《医学统计学》第一次课

绪论

医学统计学

（Medical Statistics）

Department of Health Statistics

“不明于计数，而欲举大事，犹无舟楫而欲经于水险也。”

--管仲

“统计根据科学方法，以繁赜事实，列记之，整理之；从大量，得共相；由已知，测未知；使樊然者，呈整齐之相；棼如者，有迹可循；复杂变量，烛照无遗；纷纭事态，一目了然，此其功也。”

--陈善林

“(统计学)是追求科学的人从荆棘丛生的困难阻挡中，开辟出道路的最好的工具。”

--Francis Galton

第一章 绪论

Charpter1：Introduction

“数字虽然枯燥，但却能说明问题”

统计学是一门与数据有关或研究数据的学科。

Department of Health Statistics

一、统计学与医学统计学方法

***Statistics*：** “A science dealing with the collection, analysis, interpretation and presentation of masses of numerical data”

----Webster 国际大辞典

统计学是一门收集、分析、解释与呈现数据资料的科学。

***统计学*（statistics）**

运用概率论、数理统计的原理和方法，研究数据的收集、整理、分析和推断的科学。

***医学统计学*（medical statistics）**

用统计学原理和方法研究生物医学问题的一门学科。

***生物统计学*（biostatistics）**

应用于整个生物学范畴，侧重于人的生物方面。

***卫生统计学*（health statistics）**

用于医学和卫生学领域，侧重于人的社会方面，如健康状况统计和卫生服务统计。

20世纪20年代，英国统计学家 R.A. Fisher爵士(1890-1962)创立了实验设计方法和统计分析技术，奠定现代生物统计的基础。

1948年，英国发表了评价链霉素治疗肺结核疗效的随机对照的临床试验报告，第一次采用生物统计方法进行临床干预试验。

1948年，郭祖超教授(1912~1999)编著的《医学与生物统计方法》，是我国第一部医学统计方法的教科书。

二、医学统计学中常用的数据类型

1. 计数资料

2. 计量资料

3. 等级资料

计数资料(count data):

变量值表现为按某属性划分的定性类别，清点各类别个数后得到的资料。

数据表现:

两分类: 阳性或阴性; 存活或死亡;
有效或无效; 男性或女性

多分类: 如血型: O，A，B，AB

计数资料（血型） 100 52 78 32

计量资料(measurement data):

用仪器、工具或其他定量方法准确获得的定量结果，一般带有计量单位。

(1)连续变量:

身高(cm): 1.65，1.70 1.58，…

体重(kg): 52，55，61，...

(2)比率变量

脑电图波形变化率(%): 29%，37%，...

等级资料(ordinal data)：

变量值按变化程度大小划分分类，清点各分类的个数后得到的资料。

例：病情分级(X_1)：Ⅰ，Ⅱ，Ⅲ

疗效(X_2)：痊愈、显效、有效、无效

病人满意度(X_3)：好、中、差

人数：50 25 5

观察单位 observations

个体individuals

变量 variables

住院号	年龄	身高	体重	住院天数	职业	文化程度	分娩方式	妊娠结局
2025655	27	165	71.5	5	无	中学	顺产	足月
2025654	22	160	74.0	5	无	小学	助产	足月
2025830	25	158	68.0	6	管理员	大学	顺产	足月
2022543	23	161	69.0	5	无	中学	剖宫产	足月
2022466	25	159	62.0	11	商业	中学	剖宫产	足月
2024535	27	157	68.0	2	无	小学	顺产	早产
2025834	20	158	66.0	4	无	中学	助产	早产
2019464	24	158	70.5	3	无	中学	助产	足月
2025783	29	154	57.0	7	干部	中学	剖宫产	足月

Measurement data 计量资料

Ordinal data 等级资料

Count data 计数资料

实例数据1 胆管癌患者部分指标

编号	性别	年龄（岁）	部位	分化程度	分期	肝转移	PCNA 指数	生存时间(月)
（1）	（2）	（3）	（4）	（5）	（6）	（7）	（8）	（9）
1	男	61	上	低分化	Ⅰ	阳性	52	14
2	女	58	中	高分化	Ⅱ	阴性	89	20
3	女	63	上	高分化	Ⅳ	阴性	93	19
4	女	71	下	中分化	Ⅱ	阳性	78	5
5	男	59	上	高分化	Ⅲ	阴性	85	35
…	…	…	…	…	…	…	…	…

实例数据2

体重指数	身高	班制	劳动强度	紧张程度	心率	嗜肥肉史	收缩压	舒张压	中风家族史
（1）	（2）	（3）	（4）	（5）	（6）	（7）	（8）	（9）	（10）
12.24	1.62	1	1	3	70	1	146	90	有
16.47	1.63	3	1	3	72	0	110	70	无
15.19	1.64	1	2	2	72	0	100	70	无
15.59	1.63	1	1	3	84	1	114	70	无
12.60	1.64	3	1	3	68	1	116	68	无
…	…	…	…	…	…	…	…	…	…

实例数据3

100例高血压患者治疗后临床记录

患者编号	年龄 X_1	性别 X_2	治疗组 X_3	舒张压 X_4	体温 X_5	疗效 X_6
1	37	男	A	11.27	37.5	显效
2	45	女	B	12.53	37.0	有效
3	43	男	A	10.93	36.5	有效
4	59	女	B	14.67	37.8	无效
…	…	…	…	…	…	…
100	54	男	B	16.80	37.6	无效

三种数据类型之间的转换

例：一组20~40岁成年人的血压

等级资料

<8	低血压
8~	正常血压
12~	轻度高血压
15~	中度高血压
17~	重度高血压

计量资料

计数资料

以12kPa为界分为正常与异常两组，统计每组例数

三、统计学基本概念

（一）随机变量(random variable)

简称变量(variable)，统计上习惯用大写拉丁字母表示，如*X*、*Y*、*Z*、…。

编号（ID）	性别（X）	体重（kg）（Y）	疗效（Z）
张1	1	66	0
李2	1	78	1
王3	0	57	2
…	…	…	…

随机变量的分类

- 离散型变量(discrete variable)，在一定区间变量取值为有限个，相当于计数资料
- 连续性变量(continuous variable)，在一定区间变量取值为无限个，相当于计量资料
- 有序变量(ordinal variable)，相当于等级资料

（二）同质与变异(homogeneity and variation)

同质：指事物的性质、影响条件或背景相同或非常相近。

变异：指同质的个体之间的差异。

同质与变异的例子

例1　调查2004年上海市7岁男童的身高和体重

同质：2004年、上海市、7岁男童

变异：身高和体重各不相同

例2　研究某降压药的疗效

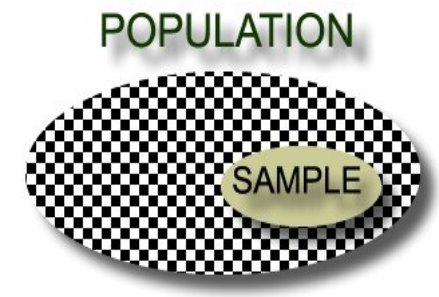
同质：高血压患者、用某药治疗

变异：疗效各不相同

（三）总体与样本(population and sample)

总体：根据研究目的确定的**同质**研究对象的**全体**(集合)。

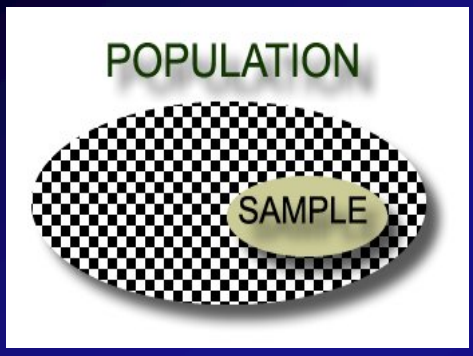
分有限总体与无限总体。



有限总体：总体中的所有研究对象事先可确定而且有限时，为一个确定的总体。

无限总体：总体中的所有研究对象事先无法全部获得而且无限时，为一个假设的总体。

（三）总体与样本(population and sample)



样本：

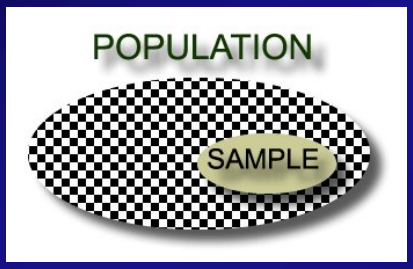
从总体中随机抽取有代表性的部分观察单位

随机抽样(random sampling)

为了保证样本的**可靠性**和**代表性**，需要采用随机的抽样方法（在总体中每个个体具有**相同的机会**被抽到）。

抽样研究(sampling research):

从总体中随机抽取样本，根据样本信息来推断总体特征的方法，即抽样研究。



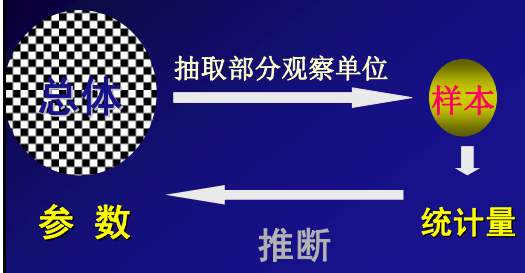
抽样研究目的:

样本观察值结论

↓

推论总体的情况

（四）参数与统计量 (parameter and statistic)



参数：总体的统计指标，如总体均数、标准差，采用希腊字母分别记为 μ、σ。固定的常数。

统计量：样本的统计指标，如样本均数、标准差，采用拉丁字母分别记为 $\bar{X}$、S。参数附近波动的随机变量。

（五）误差

误差(error)：指实测值与真值之差、样本指标与总体指标之差。

误差的类型：

1.随机误差
2.非随机误差

} 反映数据的质量

误差（Error）

	原 因	可避免否
随机误差	多种尚无法控制的因素所引起的数据变异，医学中主要是个体差异和测量误差	不能
非随机误差	过失误差：如记录、操作等人为因素。 系统误差：病例选择、仪器、方法的不一致	可控制而缩小或消除

抽样误差(sampling error):

由于样本的随机性引起的统计量与参数的差别，或同一总体的相同统计量之间的差别。

（六）频率与概率(frequency and probability)

确定性现象：在一定条件下，一定会发生或一定不会发生的现象。其表现结果为两种事件：肯定发生某种结果的叫必然事件；肯定不发生某种结果的叫不可能事件。

随机现象：在同样条件下可能会出现两种或多种结果，究竟会发生哪种结果，事先不能确定。其表现结果称为随机事件。 随机事件的特征：①随机性；②规律性：每次发生的可能性的大小是确定的。

概率：随机事件发生的可能性大小，用大写的***P***表示；取值**[0，1]**。

小概率事件

必然事件　　$P = 1$

不可能事件　　$P = 0$

随机事件　　$0 < P < 1$

$P \leqslant 0.05$（5%）称为**小概率事件**(习惯)，统计学上认为不大可能发生。

频率(frequency)：样本的实际发生率称为频率。设在相同条件下，独立重复进行n次试验，事件A出现f次，则事件A出现的频率为f / n。

实验者	投掷次数	出现“正面”的次数	频率
Buffon	4040	2048	0.5069
K.Pearson	12000	6019	0.5016
K.Pearson	24000	12012	0.5005

频率与概率间的关系：

- 样本频率总是围绕概率上下波动
- 样本含量*n*越大，频率的波动幅度越小，频率越接近概率。

用途：

- 医学中常用频率作为某事件概率的估计值。
- 统计推断结论是基于一定可信程度下的概率推断。

四、统计工作的基本步骤

根据研究目的

1.研究设计	2.收集资料	3.整理资料	4.分析资料
研究对象、指标、例数、如何准确得到数据	日常工作记录、病历。专门的调查和实验	使资料系统化，便于进一步统计分析	用统计方法分析资料，阐述规律性，得出结论

五、统计在医学中的应用

（一）变异的描述

数据的变异特征:

1. 数据的分布范围

2. 数据的分布规律

120名正常成年男子红细胞计数值(10^{12}/L)

5.12	5.13	4.58	4.31	4.09	4.41	4.33	4.58	4.24	5.45	4.32	4.84
4.91	5.14	5.25	4.89	4.79	4.90	5.09	4.64	5.14	5.46	4.66	4.20
4.21	3.73	5.17	5.79	5.46	4.49	4.85	5.28	4.78	4.32	4.94	5.21
4.68	5.09	4.68	4.91	5.13	5.26	3.84	4.17	4.56	3.52	6.00	4.05
4.92	4.87	4.28	4.46	5.03	5.69	5.25	4.56	5.53	4.58	4.86	4.97
4.70	4.28	4.37	5.33	4.78	4.75	5.39	5.27	4.89	6.18	4.13	5.22
4.44	4.13	4.43	4.02	5.86	5.12	5.36	3.86	4.68	5.48	5.31	4.53
4.83	4.11	3.29	4.18	4.13	4.06	3.42	4.68	4.52	5.19	3.70	5.51
4.64	4.92	4.93	4.90	3.92	5.04	4.70	4.54	3.95	4.40	4.31	3.77
4.16	4.58	5.35	3.71	5.27	4.52	5.21	4.37	4.80	4.75	3.86	5.69

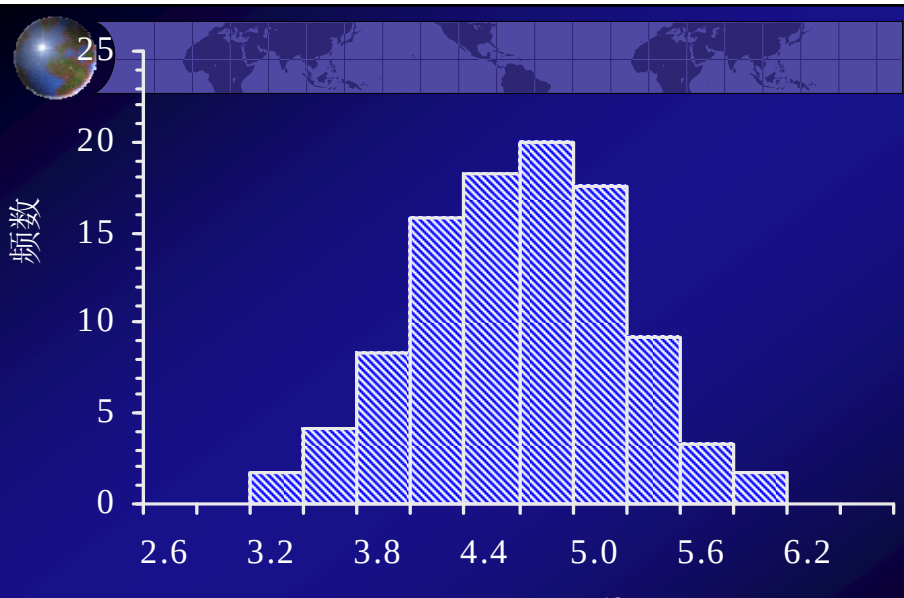


图 1-1 120名正常成年男子红细胞计数直方图

（二）观察对比

- **Farr发现地势与霍乱的关系**

当1848-1849年伦敦地区霍乱流行时，统计学家Farr对观察结果进行数据分析，发现居民居住的地势越高，霍乱的死亡率越低——“瘴气”学说(混浊空气致病)。

（二）观察对比

- Snow发现自来水污染与霍乱死亡人数的关系

1853-1854 年伦敦霍乱死亡率

水　　源	用户数	死亡人数	死亡率（1/万户）
重污染(Southwark 和 Vauxhall 公司)	40046	1263	315.4
轻污染 (Lambeth 公司)	26107	98	37.5
伦敦其它地区	256423	1422	55.5
合　　计	322576	2783	86.3

（二）观察对比

- Doll与Hill发现的吸烟与肺癌的关系

1964年Doll和Hill将接受调查的4万名英国注册医生分为吸烟和不吸烟两组，通过以后的肿瘤统计发现，吸烟组和不吸烟组肺癌的年平均发病率分别为1.66% 和0.07%，相对危险度RR=1.66/0.07=23.7，强烈提示吸烟的致癌作用，为吸烟致癌更深入的研究提供了形成研究假说的基础。

两组人群发病率比较

英格兰和威尔士男性与移民男性的发病率(1/10 万)

年龄分组	英格兰和威尔士			移民		
	人口(千人)	发病数	发病率 (1/10 万)	人口(千人)	发病数	发病率 (1/10 万)
0~4	1900	1406	74.0	26	21	80.8
5~14	3100	186	6.0	30	2	6.7
15~44	9400	1786	19.0	127	27	21.3
45~64	4900	7350	150.0	25	42	168.0
≥65	2000	17400	870.0	5	48	960.0
合 计	21300	28128	132.1	213	140	65.7

（三）实验性研究

- 英国海军军医Lind治疗牙龈溃烂的研究

1747年5月20日，Lind医生将12名病情相似的患者带到一艘船上。患者的主要症状是牙龈溃烂，皮肤有出血点，双膝无力。Lind将12名患者分为6组，分别给予A、 B、 C 、D、 E、 F组干预。

6月16日船返回英国时，所有患者的病情都有好转，其中E组恢复的最快、最好，其中一人到第6天就可以工作了。B组也有一人比登船时健康。

（三）实验性研究

- 法国医生Louis对放血疗法治疗肺炎的研究

1835年，Louis对当时流行的“放血”疗法治疗肺炎的效果进行了比较，发现“放血”的疗效不如预期的那么好，而且早期“放血”和晚期“放血”组比较，患者的诊断、病情、病程、年龄等方面的差异很大，比较平均治愈时间意义不大。因为晚期“放血”组的平均治愈时间长，但该组患者病情重、病程长、年龄大。

（三）实验性研究

- 英国评价链霉素治疗肺结核疗效的随机对照试验

1948年，英国首次采用随机对照试验评价链霉素治疗肺结核疗效。研究对象是15~30岁肺双侧进行性肺结核患者，共抽取了107例患者。试验时将患者随机分为两组，分到试验组的55例患者用链霉素治疗，分到对照组的52例患者用常规方法。治疗6个月后，试验组、对照组的生存率分别为93%、73%。

六、医学论文中的统计

- 错误统计方法产生的严重后果

胃冰冻疗法治疗胃溃疡

“非常痛心地看到，因为数据分析的缺陷和错误，那么多好的生物研究工作面临着被葬送的危险”。

——Yates和Healy

- 统计方法的应用情况

国外60-80年代调查显示，不同医学杂志发表论文有统计错误的最高达72%，最低也有20%。

1966年，据对美国医师协会杂志（JAMA）等医学杂志的来稿的统计显示，149篇投稿论文中，仅有28%可以接受，67%有统计缺陷但尚可以纠正，5%不可救药。

“调查结果反映了医学论文作者统计知识和统计水平的低下，也再次强调了生物统计学者不是令人生畏的检查官。恰恰相反，生物统计学者是我们的可贵盟友。生物统计学不是远离我们的数学，而是现代医学的一门基本学科，就像大厦中的一个支柱”。

——美国医师协会杂志编辑

- 统计数据造假问题

1976年New Science 杂志关于科研舞弊行为的调查，74%的调查表反映有不正当修改数据的情况。其中，17%拼凑实验结果，7%凭空捏造数据，2%故意曲解结果。

- 舍恩(Schon)事件—贝尔实验室
- 两种超重元素发现造假事件—劳伦斯·伯克利实验室
- 韩国克隆之父黄禹锡的人类胚胎干细胞研究造假事件（2005年）

统计学的未来

社会需要生物统计学

“统计学的年轻人有足够的就业机会，学术界、工业部门、政府机构”。

“在美国，很高比例的统计研究生出生于外国，毕业后留在美国”。

(National Science Foundation (1998) report 98-95)

统计学的未来

“对生物统计学的需求从来没有象今天这样大，特别是美国。美国 **National Research Council** 研究结论: 在所有的卫生科研行业，最最缺乏生物统计学和流行病学人才”。（**Zelen)**