

Lecture 21 Normal Approximation To Binomial Distribution & Sampling Distribution of The Sample Proportion

BIO210 Biostatistics

Xi Chen

Fall, 2024

School of Life Sciences

Southern University of Science and Technology

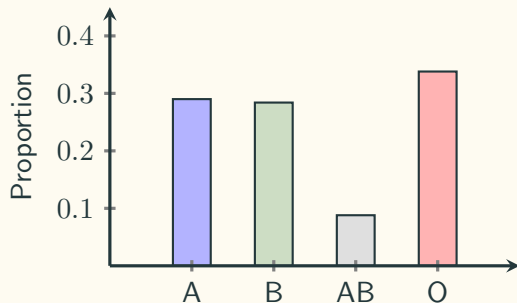


南方科技大学生命科学学院
SUSTech · SCHOOL OF
LIFE SCIENCES

ABO Blood Types Proportions In Han Chinese

Population distribution of ABO blood types in Han Chinese.

Total	A	B	AB	O
592,243	171,473	168,040	52,088	200,642
1	0.290	0.284	0.088	0.338



中国汉族人ABO血型的分布

空军成都医院* 彭德仁

研究 A B O 血型的分布在医学、法医学及人类学等方面都有重要意义。关于国人的 A B O 血型分布早在 1918 年就有人报道过^[1]。现已积累了大量的资料。1963 年, 尚书颂等曾统计分析 15 万多中国人 A B O 血型的分布资料, 提出将我国各省区的 A B O 血型分布划为 4 种类型^[2]。1982 年, 陈稚勇等收集了 1920~1979 年国内外发表的国人的 A B O 血型分布资料共 28 万多例, 通过计算各群体间的遗传距离, 将全国分为 4 个组^[3]。但这两篇文章都包含了少数民族的资料。Mourant 等所著的《人类血型分布》中也仅收集到 18 万多中国人的 A B O 血型分布资料^[4]。由于中国是一个多民族的国家, 就 A B O 血型而言, 不同的民族可有不同的分布特点, 即使是同一民族, 其分布特点因地域等原因也可能不尽相同。为了给医学、法医学及人类学等研究提供一些基本数据, 本文收集了 1920~1988 年国内外发表的有关汉族的 A B O 血型分布资料共 59 万多人, 并对其进行分析。

材料与方法

(一) 资料来源

国内发表的资料主要取自 1963~1966 年的《天津医药杂志输血及血液学附刊》、1978~

1979 年的《输血及血液学》杂志、1980~1988 年的《中华血液学杂志》、1981~1988 年的《中华医学检验杂志》等的 162 篇文献; 国外发表的资料主要取自《人类血型分布》^[4]。所收集的资料仅限于汉族, 每份资料的人数均多于 30 人且注明了居住地区, 全部资料共计 1 022 237 人。

(二) 基因频率的计算与 Hardy-Weinberg 吻合度测验

对所有收集到的有关资料用赵桐茂推荐的方法来计算 A B O 基因频率^[5]。p、q、r 分别代表 A、B、O 基因频率。为了估计调查资料的可靠性, 对每份原始资料均作了显著性测验^[6]。如果 $|D/\delta| \leq 2$, 表示观察值与期望值无显著性差异, 此时 $P \geq 0.05$ (即 Hardy-Weinberg 吻合度测验观察值与期望值吻合度很好); 如果 $|D/\delta| > 2$, 则 $P < 0.05$, 表示观察值与期望值有显著性差异。A B 型的观察值小于期望值, D/δ 为正值, 反之则为负值。所收集到的资料中除去 $|D/\delta| > 2$ 的 81 份外, 最后所选择的 327 份的数据按地区合并, 并根据 Hirschfeld 等提出的 $(A + AB)/(B + AB)$ 公式计算民族指数。但该指数只反映 A 和 B 基因的比例, 并不能反映差异程度^[7]。

(三) 遗传距离

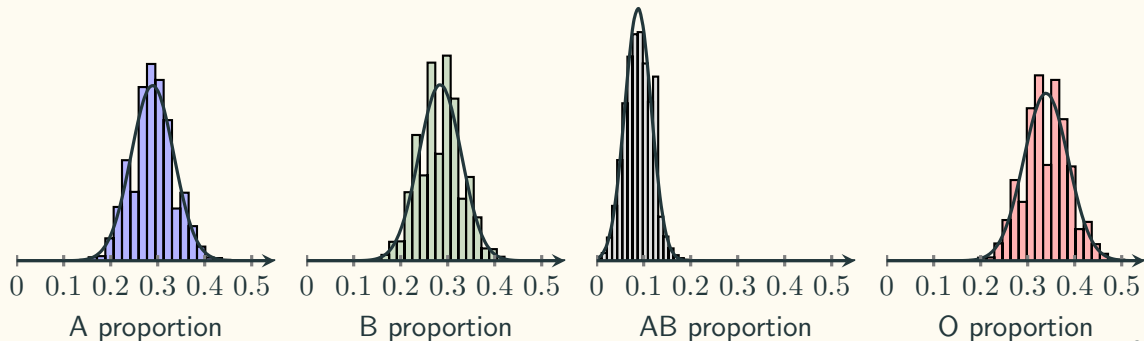
为比较 ABO 血型分布在各地区间的差异, 使用遗传距离 d, 其公式为 $d = 4(1 - \cos \theta)/\pi$

* 邮政编码 610081

Sampling Distribution of ABO Blood Type Proportions

Population distribution of ABO blood types in Han Chinese.

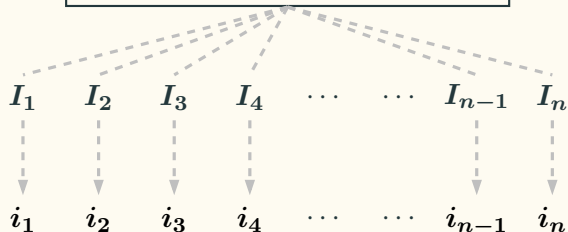
Total	A	B	AB	O
1	0.290	0.284	0.088	0.338



Sampling Distribution of The Sample Proportion

Fraction of type A blood in the population: π

A, A, B, O, O, AB, O, A, O, B, ...



$$Y = \sum_{i=1}^n I_i \quad \bar{I} = \frac{1}{n}Y$$

Dashed lines with arrowheads point from the summation symbol in the first equation to the variable y below it, and from the \bar{I} in the second equation to the variable \bar{i} below it.

...

sample

The Sum And The Mean of Indicator Variables

Meaning of Y : number of people with blood type A per n people.

Meaning of \bar{I} : The proportion of people with blood type A.

I : Indicator Variable

i.i.d.

$$I_1 \sim \text{Ber}(\pi)$$

$$I_2 \sim \text{Ber}(\pi)$$

$$I_3 \sim \text{Ber}(\pi)$$

\vdots

\vdots

$$I_{n-1} \sim \text{Ber}(\pi)$$

$$I_n \sim \text{Ber}(\pi)$$

$$Y = \sum_{i=1}^n I_i \sim ? \quad \bar{I} = \frac{1}{n} Y \sim ?$$

By definition:

$$Y \sim B(n, \pi)$$

By The Central Limit Theorem:

$$\bar{I} \dot{\sim} \mathcal{N} \left(\mu = \pi, \sigma^2 = \frac{\pi(1-\pi)}{n} \right)$$

$$Y = n\bar{I} \dot{\sim} \mathcal{N} (\mu = n\pi, \sigma^2 = n\pi(1-\pi))$$

Normal Approximation To A Binomial Distribution

Our knowledge about Han Chinese (Peng, 1991):

Total	A	B	AB	O
1	0.290	0.284	0.088	0.338

A sample from Wuhan (Xu et al., 2015): 1,188 out of 3,694 people have blood type A.

Questions:

1. When draw a random sample ($n = 3,694$), what is the probability of getting 1,100 – 1,200 people with blood type A?
2. When draw a random sample ($n = 3,694$), what is the probability of getting 1,188 people with blood type A?

Normal Approximation To A Binomial Distribution

Question 1:

Use the Binomial probability :

$$\sum_{k=1100}^{1200} \binom{3694}{k} 0.29^k 0.71^{3694-k} = 0.152949$$

Use the Normal probability :

$$\mathbb{P}(1100 \leq x \leq 1200) = \mathbb{P}\left(\frac{1100 - 1071.26}{27.58} \leq z \leq \frac{1200 - 1071.26}{27.58}\right) = 0.148681$$

Use the Normal probability with **continuity correction** :

$$\mathbb{P}(1100 - 0.5 \leq x \leq 1200 + 0.5) = 0.152923$$

Normal Approximation To A Binomial Distribution

Question 2:

Use the Binomial probability :

$$\binom{3694}{1188} 0.29^{1188} 0.71^{3694-1188} = 2.16 \times 10^{-6}$$

Use the Normal probability with continuity correction :

$$\mathbb{P}(1188 - 0.5 \leq x \leq 1188 + 0.5) = 1.86 \times 10^{-6}$$

Sampling Distribution of The Sample Proportion

- $\bar{I} \sim$ Sampling Distribution of The Sample Proportion
- Generally, we used $p = \frac{x}{n}$ to represent the sample proportion, which is an point estimate for the population parameter π .
- According to the **Central Limit Theorem**, when the sample size n is large enough, we have:

$$\mathbf{P} \dot{\sim} \mathcal{N}(\mu_{\mathbf{P}}, \sigma_{\mathbf{P}}^2), \text{ where } \mu_{\mathbf{P}} = \pi, \sigma_{\mathbf{P}}^2 = \frac{\pi(1 - \pi)}{n}$$

Sampling Distribution of The Sample Proportion

