人口数据质量评估与人口预测方法

陈华帅

2011年10月27日

人口数据质量评估与调整

人口数据误差的类型

1. 抽样误差

抽样误差（ sampling error）主要指样本设计与样本规模的选择本身，以及变量本身的随机波动性等原因造成的误差。

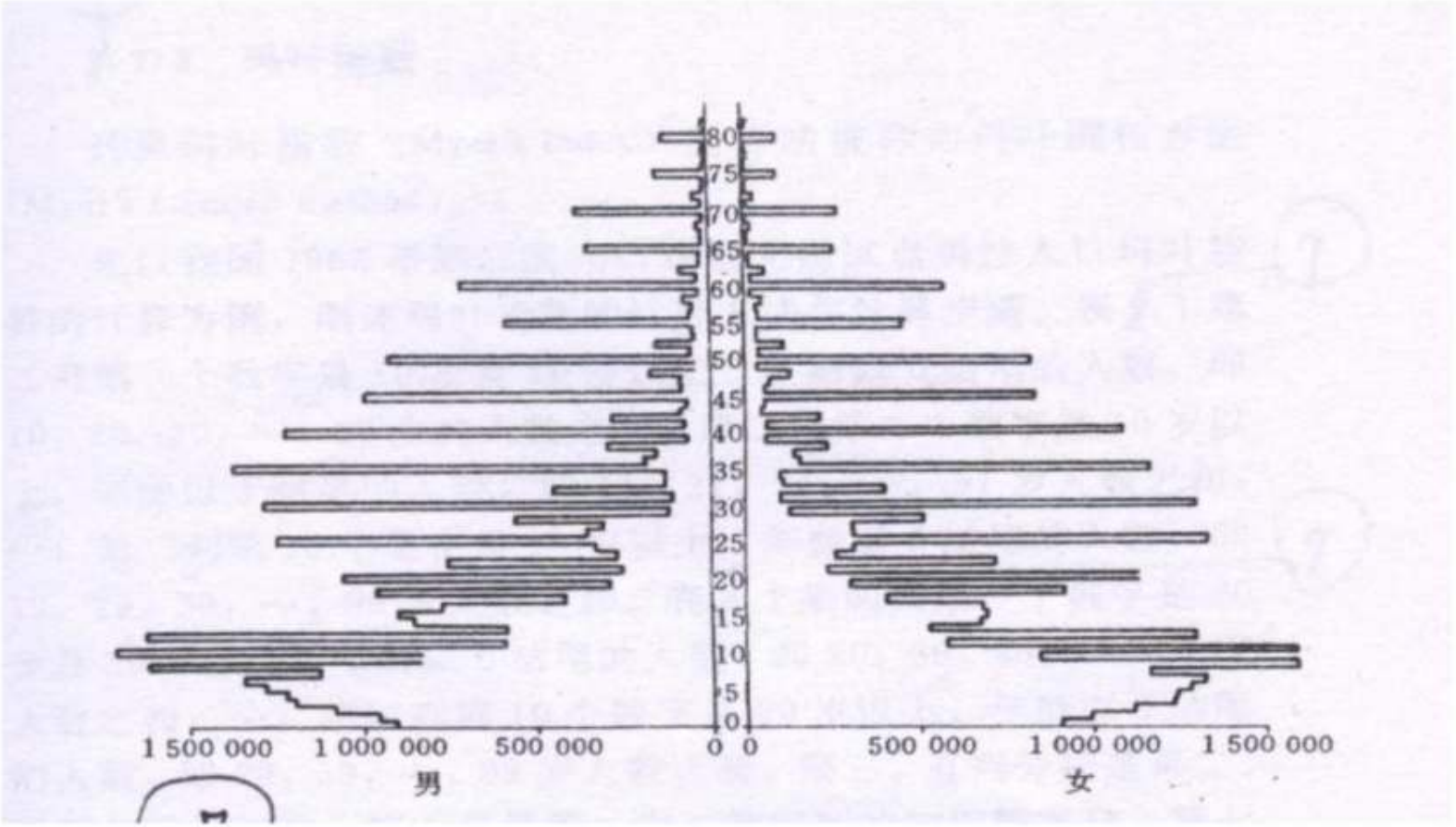
2. 系统误差

- 结构性 误差 。例如， 只能搜集调查时点存活者的现状与历史数据，无法搜集调查时点前已经死亡者的详细数据。
- 覆盖面误差 。一般 主要指人口普查未能收集到全体人口的数据资料。在抽样调查中，也可能遗漏整个局部地区或整个局部人群（组），有的流动或出差人口又可能同被原住地与流动或出差地重复登记。
- 申报 内容 误差 。 主要有年龄误报、人口事件（如婚姻、婚姻解体、生育避孕、流产、迁移、死亡等）数的漏报、人口事件发生时间的误报、性生活频率及收入等较为敏感问题申报的不实等。
- 数据录入误差

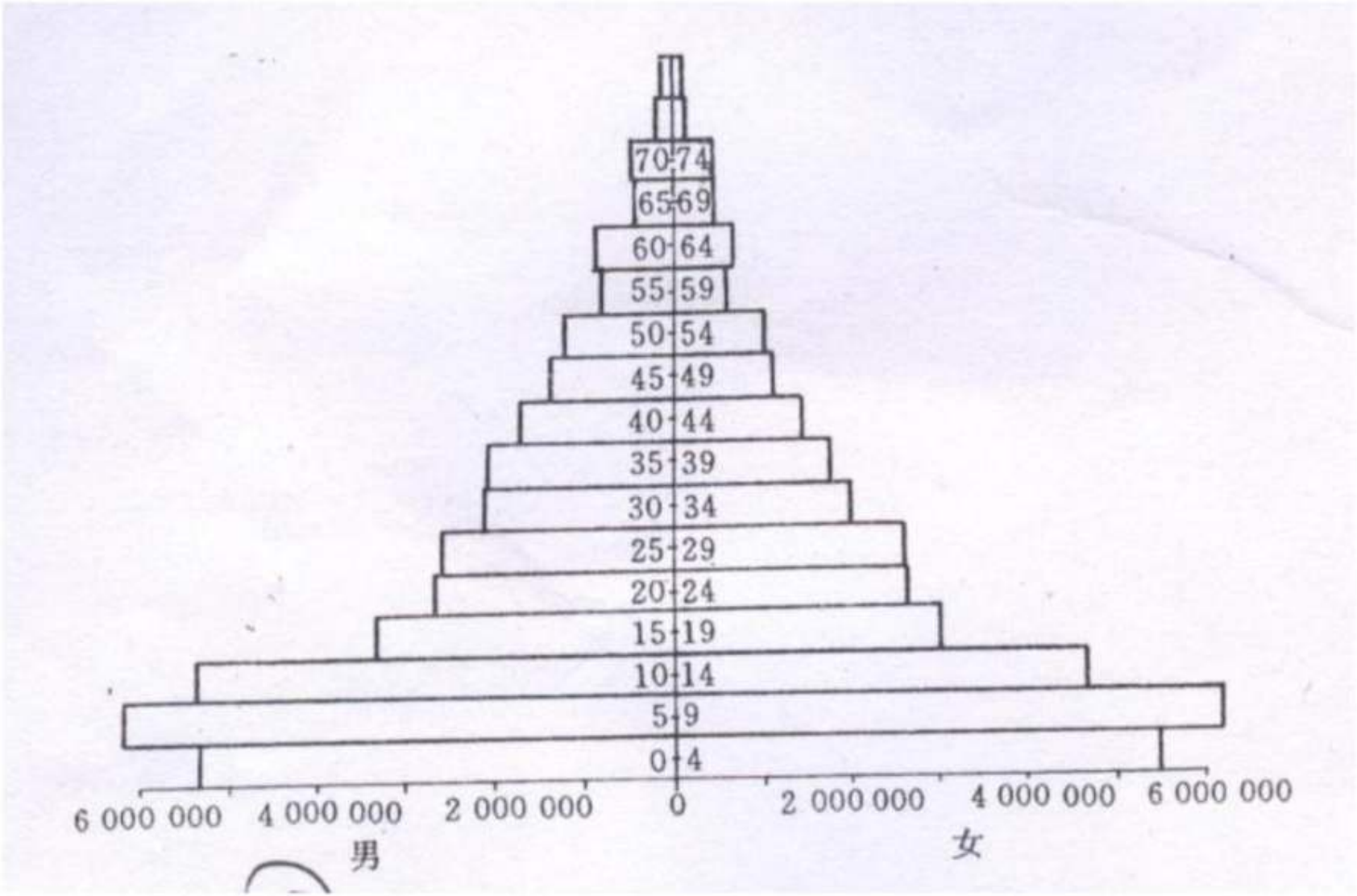
年龄申报误差及其评估指标与方法

年龄申报误差来源及类型：

数字偏好（digit preference）； “年龄堆积”（age heaping）。



孟加拉国1974年以及0与5结尾的年龄堆积现象



孟加拉国1974年5岁分组年龄构成
（以0与5岁结尾的年龄堆积现象基本消失）

完全 年龄 (completed age)：也称为上一次生日的年龄 (age last birthday)。

虚岁年龄：实际X周岁的人可能被申报为虚岁X+1岁，甚至x+2。但是，由于东方人对出生年月记得十分清楚，如果在普查或调查中询问出生年月与其它人口事件发生年月，而不是询问年龄，这种虚岁误报现象即可避免。

由于某些原因，一些人倾向于虚报自己的年龄。例如，在尊敬老人的社会里，人们以高龄为荣，因而夸大年龄。

例如，我国1982年第三次人口普查给出全国百岁以上老人中，一半以上在新疆。这显然与新疆的少数民族老人夸大年龄有关。

年龄申报质量：玛叶指数（Myer's index）

示例：我国1982年第三次人口普查无锡男性人口玛叶指数的计算

年龄结尾数a（1）	以a数结尾的人口数				调和人口数		与10%之差绝对值（8）
	自10+a起算（2）	权算（3）	自20+a起算（4）	权算（5）	（2）×（3）+（4）×（5）	百分比（7）	
0	37505	1	31110	9	317495	10.47	0.47
1	34041	2	27336	8	286770	9.46	-0.54
2	35444	3	29175	7	310557	10.24	0.24
3	32286	4	27146	6	292020	9.63	-0.37
4	32095	5	25526	5	288105	9.50	0.50
5	34626	6	27540	4	317916	70.48	0.48
6	35708	7	25291	3	325829	10.75	0.75
7	34187	8	23696	2	320774	10.58	0.58
8	29135	9	23407	1	285622	9.42	-0.58
9	28714	10	21799	0	287140	9.47	-0.53
合计					3032228	100.00	5.04

玛叶指数=5.04÷2=2.52

玛叶指数计算步骤（上表）

- 第二列第一个数字是10岁及10岁以上，年龄以0结尾的人数，即10，20，30，…90岁的人数之和；第二列第二个数字是10岁以上，年龄以1结尾的人数，即11，21，31，…91岁人数之和；…；
- 第四列第一个数字是20岁及20岁以上，年龄以0结尾的人数，即20，30，40，…，90岁人数之和；…；第四列第10个数字是20岁以上，年龄以9结尾的人数，即29，39，…，99岁人数之和。
- 第三、五列分别是第二、四列各数的权数。
- 第六列是2、4列的加权数求和，第七列是第六列各数字占该列总数的百分比。
- 如果人口的年龄申报绝对准确，且人口年龄构成未受到过去年份不正常的出生率、死亡率与迁移率大起大落的影响，第七列各数字应十分接近于10%。与10%差异越大，说明年龄申报越不准确（当然也包含过去年份出生、死亡或迁移大起大落的影响），
- 第八列给出的是第七列各数与10%之差的绝对值。
- 第八列的和除以2即为玛叶指数。显然玛叶指数的取值范围是0至90。
- 当玛叶指数为0时，表示年龄数据不但无任何误报堆积现象，而且完全符合关于生育、死亡、迁移在过去很长时期内稳定不变，无不规则波动的理论假设。
- 当玛叶指数为90时，说明所有人都申报为以同一数字结尾的年龄，这当然是不可能出现的。
- 一般来说，玛叶指数小于10的人口年龄申报质量较好；玛叶指数大于20的人口年龄申报质量被认为很差，是不可接受的，必须予以调整后才能使用。而玛叶指数在10到20之间被认为基本可以接受，其五岁组年龄数据是可用的，但单岁年龄数据则不一定完全可用。

年龄申报质量：韦伯指数（Whipple's index）

- 目的：为量测这种对以0与5结尾的年龄偏好与堆积而设计的。
- 只利用23至62岁的年龄数字来计算韦伯指数。

- 如果一个人口的年龄构造不受过去年份生育、死亡、迁移大起大落影响；当没有任何0岁与5岁的年龄偏好造成的堆积时，韦伯指数应为100。当所有人都申报以0或5结尾的年龄时，韦伯指数为500。

联合国关于年龄申报质量的评价标准

质量	韦伯指数
很准确	<105
较准确	105—110
大致合格	110—125
较差	125—175
很差	>175

玛叶指数与**韦**伯指数**优**点：

- 清楚明了，易于计算，便于比较

玛叶指数与**韦**伯指数缺点：

- 当人口年龄构成受到过去生育高峰、战争、灾荒或迁移高峰影响时，玛叶指数值会受到影响而不能客观反映年龄申报的准确程度。

判别人口事件数的漏报及事件

- 人口事件数的漏 **报**
  - ✓ 年老妇女漏报曾生子女数相对严重。他们往往有意或无意遗忘已经死亡或已经离家独立生活，特别是在外地的子女。
  - ✓ 由于重男轻女观念及计划生育影响，出现女孩漏报，致使出生性别比过高。
- 人口事件**发生时间**的**错报**
  - ✓ 错报人口事件发生时间可能因为被调查者记忆不清或与其它事件发生时间混淆而造成。若问：“您什么时候生第一个孩子？”有的妇女回答“5年以前”。而5年以前可能是5年多甚至6、7年，或不到5年，甚至4年，被调查者估摸着“大约是5年吧”。因此，这个问题应改为“您是何年何月生第一个孩子的？”，以避免上述误差。

评估人口普查数据质量

1. 人口分析法（demographic analysis）。

- 普查数据本身的内部一致性检验；
  - 例如：男性有配偶总人数应等于女性有配偶总人数；普查时点前一年内出生的人数减去前一年内0岁死亡人数应等于登记到的0岁人数；甲地迁往乙地的人数应等于乙地从甲地迁入的人数；
- 与按人口分析得到的期望值的比较；
  - 例如：1990年粗出生率不会比1988年大幅度下降，如果实际上得到了这种大幅度下降的粗出生率，可能出现了漏报瞒报出生数。
- 与其他来源数据的比 较

评估人口普查数据质量

2. 事后质量抽查（post enumeration sampling）

3. 匹配研究（Matching study）

	登记了户籍	未登记户籍
普查时被登记	C	N_2
普查时未被登记	N_1	X

实际总人数（或事件数）$N=C+N_1+N_2+X$

普查登记的完整率=（$C+N_2$）/N

户籍（或事后抽查）登记的完整率=（$C+N_1$）/N_1