偏倚和混杂

研究的真实性

- 研究的真实性(accuracy)可通过衡量研究中是否存在误差及误差的影响程度来反映，理论上要求在有限的资源条件下达到最小误差

误差

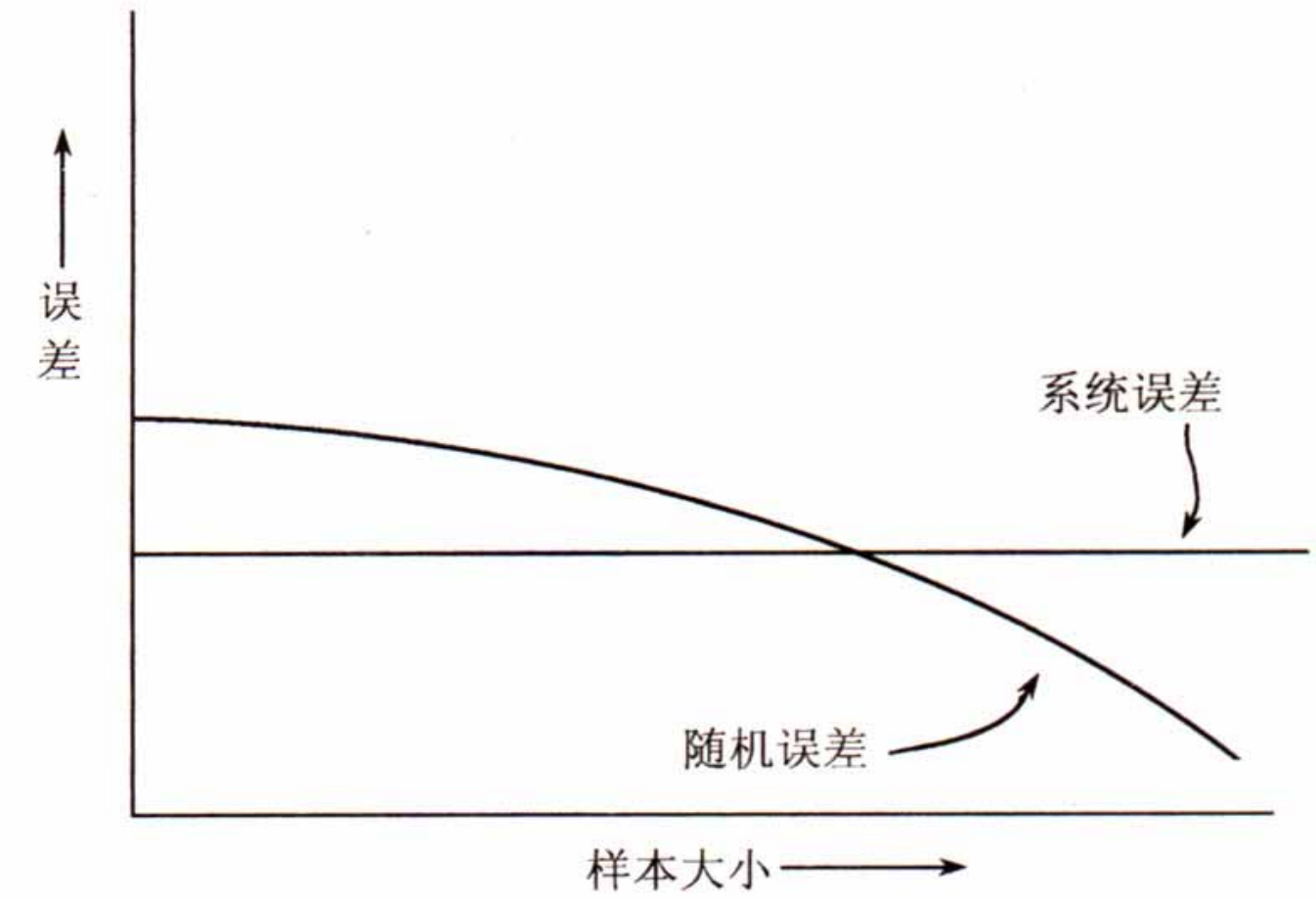
- 误差(error)是指研究的测得值和真实值之间的偏离，包括随机误差和系统误差两类。
- 随机误差（random error）是由抽样而产生的误差，可影响研究的精确性，一般可通过统计学方法予以估计或评价。
- 系统误差（systematic error），又称偏倚（bias），发生在研究的设计、实施、分析、推断等各阶段，可影响研究的有效性。

偏倚的定义

- 偏倚指的是研究设计、实施、分析和推断过程中存在的各种对暴露因素与疾病关系的错误估计，它系统地歪曲了暴露因素与疾病间的真实联系。

- 偏倚是一种系统误差，它与随机误差不同，即使样本增加至无穷大，系统误差仍维持原样（图10-1）。

图10－1



偏倚的方向

- 偏倚是有方向的。
- 当研究结果因偏倚而被夸大时，称为正偏倚；
- 而当研究结果因偏倚而被缩小时，称为负偏倚。

- 相对于危险因素，正偏倚时，$RR_{偏}>RR_{真}$；负偏倚时，$RR_{偏}<RR_{真}$。
- 相对于保护因素，正偏倚时，$RR_{偏}<RR_{真}$，负偏倚时，$RR_{偏}>RR_{真}$。

偏倚的种类

- 选择偏倚
- 信息偏倚
- 混杂偏倚

选择偏倚的定义

- 选择偏倚是由被选入到研究中的研究对象与没有被选入者在暴露或疾病有关的特征上的差异所造成的系统误差。

- 流行病学研究中，当按一定的条件识别研究对象时，从所纳入的研究对象中获得的有关因素与疾病的联系系统地偏离了源人群中该因素与疾病之间的真实联系，即认为有选择偏倚(**selection bias**)存在。

选择偏倚的种类

- 检出偏倚或检出症候偏倚
- 诊断偏倚
- 入院率偏倚
- 纳入/排除偏倚
- 奈曼偏倚
- 志愿者偏倚
- 健康工人效应
- 失访偏倚
- 无应答偏倚

检出偏倚或检出症候偏倚

(detection bias/detection signal bias)

- 有检出症候者：若暴露于所研究因素可以产生某种临床症候，则具有该因素暴露史的病例相对于无暴露史的病例可能更早就诊，有更大的可能被选择性地纳入样本。
- 例：雌激素与子宫内膜癌

  雌激素 → 子宫内膜生长 → 出血 → 就诊 → 检出。

- 当病例对照研究中的病例主要为有检出症候者，而对照来自产生所有病例（有检出症候和无检出症候病例）的人群时，则通常可以夸大暴露的危险性，产生偏倚，这种偏倚被称为检出偏倚或检出症候偏倚。

诊断偏倚(diagnostic bias)

- 当临床医生或疾病报告系统对暴露者和非暴露者采用不同的诊断标准时，由此造成的诊断上的偏倚称为诊断偏倚。

入院率偏倚

- 入院率偏倚又称伯克森偏倚(Berkson’s bias)，是指利用医院就诊或住院病人作为病例对照研究对象时，由于病例和对照入院率不同而导致的偏倚。

入院偏倚

	有因素(X)	无因素	总人数	%
病例组(A)	200	800	1,000	20
对照组(B)	200	800	1,000	20

假定：因 X，A，B的入院率分别为0.4，0.5及0.2

则：1 具有X及A的人在人群中有200，医院中则有
200*0.4=80，(200-80)*0.5=60，共140人

2 有A但无X的在人群中有800，医院中有
800*0.5=400

3 有B及X的人
200*0.4=80，(200-80)*0.2=24，共104人

4 有B但无X
800*0.2=160

故医院中的样本为：

	X	$\bar{X}$
A	140	400
B	104	160

$$OR=\frac{140*160}{104*400}=0.54$$

人群中　$OR=AD/BC$

医院中　$\hat{OR}=b*OR$

$b=(S_1*S_4)/(S_2*S_3)$

$\hat{OR}=[(S_1*S_4)/(S_2*S_3)]*OR$

暴露状况	疾病状况	人群中频数	入院率	医院中频数
有	病例	A	S_1	S_1A
	对照	B	S_2	S_2B
无	病例	C	S_3	S_3C
	对照	D	S_4	S_4D

$$\hat{OR}=\frac{a*d}{b*c}=\frac{S_1A*S_4D}{S_2B*S_3C}=[(S_1*S_4)/(S_2*S_3)]*OR$$

纳入/排除偏倚(inclusion/exclusion bias)

- 病例对照研究中由于系统性地纳入或排除患有已知与暴露有关疾病的对象所致的偏倚称为纳入或排除偏倚。

奈曼偏倚(Neyman bias)

- 又称现患- 新发病例偏倚(**prevalence-incidence bias**)。病例对照研究往往纳入现患病例或存活病例，即同时纳入新、旧病例而不包括死亡病例和那些病程短的病例。由此而产生的偏倚称为奈曼偏倚。因为：
- ⑴现患病例与新病例的暴露状况、病情、病型、病程和预后等都不尽相同。
- ⑵现患病例可能是“生物学上的强者”。
- ⑶现患病例往往对自身所患疾病有所了解，有时会主动更改其对危险因素的暴露，导致了对危险因素与疾病关系的低估。

美国弗明汉地区男性居民血胆固醇水平与冠心病关系研究

胆固醇	队列研究(第 6 次检查)			病例对照研究(第 6 次检查)		
百分位	冠心病	非冠心病	合计	冠心病	非冠心病	合计
>75	85	462	547	38	34	72
<=75	116	1511	1627	113	117	230
合计	201	1973	2174	151	151	302
	OR=2.40			OR=1.16		

志愿者偏倚(volunteer bias)

- 当研究的暴露组或治疗组对象为志愿者时，在暴露的志愿者和非暴露的对照（主要为非志愿者）间的比较可能受到志愿者偏倚的影响。因为：
- ⑴除了暴露状态不同外，在与疾病发生相关的其他很多方面也可能不同，如志愿者具有更强的自我保健意识等。
- ⑵志愿者由于对疾病及其危险因素的了解较多，在回忆暴露情况时可能会过分强调其暴露程度；或因未患所研究疾病而对回忆暴露史不感兴趣。

健康工人效应 (health worker effect)

- 在职业流行病学研究中，常常碰到的一个问题是健康工人效应。

- 通常，受雇佣的工人比失业者健康；有些行业还对雇员的健康有专门的要求。因此，由于健康工人效应，可能反而会得出暴露组疾病危险性低于非暴露组的结果，而其实是因为这些健康工人比一般工人或失业者健康，且对暴露因素的易感性可能低于一般工人或失业者。

美国某交通企业男性五人与全美男五人
死亡率比较

年龄组	59-63 交通企业	59-61 全美男性五人	死亡率比较 (校正系数)
40-	2.0-3.0	3.75-5.51	0.536
45-	3.4-5.4	6.05-9.11	0.574
50-	6.2-10.1	10.14-14.40	0.658
55-	11.2-15.8	15.49-21.54	0.735
60-	17.0-21.8	23.50-32.26	0.700

美国橡胶五人SMR“健康五人效应”校正

年龄	64-72 观察死亡数	未校正 期望数	SMR	校正数	SMR
40-	17	12.9	1.32	6.9	2.46
45-	42	49.0	0.86	28.1	1.49
50-	73	87.1	0.84	57.3	1.27
55-	144	141.6	1.02	104.1	1.38
60-	213	234.3	0.91	164.0	1.30
合计	489	524.9	0.93	336.5	1.45

失访偏倚(loss-to follow-up bias)

- 研究对象在随访过程中发生影响疾病危险性评价的失访时，如因健康原因、死亡、不合作、迁出等失访，则可发生失访偏倚。

- 失访偏倚对研究结果的影响取决于失访的程度、失访者在所比较组的分布和失访原因与所研究结果的关联程度等。

无应答偏倚(non-response bias)

- 无应答偏倚主要发生于现况调查，表现为调查对象不合作或不参与。

- 这些无应答对象通常不能代表所研究人群，且无法判断其暴露或疾病状况，因此当无应答率较高时，如大于15%，由于选择偏倚的存在，从应答人群中得出的有关研究因素与疾病的联系不能反映两者间的真实联系。

选择偏倚的控制

1. 研究设计阶段
- 建立和利用健康监测系统信息，尽可能使用发病率资料。
- 采用严格科学的研究设计。
- 明确对象纳入标准、统一疾病诊断和监测程序。

2. 资料收集阶段
- 加强随访、提高应答率。
- 在资料收集阶段尽可能多地收集有关暴露史的各种信息
- 确保疾病的诊断不是依据暴露史而得出。

3. 数据分析阶段

信息偏倚的定义

- 信息偏倚(information bias)又称观察偏倚(observational bias)，指在研究的实施阶段从研究对象获取研究所需信息时所产生的系统误差。

- 信息偏倚可发生于各种类型的流行病学研究，可来自研究对象，也可来自研究者本身，或来自用于测量的仪器、设备和方法。

信息偏倚的种类

- 错分偏倚
- 均数回归趋势
- 生态学偏倚等

错分偏倚(misclassification bias)

- 由于研究中的测量误差如资料收集不准确或不完整等造成对研究对象的暴露程度或疾病结果的错误归类，影响了结果估计的有效性，此类偏倚统称为错误分类偏倚或错分偏倚。

- 错分偏倚由发生在不同类型研究中的系统误差所致，包括回忆偏倚、报告偏倚、诊断怀疑偏倚、暴露怀疑偏倚和测量偏倚等。

回忆偏倚(recall bias)

- 回忆偏倚多见于病例对照研究和回顾性队列研究。
- 由于所调查的因素发生于过去，回忆的准确性和完整性受回忆间期长短、所回忆因素对研究对象的意义和该因素的发生频率的影响，造成对研究结果的有偏估计。护士乳腺癌和围产期特征关系的病例对照研究
- 而且既往经历对病例和非病例的意义往往迥然不同，病例组对既往暴露情况的记忆深度和详细程度通常较对照组为甚，由此造成了回忆偏倚在各比较组中分布不同。
- 代理者的记忆和对对象的了解程度（代理者偏倚）

风湿性关节炎家族史

	有风湿史 (%)	无风湿史 (%)
双亲都无	16	55
一位有	53	37
双亲都有	31	8

患者兄弟（姐妹）提供家族史

	提供者有病史	提供者无病史
双亲都无	27	50
一位有	58	42
双亲都有	15	8

Schull and Cobb

报告偏倚(reporting bias)

- 与回忆偏倚不同，报告偏倚是因为对象有意夸大或隐瞒某些信息导致了对疾病或暴露程度的错误分类。

诊断怀疑偏倚和暴露怀疑偏倚

(diagnostic suspicion bias and exposure suspicion bias)

- 由于研究者或被研究者的主观倾向、愿望或偏见所导致的对暴露因素和/或疾病结果的错误判断，从而歪曲了暴露同疾病间的真实联系，分别称为诊断怀疑偏倚或暴露怀疑偏倚。

- 在队列研究或实验中，如果研究者事先已认为暴露于研究因素可能与疾病的发生有关，则可能对暴露或干预组进行非常严格细致的检查，而对非暴露组则不然，造成对研究结果判断的偏倚，此类偏倚称为诊断怀疑偏倚。

调查者偏倚(interviewer bias)

- 调查者在收集、记录和解释来自研究对象的信息时发生的偏倚称为调查者偏倚。

测量偏倚(measuring bias)

- 由于研究中所使用的仪器、设备、试剂、方法和条件的不精良、不标准、不统一或研究指标设定不合理、数据记录不完整造成的研究结果系统地偏离其真值的现象称为测量偏倚。

- 测量偏倚可发生在各种流行病学研究的设计、实施和资料处理过程中。

双侧颈动脉狭窄外科与内科疗效比较

	复发一时性缺血、卒中、死亡		总数
	有	无	
外科	43	36	79
内科	53	19	72
			151

外科手术减少的危险度:
〔(53/72) - (43/79)〕/ (53/72) =27%, X^2=5.98

统计所有167个病人

外科	58	36	94
内科	54	19	73
			167

外科手术减少的危险度:
〔(54/73) - (58/94)〕/ (54/73) =16%, X^2=2.80

1970年 Fields

均数回归趋势(regression to the mean)

- 以连续变量表示的某些测量值，由于随机误差的存在，在初次测量时可能表现为极端值，即远远地高于或低于人群中的其他对象的相应值，但在以后的多次重复测量中，该对象的上述测量值会出现向这一变量的人群均数靠拢的倾向，称为均数回归趋势。
- 均数回归趋势所呈现的变化可能会被当作真实的变化而错误地归因于某种干预措施的效果。

生态学偏倚(ecologic bias)

- 生态学研究中个体水平的生物学信息由于被结合于群体（组群）水平的暴露与疾病结果的推断中而丧失。由于每个组群内部的暴露状态并不一致，因此，由组群间暴露水平与疾病发生的差异得出的生态学联系可能与相应的个体暴露水平与疾病发生的关系迥然不同，从而导致生态学谬误。

信息偏倚的控制

- 错分偏倚的控制
- 均数回归趋向的控制
- 生态学偏倚的控制

错分偏倚的控制措施

- 首先，在研究设计中对暴露因素必须有严格、客观的定义，并力求指标定量化。

- 其次，在资料收集阶段，应尽量选用客观定量指标，可选用回忆指征帮助对象回忆，也可利用实物或照片来准确获取信息。

（2）错分偏倚的校正：

- 错分偏倚在所比较组内的分布可以相同，也可以不同，可用错分的灵敏度和特异度来表示。

- 在病例对照研究中，错分的灵敏度指正确查出有暴露史者占实际有暴露史人数的比例；特异度指正确查出无暴露史者占实际无暴露史人数的比例。

表 1．真实的暴露状况

	病例	对照	合计
暴露	60	30	90
非暴露	40	70	110
合计	100	100	200

OR=60*70/（40*30）=3.5

表 2．错分的暴露状况

错分后暴露状况	真实暴露状况					
	病例			对照		
	暴露	非暴露	合计	暴露	非暴露	合计
暴露	54	12	66	18	7	25
非暴露	6	28	34	12	63	75
合计	60	40	100	30	70	100

错分后的 OR=66*75/（25*34）=5.8 病例组：Se=54/60=0.9，Sp=28/40=0.7

对照组：Se=18/30=0.6，Sp=63/70=0.9

- 当各比较组发生错分的灵敏度和特异度分别相同时，产生的错分偏倚称为均衡性错分(non-differential misclassification)，又称无差异错分或非特异性错分。

- 当各比较组发生错分的灵敏度和特异度各不相同时，称为非均衡性错分(differential misclassification)，又称差异错分或特异性错分。

均数回归趋向的控制

- 在实验研究中，可以通过设立对照组、尤其是随机化分组的对照组来控制均数回归趋向的影响。
- 另外一个有效的方法是不论在基线时还是随访过程中，采用一组重复测量值的均数来代替对象的相应指标测量值。重复测量的次数越多，所获值越稳定，受均数回归趋向的影响越小，当然也需考虑测量的成本效益。
- 在分析过程中也可通过各种统计分析方法来估计均数回归趋向的程度。

生态学偏倚

- 很难避免出现生态学偏倚。但其意义在于为进一步的分析性流行病学研究提供线索，因此，只要充分注意到生态学研究的局限性，并运用适当的统计学方法来估计生态学偏倚的影响程度，必要时开展纵向的生态学趋势研究，生态学研究结果还是可以获得合理的应用。

混杂偏倚 (confounding)

流行病学研究中，由于一个或多个外来因素（又称第三因子）的存在，掩盖或夸大了研究因素与疾病（或事件）的联系，从而部分或全部地歪曲了两者之间的真实联系，称为混杂偏倚或混杂,引起混杂偏倚的外来因素称为混杂因素(**confounder**)。

例：吸烟、肺癌、年龄

混杂因素的特点：

（1）混杂因素必须与所研究疾病的发生有关，是该疾病的危险因素之一。

（2）混杂因素必须与所研究因素有关。

（3）混杂因素必须不是研究因素与疾病病因链上的中间环节或中间步骤。

中年人少量饮酒与心肌梗塞危险性 – 混杂因素示例

	少量饮酒	不饮酒
心肌梗塞发病（例）	140	100
随访人年（人年）	30,000	30,000
发病率（1/千）	4.67	3.33
	RR=1.40	

不同性别中年人少量饮酒与心肌梗塞危险性 – 混杂因素示例

	男性			女性	
	少量饮酒	不饮酒		少量饮酒	不饮酒
心肌梗塞发病（例）	120	60		20	40
随访人年（人年）	20,000	10,000		10,000	20,000
发病率（1/千）	6.00	6.00		2.00	2.00
	RR=1.0			RR=1.0	

图10－2

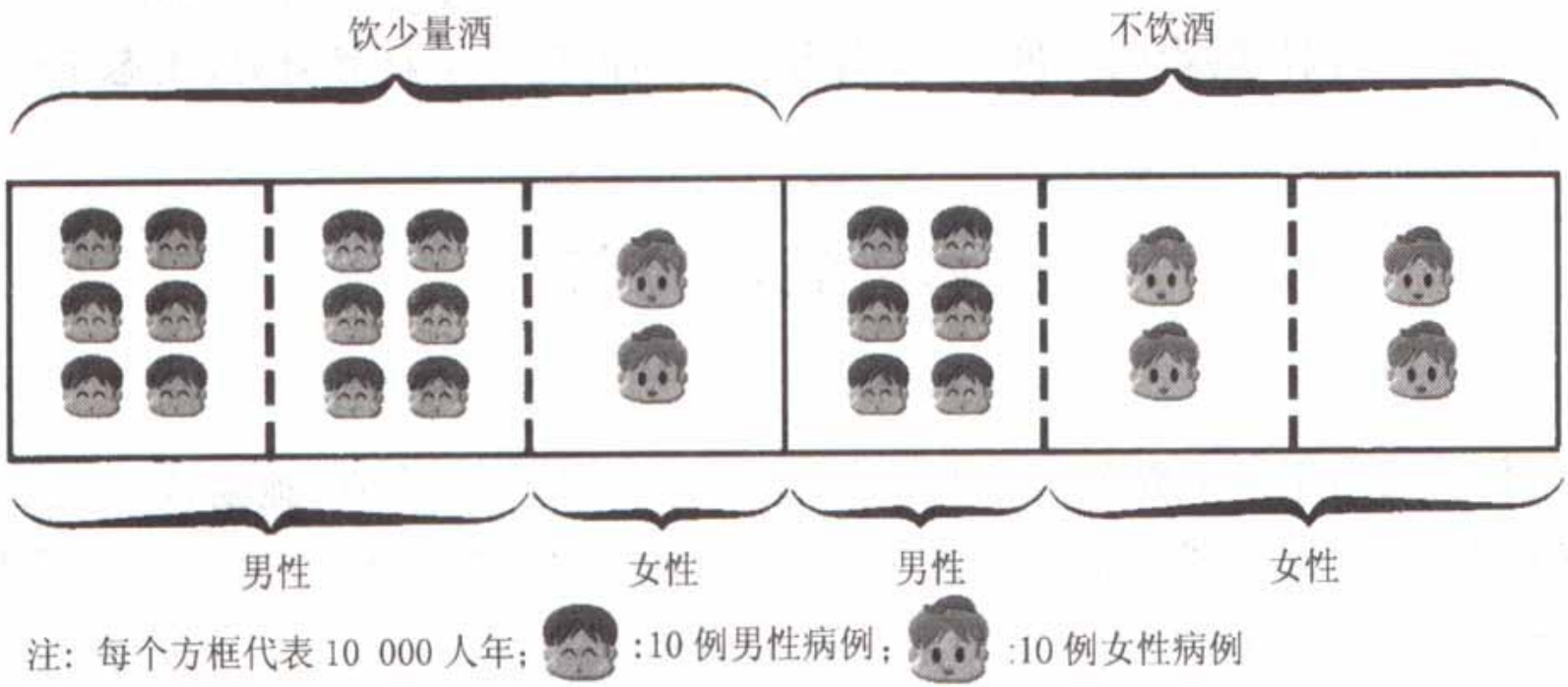


图 10-2　不同性别对象饮酒与心肌梗死发病危险性－混杂因素示例图

混杂偏倚的判断和测量

- 判断和测量某一可疑混杂因素的混杂作用，可以通过比较含有该因素时研究因素与疾病的效应估计值，如RR或OR，与排除该因素后的效应估计值来实现。

假设：在一次非配对的病例对照研究中，暴露因素为X，疾病为D，潜在混杂因素为年龄。

表 1　因素 X 在各比较组的分布

因素 X	病例组	对照组
有	30	18
无	70	82
合计	100	100

cOR(crude odds ratio)=1.9

表 2　因素 X 与疾病 D 按年龄的分层分析

因素 X	<40 岁		>=40 岁	
	病例组	对照组	病例组	对照
有	5	8	25	10
无	45	72	25	10
合计	50	80	50	20

$OR_{<40}=1.0$　　$OR_{>=40}=1.0$

$aOR_{M\text{-}H}=1$

表 3 无因素 X 者中年龄与疾病的关系

年龄(岁)	病例组	对照组
>=40	25	10
<40	45	72
合计	70	82

OR_{DFX}= 25*72/（10*45）=4.0

表 4　各比较组的年龄分布

年龄(岁)	病例组	对照组
>=40	50	20
<40	50	80
合计	100	100

病例组>=40 为 50%，对照组为 20%。

表 5　对照组中因素 X 与年龄的关系

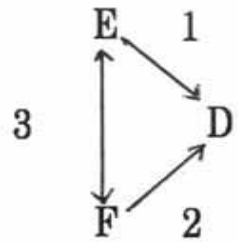
年龄(岁)	有 X	无 X
>=40	10	10
<40	8	72
合计	18	82

OR_{XFD} =10*72/（10*8）=9

混杂偏倚的判断和测量

- 当cRR=*aRR*(*f*)或cOR=*aOR*(*f*)时，则*f*无混杂作用。
- 当cRR≠*aRR*(*f*)或cOR≠*aOR*(*f*)时，且分层后有*f*层和无*f*层的分层*RRi*或*ORi*相同，则*f*有混杂作用。

产生混杂的条件:



1 CRR≠ARR　　定群研究
COR≠AOR　　病例对照研究

粗相对危险度不等于调整后的相对危险度。

2 RR_{df} | $\bar{E}\neq 1$　在定群研究的非暴露组中
OR_{df} | $\bar{E}\neq 1$　在病例对照研究中无暴露情况下

3 $RR_{ef}\neq 1$

暴露与非暴露组中，可疑的混杂因素分布不均匀，即混杂因素与研究因素之间有联系。

OR_{ef} | $\bar{D}\neq 1$　在对照组中

饮酒与心肌梗塞的关系

	病例	对照
饮　酒	71	52
不饮酒	29	48

$COR=2.26$，　$X^2=7.62$　　$P=0.006$

	吸烟 F_+		不吸烟 F_-	
	饮酒	饮酒	饮酒	饮酒
病例	63	7	8	22
对照	36	4	16	44

$OR_{F+}=1.0$　　$OR_{F-}=1.0$

$$AOR_{(F)}=\frac{\sum(ai*bi/Ni)}{\sum(bi*ci/Ni)}=1.0$$

$$OR_{DF}\mid\bar{E}=\frac{7\times44}{4\times22}=3.5\qquad X^2=4.013$$

$$OR_{EF}\mid\bar{D}=\frac{36\times44}{4\times16}=24.75\qquad X^2=37.184$$

混杂偏倚的控制

- 限制(restriction)
- 匹配(matching)
- 随机化(randomization)
- 统计处理
- 灵敏度分析

限制(restriction)

- 一个提高可比性的方法是在选择研究对象时，限制在具有一定特征的对象中进行观察，以排除其他因素的干扰。

- 但用这种方法来控制偏倚所获得的结论常有很大局限性，影响研究对象的代表性，使研究结果外推至一般人群时受限。

匹配(matching)

- 匹配是指在为研究对象设立对照时，使病例和其对照在一个或多个潜在混杂因素上相同或相近，从而消除混杂因素对研究结果的影响。

匹配(matching)

- 匹配的目的是为了控制混杂、提高研究的统计学效率。
- 病例对照研究、队列研究和实验研究均可采用匹配。尤其是队列研究，使用匹配可达到直接控制混杂的效果，但病例对照研究中仍需进行分层分析来较好地控制混杂。

随机化(randomization)

- 随机化是指以随机化原则将研究对象以同等的机率被分配在各处理组中，从而使潜在的混杂因素在各组间分布均衡。
- 随机化多用于实验研究，尤其是临床试验。

统计处理：

常用的估计和控制混杂偏倚的统计处理方法有：

- 分层分析
- 标准化
- 多因素分析

分层（stratification）

- 分层是指将研究所获资料按混杂因素分成数层（亚组）进行分析。
- 分层是最常用的检出和控制偏倚的方法之一。
- 可以通过Mantel-Haenszel分层分析法进行分析。但如果欲控制的混杂因素较多，则分层分析对样本量的要求较大，此时，可以应用多因素分析方法如Logistic回归分析等来估计和控制混杂。

标准化（standardization）

- 当比较两个率时，如果两组对象内部构成存在的差别足以影响结论，可用率的标准化加以校正，亦即使可能影响结果的因素受到同等的加权，使这两个率可比、无偏倚，这种方法称为标准化。