = 2p_A p_a,\ p_{aa} = p_a^2$

Testing Fit to HWP

- **Chi-square Goodness-of-fit test**
- **Likelihood Ratio test**
- **Exact test**

Chi-square Goodness-of-fit test

O_i = observed # with the i^{th} genotype
E_i = expected # with the i^{th} genotype, under H_0: HWP

$$X^2_{HW} = \sum (O_i - E_i)^2 / E_i$$

Example:

$$n_{aa} = 26, \; n_{Aa} = 9, \; n_{AA} = 5$$

$$n = n_{aa} + n_{Aa} + n_{AA} = 40$$

Likelihood Ratio test

$$P(n_{AA}, n_{Aa}, n_{aa}) = \frac{n!}{n_{AA}!\, n_{Aa}!\, n_{aa}!} \left(p_{AA}\right)^{n_{AA}} \left(p_{Aa}\right)^{n_{Aa}} \left(p_{aa}\right)^{n_{aa}}$$

The likelihood, L, of the data is just P() considered as a function of (p_{AA} , p_{Aa}, p_{aa}) for a fixed set of observed genotypes, (n_{AA} , n_{Aa}, n_{aa}).

- **L_0 is the likelihood computed under H_0 : HWP**
  - $p_{AA} = p_A^2 = (n_A / 2n) = (2n_{AA} + n_{Aa}) / 2n, \ldots$
- **L_1 is the likelihood computed under H_1: no restrictions on genotype frequencies**
  - $p_{AA} = (n_{AA} / n), \ldots$

$$S = -2 \ln\left(\frac{L_0}{L_1}\right) \sim \chi_1^2$$

Example (continued):

- **All possible arrangements with allele frequencies fixed at the observed values.**

AA	Aa	aa	Probability	Cumulative Probability	
9	1	30	0.0000	0.0000	
8	3	29	0.0000	0.0000	
7	5	28	0.0001	0.0001	
6	7	27	0.0023	0.0024	
5	9	26	0.0205	0.0229	←Observed Data
0	19	21	0.0594	0.0823	
4	11	25	0.0970	0.1793	
1	17	22	0.2308	0.4101	
3	13	24	0.2488	0.6589	
2	15	23	0.3411	1.0000	

Example (continued):

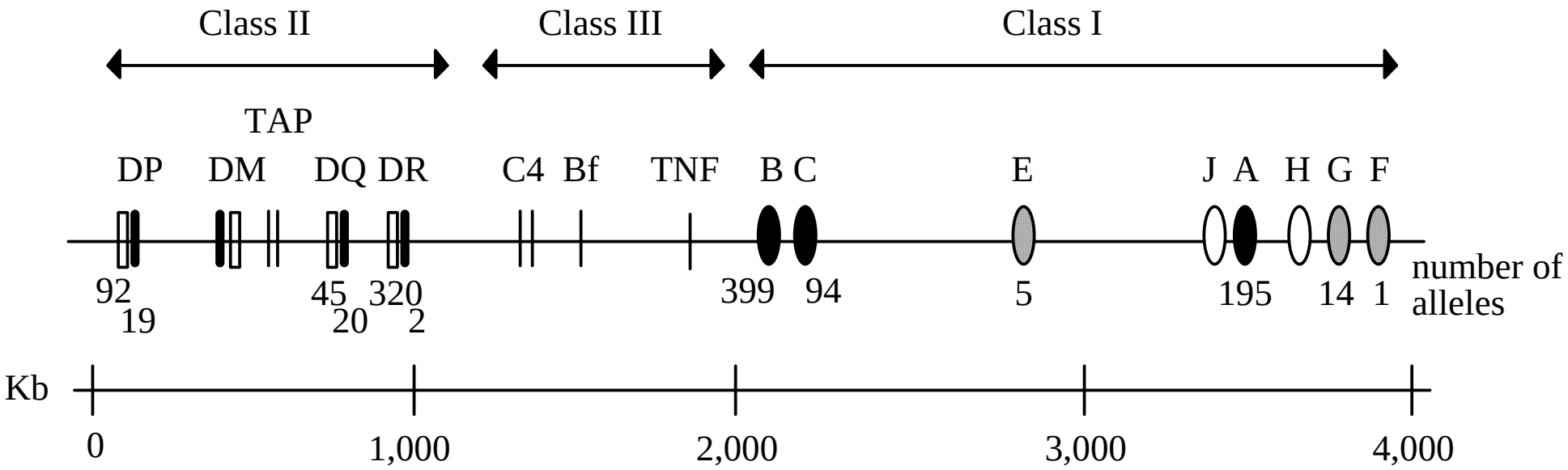
- **Observed table**

	A	a
A	5	
a	9	26
Total	14	26

- **When there are many alleles enumeration of all possible tables is not feasible and permutations may not adequately sample the space of possible tables.**

- **Sampling tables via Markov-Chain Monte Carlo (MCMC)**
  - **Only tables with fixed marginal totals (allele frequencies) are considered.**
  - **Transition probabilities between any two tables are a function of the ratio of probabilities of each.**

Map of the HLA region (Source: Rhodes and Trowsdale 1999)



Class I – ovals:

classical (Ia) filled, non-classical (Ib) stippled, pseudogenes open.

Class II–rectangles:

B genes light, A genes dark.

Allele numbers from http://www.anthonynolan.com/HIG/index.html, Oct. 13, 2000.

HLA VARIATION: EVIDENCE FOR SELECTION

- **There is considerable functional, population, and evolutionary evidence for selection operating in the HLA region.**
- **The actual selective pressures are not known, but heterozygote advantage in response to a wide variety of pathogens is believed to be a major evolutionary force.**

(1) Extensive polymorphism with high variability at ARS
(2) Deviation from Hardy Weinberg
(3) Relatively even allele frequencies
(4) Even amino acid frequencies
(5) Non-synon. aa changes more frequent than synon. in ARS
(6) The great age of alleles
(7) Distinctive linkage disequilibrium patterns such as A1-B8-DR3
(8) Maternal-fetal incompatibility
(9) Mate choice
(10) Direct involvement in autoimmune diseases, cancers, infectious diseases, and others

Bootstrap Version of the Two-Sample *t*-Test

	Group1	Group2
	11	2
	14	9
	7	0
	8	5
mean	10	4

- H_0: Population means are equal
  - Variances not assumed to be equal
- Statistic: $t = (\overline{X_1} - \overline{X_2}) / \sqrt{s_1^2 / n_1 + s_2^2 / n_2}$

Bootstrap Procedure for generating the Null Distribution

- **Create "amended data" compatible with H_0**
  - **Subtract off group mean from observations in each group**
- **Generate a large number, *R*, of bootstrap values for *t***
  - **Take an SRS of n_1 from "amended" group 1 with replacement**
  - **Take an SRS of n_2 from "amended" group 2 with replacement**
  - **Compute $t_{Bootstrap}$**
- $$pvalue = (\#t_{Bootstrap} \geq t_{Observed}) / R$$