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

BIO210 Biostatistics

Xi Chen

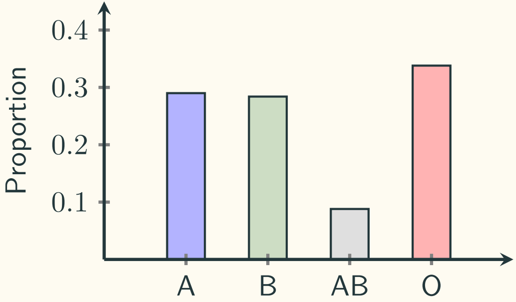
Spring, 2024

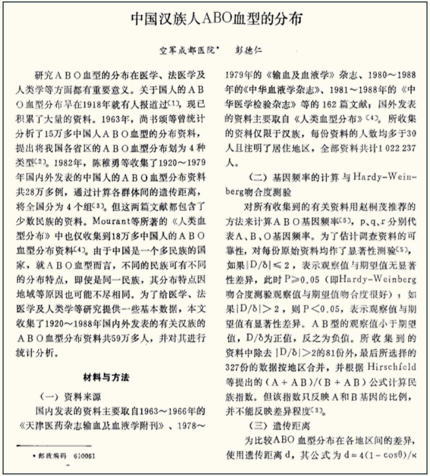
School of Life Sciences
Southern University of Science and Technology

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

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ＡＢＯ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ＡＢＯ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了15万多中国人ＡＢＯ血型的分布资料，提出将我国各省区的ＡＢＯ血型分布划为４种类型[2]。1982年，陈稚勇等收集了1920～1979年国内外发表的中国人的ＡＢＯ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]。但这两篇文献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ＡＢＯ血型分布资料[4]。由于中国是一个多民族的国家，就ＡＢＯ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ＡＢＯ血型分布资料共59万多人，并对其进行统计分析。

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ＡＢＯ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[5]，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。ＡＢ型的观察值小于期望值，D/δ为正值，反之为负值。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[3]。

（三）遗传距离

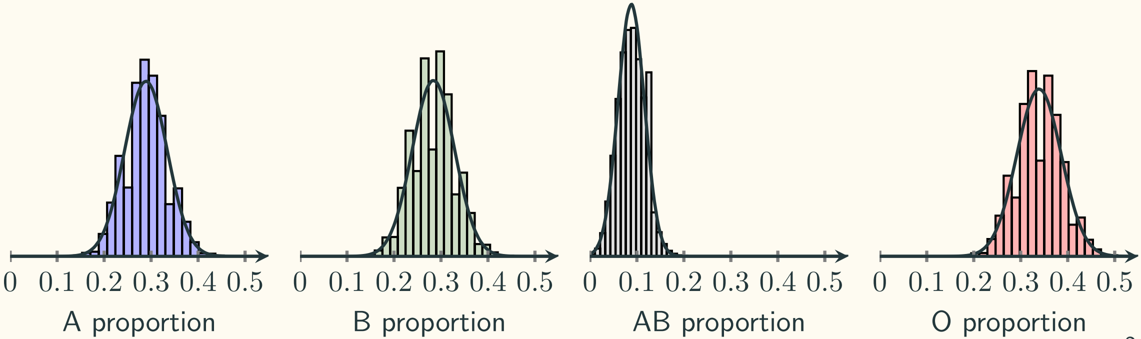
为比较ABO血型分布在各地区间的差异，使用遗传距离d，其公式为$d=4(1-\cos\theta)/\kappa$

* 邮政编码 610061

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

$\boxed{\text{A, A, B, O, O, AB, O, A, O, B, ...}}$

$$
\begin{array}{cccccccccc}
I_1 & I_2 & I_3 & I_4 & \cdots & \cdots & I_{n-1} & I_n & Y = \sum_{i=1}^{n} I_i & \bar{I} = \frac{1}{n} Y \\
\downarrow & \downarrow & \downarrow & \downarrow & & & \downarrow & \downarrow & \downarrow & \downarrow \\
i_1 & i_2 & i_3 & i_4 & \cdots & \cdots & i_{n-1} & i_n & y & \bar{i} \\
1 & 0 & 1 & 1 & \cdots & \cdots & 1 & 0 & 0.3n & 0.3 & \text{sample}
\end{array}
$$

The Sum And The Mean of Indicator Variables

$\boldsymbol{I}$: Indicator Variable

$$i.i.d.$$

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

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

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

$$\vdots$$

$$\vdots$$

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

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

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.

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

By definition:

$$\boldsymbol{Y} \sim B(n, \pi)$$

By The Central Limit Theorem:

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

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

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?

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 :

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

Use the Normal probability with continuity correction :

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

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 :

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

- $\bar{I} \sim$ **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:

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

Sampling Distribution of The Sample Proportion