



成都中医药大学
CHENGDU UNIVERSITY OF TCM

医学统计学

2021级九年制中医学（李斯炽班）

授课人：张杨、康娟娟

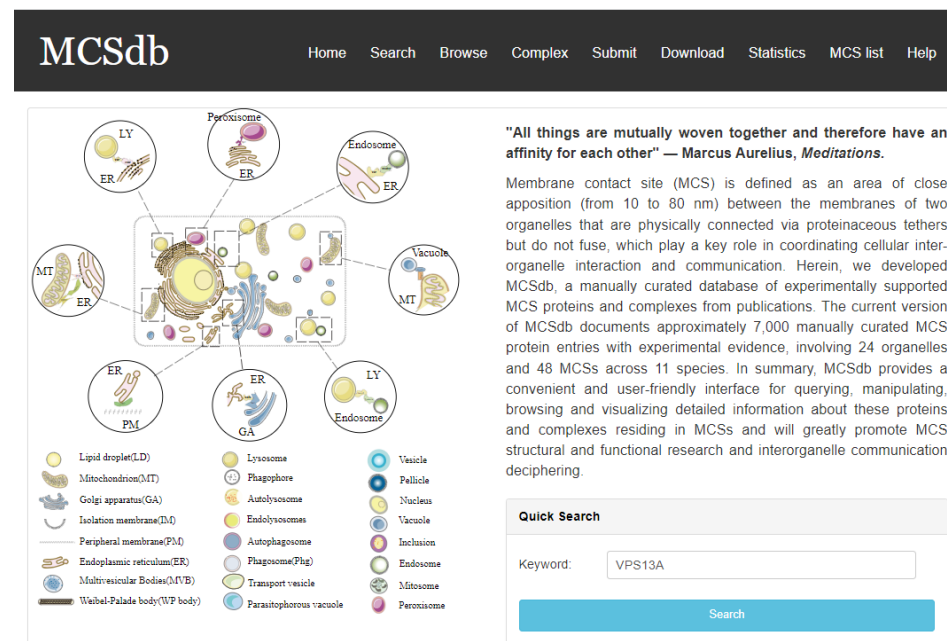


人民卫生出版社
PEOPLE'S MEDICAL PUBLISHING HOUSE

- 张杨 特聘副教授 中医药创新研究院
- 研究方向：生物信息学
- 关键词：数据挖掘、机器学习、人工智能、数据库

<https://www.cellknowledge.com.cn/mcsdb/>

- 电话：13308075867
- 邮箱：yangzhang@cdutcm.edu.cn
- 实验室网站：<https://www.zhangy-lab.cn/>



第一作者: 医学技术学院 生物科学2020级 潘贤润

课程安排（必修课）

- 理论36学时

共计十五章（张杨：1-3章、康娟娟：4-15章）

教学要求

1) 仔细阅读教材，做到**课前预习**、**课后复习**

2) 认真、按时完成并提交**作业**

3) **课堂准备：**

教材

草稿纸、笔

教学资源

- 教材、PPT
- 慕课资源
- B站资源

考核办法

• 形成性考核

- 考勤—20%
- 课堂表现—10%
- 作业—10%

终结性考核

- 期末考试—60%



第一章

绪 论

授课人：张杨



人民卫生出版社
PEOPLE'S MEDICAL PUBLISHING HOUSE

目录

第一节 医学统计学的作用

第二节 医学统计学的基本内容

第三节 医学统计学中的基本概念



学习要求

掌握

统计数据的三种基本类型：定量数据、定性数据和有序数据。
基本概念：同质与变异、总体与样本、系统误差和随机误差。

熟悉

医学统计学研究的对象、内容和作用。统计描述和统计推断的核心问题。

第三节

第一、二节





第一节

医学统计学的作用



人民卫生出版社
PEOPLE'S MEDICAL PUBLISHING HOUSE

1. 医学统计学的定义

医学统计学(medical statistics): 临床医学、基础医学、公共卫生学和医疗卫生服务研究中的一门基础学科，是关于**收集数据、分析数据和由数据得出结论**的一组**概念、原则和方法**。

- **医学统计学：用概率论和数理统计的原理，结合医学实践，研究医学资料的收集、整理、分析和推断。**

统计学的重要作用在于能够透过偶然现象来探测其规律性，使研究结论具有科学性。

- 具体来讲，统计学讲究如何从一定范围的**样本**中总结规律并且推广到更大范围（即**总体**），统计学的技术正是从总体中进行恰当的**抽样**，对样本进行试验、观察，总结归纳，**形成关于总体的结论**。

2. 医学统计学的作用

- **案例1：**1962年美国医学会杂志（JAMA）曾发表了一篇关于胃溃疡治疗新技术的报告，该报告根据动物实验和24名患者的临床试验结果得出结论，即将冷冻液导入胃中使胃冷却可以缓解溃疡症状，之后这一研究成果在临床中被广泛应用。后证实这种方法无效甚至有害。
- **案例2：**20世纪80年代早期，两项观察性研究结果提示孕妇在怀孕期间补充维生素可以降低新生儿神经管缺陷（NTD）的风险，但一直无法证实。该项研究使用统计学方法确定了服用叶酸组与对照组的差别不是简单偶然出现的，而是归因于叶酸的作用。



☑ 问题：

- 胃溃疡治疗新技术的研究结论为什么会出现错误？
- 为了证明服用叶酸的作用，应如何进行分组？
- 如何准确地估计两组出现神经管缺陷的发病率？
- 如何比较两组NTD的发病率是否有差别？

上述问题的解决，都需要运用统计学设计**原则**和统计学**分析方法**

总结

分析数据并得出结论

- 医学统计学的**研究对象**是医学中具有不确定性结果的事物，其主要作用是**通过数据的偶然性揭示医学研究内在的规律性**。
- 统计分析的要点是**正确地选用统计学分析方法**，并结合医学专业知识和研究目的**做出科学的结论**。

本门课程要求

◆掌握概念、原则、方法是首位

◆重在理解和解释结果

通过计算，体验统计概念和思想
正确理解、解释和表达计算结果

最关键是正确的理解公式、能够解释和表达计算结果（解读）。



第二节

医学统计学的基本内容



人民卫生出版社
PEOPLE'S MEDICAL PUBLISHING HOUSE

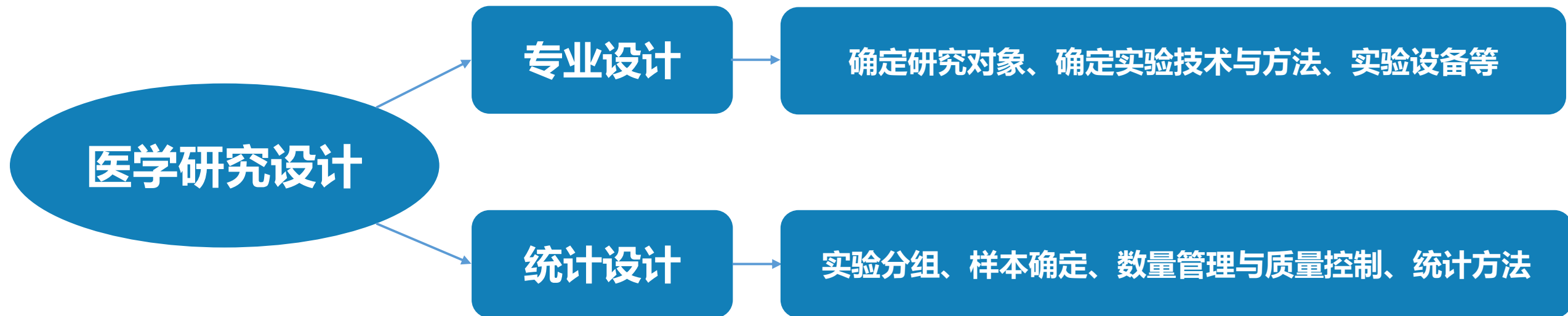
☑ 医学统计学基本内容

- 统计设计
- 数据整理与核查
- 统计描述
- 统计推断



☑ 医学统计学基本内容

- **统计设计** 是包含在医学研究设计中的一环，应该在研究开始时就参与。主要包括实验分组或抽样方法、样本含量估计、数据管理与质量控制、拟使用的统计分析方法等。**统计设计能够提高研究效率，并使结果更加准确和可靠。**



☑ 医学统计学基本内容

- **数据整理与核查** 主要是指对数据进行分组、对数据质量进行检查，考虑数据分布及变量转换，检查异常值及数据是否符合特定的统计分析方法要求等。

举例：

- ◆ 数据分组：如根据性别分组、年龄进行分组研究
- ◆ 数据核查：核查性别中是否包含第三类（错误），年龄小于0等情况
- ◆ 数据剔除需谨慎：异常值的剔除需仔细考虑

☑ 医学统计学基本内容

- **统计描述** 描述及总结一组数据的重要特征，目的是使实验或观察得到的数据表达清楚并便于分析。

统计描述结果的表达方式主要是：

- ◆ **统计指标**：大量数据→简单的数字、表达重要特征（均值、方差）
- ◆ **统计表**：代替冗长的文字描述，便于分析、对比
- ◆ **统计图**：生动、形象地表达结果

☑ 医学统计学基本内容

➤ **统计推断** 指由样本数据的特征推断总体特征的方法，包括参数估计和假设检验。

◆ **参数估计：** 点估计（通过样本均值估计总体的均值）、区间估计

◆ **假设检验：** 判断或比较不同样本是否来自同一个总体。

本课程讲解内容

▸ 统计描述

✓ 统计指标

✓ 统计表、统计图

第二章-第五章

▸ 统计推断

✓ 参数估计、假设检验

✓ t检验、方差分析、..线性回归与相关

第六章-第十一章

▸ 现代统计学方法

✓ 多重回归、logistic回归、生存分析

第十二章-第十四章

▸ 统计设计

第十五章





第三节

医学统计学中的基本概念



人民卫生出版社

PEOPLE'S MEDICAL PUBLISHING HOUSE

☑ 同质与变异

- **同质 (homogeneity)**：指**观察单位**或研究个体间具有**相同或相近**的性质，通常要求主要研究指标的影响因素相同或基本相同。
- **观察单位 (observed unit)**：研究对象的基本单元，可以是一个人、一个地点、一只动物、一份生物样品等。
- **变异 (variation)**：指同一种测量在总体中不同观察单位或个体之间的差别。

☑ 变量与数据类型

- **变量 (variable)**：对研究对象的某个或某些特征实施观测，这些特征称为变量，变量的观测值称为数据 (data)。
- 变量的类型有数值型变量、定性变量、有序变量。
- 对应的数据类型为定量数据、定性数据、有序数据。

☑ 数据类型

- 定量数据（quantitative data）：也称计量资料。变量的观测值是定量的，其特点是能够用**数值大小**衡量其水平的高低，一般有计量单位。根据变量的取值特征可分为：

◆ 连续型数据

可以取实数轴上任何数值，具有无限可能的值

◆ 离散型数据

通常只能取正整数值



- 定性数据（qualitative data）：也称计数资料。变量的观测值是定性的，表现为**互不相容的类别**或属性。例如，血型分为A、B、O、AB，性别分为男、女等。
- 有序数据（ordinal data）：也称半定量数据或**等级资料**。变量的观测值是定性的，但各**类别（属性）**之间有程度或顺序上的差别，如尿糖的化验结果分为 -、+、++、+++。

统计分析方法的选用与数据类型有密切的关系。根据分析的需要，不同类型的变量或数据之间可以进行转换。

原始的血红蛋白含量为定量数据；

如果分为正常和异常，转换为 **定性数据** ；

如果分为正常、轻度贫血、中度贫血、重度贫血四个等级，转换为 **有序数据** 。

☑ 总体和样本

- 总体（population）：指研究对象的全体，由所有的同质观察单位或个体组成。
描述总体特征的指标称为**参数**（parameter）。
- 样本（sample）：从研究总体中抽取部分有代表性的观察单位，对变量进行观测构成一个样本。描述样本特征的指标称为**统计量**（statistic）。

☑ 误差 (error)

是指观测值与真实值、样本统计量与总体参数之间的差别。

☑ 误差类型

- 系统误差 (systematic error)
- 随机误差 (random error)
 - ◆ 随机测量误差 (random measurement error)
 - ◆ 抽样误差 (sampling error)

系统误差（systematic error）

- **系统误差：**误差大小通常恒定或按照一定规律变化，具有明确的方向性。
- **产生原因：**由一些固定因素产生，如仪器未归零校正、研究对象不合适
- **消除与控制：**可通过周密的研究设计和测量过程标准化等措施加以消除



随机误差 (random error)

➤ 随机测量误差

➤ 抽样误差



随机测量误差（random measurement error）

- 随机测量误差：由各种偶然因素造成误差，误差没有固定的大小和方向，但具有一定的统计规律。
- 产生原因：各种不可预知因素
- 消除与控制：多次测量，取平均值

抽样误差（sampling error）

- 抽样误差（sampling error）：由于抽样而引起的样本统计量与总体参数间的差异。
- 产生原因：来源于个体的变异
- 消除与控制：不可避免，加大样本量

☑ 概率与概率分布

- **概率 (probability)**：描述随机事件发生可能性大小的定量度量。
- 事件A发生的概率可以写成 $P(A)$ ，其取值为 $0 \leq P(A) \leq 1$ ，
- $P(A) = 0$ 表示该事件不可能发生， $P(A) = 1$ 表示该事件必然发生。

➤ 古典概率定义:

$$P(A) = m/n$$

其中n表示可能出现基本结果的总数目, m表示事件A包含的基本结果数, 条件是每个基本结果出现的可能性相同。(频率)

频率是有限次数的试验所得的结果,
概率是频数无限大时对应的频率.

假设事件A的概率是0.3, 在100次中发生28次, 那么它的频率是
 $28/100=0.28$

➤ 统计概率, 医学研究中经常使用的是统计概率, 定义为:

$$P(A) = \lim_{N \rightarrow \infty} \frac{n_A}{N}$$

其中N 和 n_A 分别表示观察的总例数和发生事件A的例数。

频率是在一定数量的某件事情上面, 发生的数与总数的比值.

概率是一个稳定的数值, 也就是某件事发生的概率是多少。



- **概率分布 (probability distribution)**：随机变量**所有可能的取值**与各取值下所发生的**概率**之间的对应关系，用以全面地表述随机变量取值的概率。

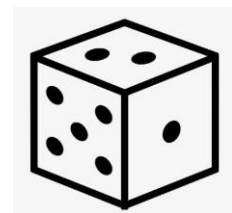
随机变量 (Random Variable)：在随机试验中，确定一个对应关系，使得每一个**试验结果**都用一个确定的**数字**表示。

随机变量举例：

抛硬币的结果是一个随机变量，可用数字1，表示正面，数字0，表示背面



丢色子的结果是一个随机变量，可用数字1，表示丢的结果为1，...数字6，表示丢的结果为6



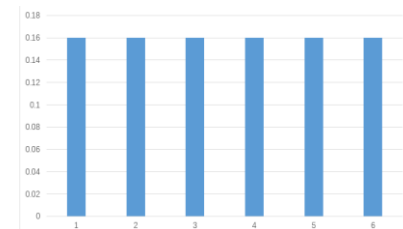
血型的结果是一个随机变量，可用数字1，表示A型，数字2，表示B型，数字3，表示O型，数字4，表示AB型

概率分布举例：

- 抛硬币的结果是一个随机变量，可用数字1，表示正面，发生概率为1/2，数字0，表示背面，发生的概率为1/2

X	0	1
P	1/2	1/2

- 丢色子的结果是一个随机变量，可用数字1，表示丢的结果为1，发生的概率是1/6,...数字6，表示丢的结果为6,发生概率是1/6



- 血型的结果是一个随机变量，可用数字1，表示A型，数字2，表示B型，数字3，表示O型，数字4，表示AB型, O型:34.11%， B型:28.98%， A型:28.29%， AB型:8.69%

- 在统计学上，统计结论都是基于一定概率得出的，习惯上将 $P \leq 0.05$ 的事件称为小概率事件，表示在一次试验中发生的可能性很小。如果小概率事件在现实中出现，就要追究其原因。
- 简单例子，例如我们身高取值集中在1.55-1.75，假如出现身高1.4，考虑是有原因的（疾病因素）



第四节

概率的基本运算法则



人民卫生出版社
PEOPLE'S MEDICAL PUBLISHING HOUSE

1. 概率的乘法法则

- 如果有两个事件A和B相互独立，即事件A的出现并不影响事件B出现的机会，则两者同时出现的概率等于各自出现的概率乘积：

$$P(A B) = P(A)P(B)$$

- 例如，某地40岁以上人群中 2 型糖尿病患病（A）的概率为22.5%，患有甲状腺结节（B）的概率为20.2%，则患有 2 型糖尿病同时患有甲状腺结节疾病的概率为 $0.225 \times 0.202 = 0.045$ ，即约等于4.5%。

2. 概率的加法法则

- 如果有事件A和B相互独立而且互不相容，则出现任一事件的概率

$$P(A \cup B) = P(A) + P(B)$$

- 如果事件的出现相互独立但可能互相包容，则上述公式需要改为

$$P(A \cup B) = P(A) + P(B) - P(AB)$$

- 其中 $P(AB)$ 表示A和B两事件同时出现的概率。
- 例如，某地40岁以上人群中2型糖尿病患病（A）的概率为22.5%，患有甲状腺结节（B）的概率为20.2%，则出现任一种疾病的概率为

$$P(A \cup B) = 0.225 + 0.202 - 0.225 \times 0.202 = 0.38155$$

即约等于38.2%。

3. 条件概率

- 如果事件A和B之间不独立，即事件B与事件A存在某种关联，则在A发生的条件下B的概率称为条件概率，记为 $P(B|A)$ ；反之，在B发生的条件下A发生的概率记为 $P(A|B)$

$$P(A|B) = \frac{P(AB)}{P(B)}$$

两者的联合概率： $P(AB) = P(A)P(B|A) = P(B)P(A|B)$

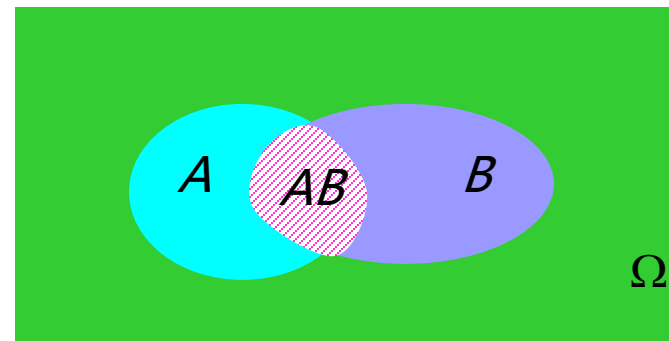
$P(AB)$ 与 $P(A|B)$ 的区别：

$$P(AB) = \frac{AB \text{ 包含的样本点数}}{\text{样本空间} \Omega \text{ 包含的样本点数}}$$

$P(AB)$ ：以 Ω 为样本空间；

$$P(A|B) = \frac{AB \text{ 包含的样本点数}}{B \text{ 包含的样本点数}}$$

$P(A|B)$ ：以 $\Omega_B = B$ 为样本空间；



- 例如，某地40岁以上人群中2型糖尿病（A）患病的概率为12.5%，患有高血压（B）的概率为20.2%，患有糖尿病的患者中有65.2%的患者合并患有高血压，则同时患有糖尿病和高血压的概率为

$$P(AB) = P(A) P(B | A) = 0.125 \times 0.652 = 0.0815$$

同时可以得出在高血压的患者中患有糖尿病的概率

$$P(A|B) = \frac{P(AB)}{P(B)} = \frac{0.0815}{0.202} = 0.4035$$

3. 全概率公式

定义 设 Ω 为实验 E 的样本空间, $A_1, A_2, \dots, A_n$ 为 E 的一组事件, 若

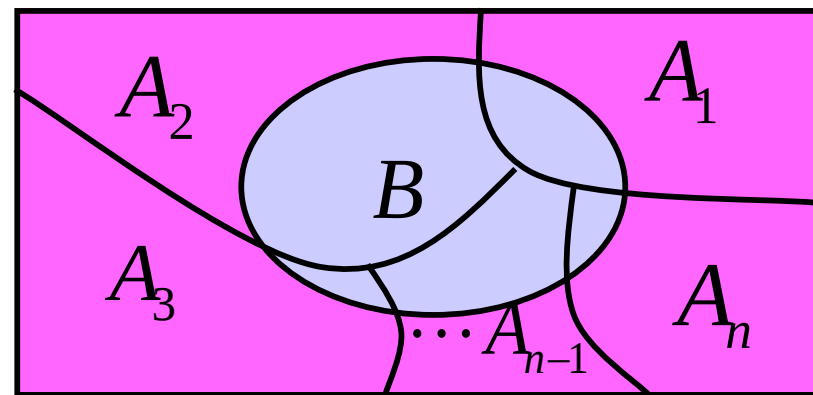
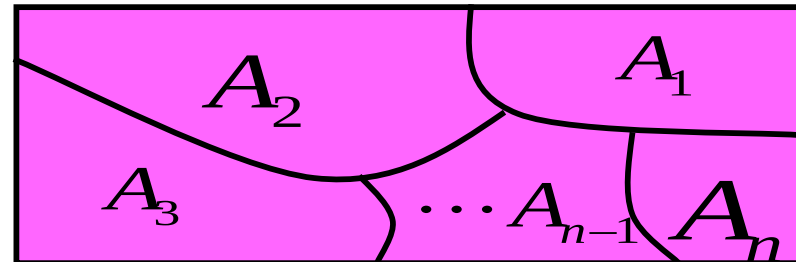
$$(1) A_i A_j = \Phi, \quad i \neq j, \quad i, j = 1, 2, \dots, n;$$

$$(2) A_1 \cup A_2 \cup \dots \cup A_n = \Omega.$$

则称 $A_1, A_2, \dots, A_n$ 为样本空间 Ω 的一个划分.

B 为 E 的事件, $A_1, A_2, \dots, A_n$ 为 Ω 的一个划分, 且 $P(A_i) > 0 (i = 1, 2, \dots, n)$, 则有:

$$P(B) = P(B|A_1)P(A_1) + P(B|A_2)P(A_2) + \dots + P(B|A_n)P(A_n) = \sum_{i=1}^n P(A_i)P(B|A_i)$$



☑ 全概率公式的证明:

$$B = B \cap \Omega = B \cap (A_1 \cup A_2 \cup \cdots A_n)$$

$$= BA_1 \cup BA_2 \cup \cdots \cup BA_n.$$

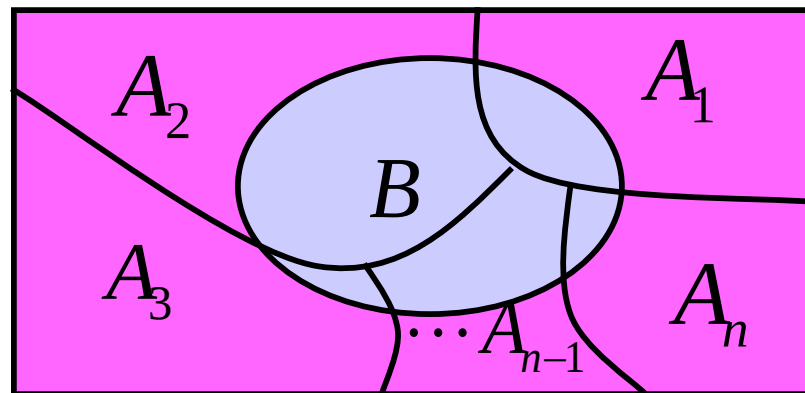
$$\text{由 } A_i A_j = \emptyset \Rightarrow (BA_i)(BA_j) = \emptyset$$

$$\Rightarrow P(B) = P(BA_1) + P(BA_2) + \cdots + P(BA_n)$$

$$\Rightarrow P(B) = P(A_1)P(B|A_1) + P(A_2)P(B|A_2) + \cdots + P(A_n)P(B|A_n) = \sum_{i=1}^n P(A_i)P(B|A_i)$$

某事件 B 的发生由各种可能的“原因” $A_i (i=1,2,\cdots,n)$ 引起, 而 A_i 与 $A_j (i \neq j)$ 互斥, 则 B 发生的概率 $P(A_i|B) (i=1,2,\cdots,n)$ 有关, 且等于它们的总和。

全概率公式的主要用处在于: 它可以将一个**复杂事件**的概率计算问题, **分解**为若干个**简单事件**的概率计算问题, 最后应用概率的可加性求出最终结果。



化整为零
各个击破



5. Bayes公式

设 Ω 为试验 E 的样本空间, B 为 E 的事件, $A_1, A_2, \dots, A_n$ 为 Ω 的一个划分,且 $P(B) > 0, P(A_i) > 0 (i = 1, 2, \dots, n)$, 则

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{\sum_{j=1}^n P(B|A_j)P(A_j)}, \quad i = 1, 2, \dots, n.$$

证明:

$$P(A_i|B) = \frac{P(A_i B)}{P(B)} = \frac{P(A_i)P(B|A_i)}{P(B)} = \frac{P(A_i)P(B|A_i)}{\sum_{i=1}^n P(A_i)P(B|A_i)}.$$



- 例如，根据以往的数据，某地肝癌发病率为0.4%，已知肝癌病人用维生素K缺乏诱导蛋白（PIVKA-II）和甲胎蛋白（AFP）联合检验呈阳性为85%，健康人用该法检验呈阴性为90%。若以A表示患有肝癌，B表示检验呈阳性，则在检测结果呈阳性的人中，真患有肝癌病的可能性为

$$P(A | B) = \frac{0.85 \times 0.004}{0.85 \times 0.004 + 0.10 \times 0.996} = 0.033$$

从这一计算结果可以看出什么？

6.先验概率与后验概率

上题中概率 0.004 是由以往的数据分析得到的, 叫做先验概率。

而在得到信息之后再重新加以修正的概率 0.033, 叫做后验概率。

☑ 注意

- 统计推断结论都是基于一定概率得出的，概率的计算则是依据样本统计量的概率分布得到相应的概率值 P 。习惯上将 $P \leq 0.05$ 的事件称为小概率事件，表示在一次试验中发生的可能性很小。如果小概率事件在现实中出现，就要追究其原因。

学习医学统计学目的

- **统计思维**

由样本推断总体，结论具有不确定性—有可能犯错误

- **统计学任务**

限定犯错误概率的大小，在此前提下，作决策。

本章小结

1. 医学统计学是关于收集数据、分析数据和由数据得出结论的一组概念、原则和方法。其重要作用在于能够透过偶然现象来探测其规律性，使研究结论具有科学性。
2. 医学统计学的基本内容包括统计设计、数据整理与核查、统计描述和统计推断。这四项基本内容相互联系。
3. 由实验或观察得到的数据可分为定量数据、定性数据和有序数据。统计分析方法的选用与数据类型有密切的关系。根据分析的需要，不同类型的变量或数据之间可以进行转换。



本章小结

4. 误差是指观测值与真实值、样本统计量与总体参数之间的差别。根据误差的性质和来源主要可以分为系统误差、随机测量误差和抽样误差几种类型。
5. 变异和概率是统计学中两个最基本的概念。变异可以来自生物个体、测量等多个方面，由于有变异，则结果不定，对此可以使用概率进行描述。



课后习题一

医学统计学的主要作用是什么？

其主要作用是通过**数据的偶然性**揭示医学研究**内在的规律性**，**正确地选用统计学分析方法**，并结合医学专业知识和研究目的**做出科学的结论**。

课后习题二

医学统计学的基本内容包括：

统计设计 数据整理与核查 统计描述 统计推断

课后习题三

统计描述与统计推断的主要特点是什么：

统计描述指统计指标、统计图、统计表

统计推断指由样本数据的特征推断总体特征的统计学方法

课后习题四

统计量与参数的差别是什么？

描述样本特征的指标称为统计量，由样本数据计算得到

描述总体特征的指标称为参数，由“全体”数据算出

课后习题五

常见的三类误差？ 如何控制？

系统误差

通过周密的研究设计、测量过程标准化等措施进行控制

随机测量误差

通过多次测量取平均值进行控制

抽样误差

通过加大样本量进行控制

课后习题六

概率分布的涵义和意义？

随机变量所有可能的**取值**与各取值下所发生的**概率**之间的对应关系，用以全面地表述随机变量取值的概率。



人民卫生出版社

PEOPLE'S MEDICAL PUBLISHING HOUSE

谢谢观看