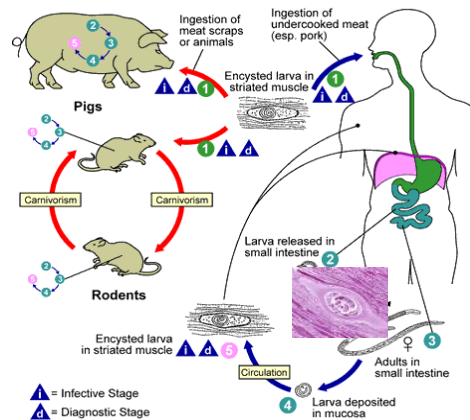




Design of studies

María J. Vilar

Dep. of Food Hygiene and Environmental Health



28-31. 05. 2013

At the end of the course, the student should be able to:

- Describe the epidemiologic concepts
- Describe patterns of occurrence of foodborne pathogens
- Differentiate between proportion, ratio, and rate
- Calculate and interpret measures of association and frequency

Questionnaires and Sampling

Sample size and type of sampling

Types of studies

Experimental

Observational

Population and samples

Questionnaires and Sampling

All individuals -> Census

Representative -> Sample -> Sampling : which and how many individuals?

Sampling

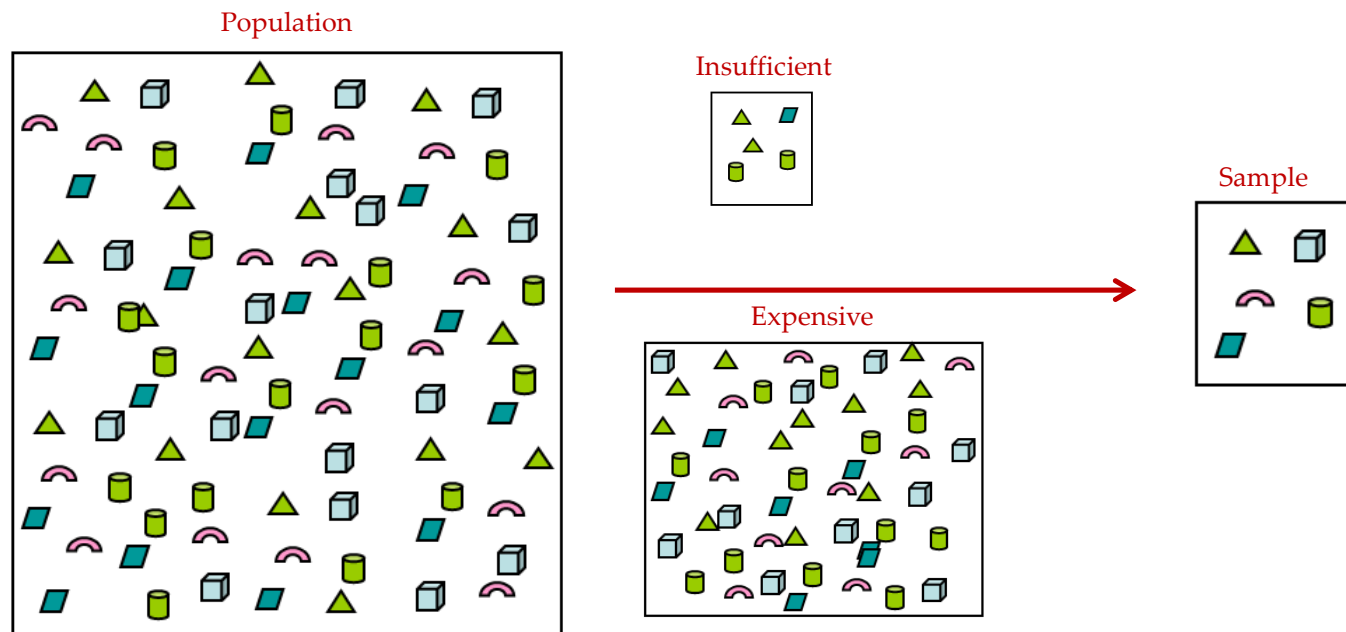
Which individuals: convenience, simple random, systematic, stratified, cluster, mixed

How many animals: detect the pathogen or find out the prevalence of the pathogen

How many animals

Statistical factors: size of population (N), estimated prevalence (p), precision (B), confidence (Z)

Non-statistical factors: manpower availability



Sample size

size of population (N)

estimated prevalence (p): data from other surveys

precision (E): error assumed

confidence (Z): Z=1.96 if 95%CI

Detect the pathogen

$$n = \left[1 - (1 - Z)^{1/D} \right] \left[N - \frac{D - 1}{2} \right]$$

n: sample size

D: minimum number of affected individuals expected

Example

Sample size to detect if there is *Yersinia* spp in a farm of 100 pigs, guessing that if *Yersinia* is present, at least 25% of pigs are infected

$$n = \left[1 - (1 - Z)^{1/D} \right] \left[N - \frac{D - 1}{2} \right] \quad n = \left[1 - (1 - 0.95)^{1/25} \right] \left[100 - \frac{25 - 1}{2} \right] = 9.9 \quad n = 10 \text{ pigs}$$

n: sample size

Z: confidence

D: number of infected individuals

N: population size

Table. (i) Sample size required for detecting disease where the probability of finding at least one case in the sample is 0.95; (ii) upper 95% confidence limits for number of cases (From Cannon and Roe, 1992)

Population size (N)	(i) Percentage of diseased animals in population (d/N) OR (ii) Percentage sampled and found clean (n/N)											
	50%	40%	30%	25%	20%	15%	10%	5%	2%	1%	0.5%	0.1%
10	4	5	6	7	8	10	10	10	10	10	10	10
20	4	6	7	9	10	12	16	19	20	20	20	20
30	4	6	8	9	11	14	19	26	30	30	30	30
40	5	6	8	10	12	15	21	31	40	40	40	40
50	5	6	8	10	12	16	22	35	48	50	50	50
60	5	6	8	10	12	16	23	38	55	60	60	60
70	5	6	8	10	13	17	24	40	62	70	70	70
80	5	6	8	10	13	17	24	42	68	79	80	80
90	5	6	8	10	13	17	25	43	73	87	90	90
100	5	6	9	10	13	17	25	45	78	96	100	100
120	5	6	9	10	13	18	26	47	86	111	120	120
140	5	6	9	11	13	18	26	48	92	124	139	140
160	5	6	9	11	13	18	27	49	97	136	157	160
180	5	6	9	11	13	18	27	50	101	146	174	180
200	5	6	9	11	13	18	27	51	105	155	190	200
250	5	6	9	11	14	18	27	53	112	175	228	250
300	5	6	9	11	14	18	28	54	117	189	260	300
350	5	6	9	11	14	18	28	54	121	201	287	350
400	5	6	9	11	14	19	28	55	124	211	311	400
450	5	6	9	11	14	19	28	55	127	218	331	450
500	5	6	9	11	14	19	28	56	129	225	349	500
600	5	6	9	11	14	19	28	56	132	235	379	597
700	5	6	9	11	14	19	28	57	134	243	402	691
800	5	6	9	11	14	19	28	57	136	249	421	782
900	5	6	9	11	14	19	28	57	137	254	437	868
1000	5	6	9	11	14	19	29	57	138	258	450	950
1200	5	6	9	11	14	19	29	57	140	264	471	1102
1400	5	6	9	11	14	19	29	58	141	269	487	1236
1600	5	6	9	11	14	19	29	58	142	272	499	1354
1800	5	6	9	11	14	19	29	58	143	275	509	1459
2000	5	6	9	11	14	19	29	58	143	277	517	1553
3000	5	6	9	11	14	19	29	58	145	284	542	1895
4000	5	6	9	11	14	19	29	58	146	268	556	2108
5000	5	6	9	11	14	19	29	59	147	290	564	2253
6000	5	6	9	11	14	19	29	59	147	291	569	2358
7000	5	6	9	11	14	19	29	59	147	292	573	2437
8000	5	6	9	11	14	19	29	59	147	293	576	2498
9000	5	6	9	11	14	19	29	59	148	294	579	2548
10 000	5	6	9	11	14	19	29	59	148	294	581	2588
∞	5	6	9	11	14	19	29	59	149	299	598	2995

Sample size

size of population (N)

estimated prevalence (p): data from other surveys

precision (E): error assumed

confidence (Z): Z=1.96 if 95%CI

Detect the pathogen

$$n = \left[1 - (1 - Z)^{1/D} \right] \left[N - \frac{D - 1}{2} \right]$$

n: sample size

D: minimum number of affected individuals expected

Prevalence

$$n = \frac{z^2 p (1 - p)}{E^2}$$

n: sample size

but, in small populations: n' corrected sampled size

$$\frac{1}{n'} = \frac{1}{n} + \frac{1}{N}$$

Table. The approximate sample size required to estimate prevalence in a large population with the desired fixed width confidence limits (modified from Cannon and Roe, 1982).

Expected prevalence	Level of confidence								
	90%			95%			99%		
	Desired absolute precision			Desired absolute precision			Desired absolute precision		
	10%	5%	1%	10%	5%	1%	10%	5%	1%
10%	24	97	2435	35	138	3457	60	239	5971
20%	43	173	4329	61	246	6147	106	425	10 616
30%	57	227	5682	81	323	8067	139	557	13 933
40%	65	260	6494	92	369	9220	159	637	15 923
50%	68	271	6764	96	384	9604	166	663	16 587
60%	65	260	6494	92	369	9220	159	637	15 923
70%	57	227	5682	81	323	8067	139	557	13 933
80%	43	173	4329	61	246	6147	106	425	10 616
90%	24	97	2435	35	138	3457	60	239	5971

Example

100 (N) individuals in population, precision of 5% (B) and confidence of 95% (z=1.96), guessed that 30% (p) individuals are infected.

$$n = \frac{z^2 p(1-p)}{E^2}$$

n: sample size

p: estimated prevalence

Z: confidence

D: number of infected individuals

N: population size

n': corrected sample size

$$n = \frac{1.96^2 \times 0.3 \times (1 - 0.3)}{0.05^2} = 322.7$$

n = 323

$$\frac{1}{n'} = \frac{1}{n} + \frac{1}{N}$$

$$\frac{1}{n'} = \frac{1}{323} + \frac{1}{100} = \frac{1}{0.131} = 76.36$$

n = 77

WinEpi



Example

Sample size for one sample, continuous outcome

An investigator wants to estimate the mean antibody levels of *Yersinia* in pigs which are between 5 months and slaughtered age. How many individuals should be enrolled in the study?

The investigator plans on using a 95% confidence interval and wants a margin of error of 5. The standard deviation of antibody levels is unknown, but the investigators conduct a literature search and find that the standard deviation in slaughtered-age pigs is between 15 and 20. To estimate the sample size, we consider the larger standard deviation in order to obtain the most conservative (largest) sample size. What if the standard deviation was the lowest?

$$n = \left(\frac{Z\sigma}{E} \right)^2$$

Population size: unknown
Expected SD: 20
Expected absolute error: 5
Level of confidence: 95%

Result, sample size = 62

Summary

Characteristics			Sample size
Outcome variable	No. of groups	Confidence Interval	Tests hypothesis
continuous	one sample	$n = \left(\frac{Z\sigma}{E}\right)^2$	$n = \left(\frac{Z_{1-\alpha/2} + Z_{1-\beta}}{ES}\right)^2$ $ES = \frac{ \mu_1 - \mu_0 }{\sigma}$
continuous	two independent samples	$n_i = 2\left(\frac{Z\sigma}{E}\right)^2$	$n_i = 2\left(\frac{Z_{1-\alpha/2} + Z_{1-\beta}}{ES}\right)^2$ $ES = \frac{ \mu_1 - \mu_2 }{\sigma}$
continuous	two matched samples	$n = \left(\frac{Z\sigma_d}{E}\right)^2$	$n = \left(\frac{Z_{1-\alpha/2} + Z_{1-\beta}}{ES}\right)^2$ $ES = \frac{\mu_d}{\sigma_d}$
dichotomous	one sample	$n = p(1-p)\left(\frac{Z}{E}\right)^2$	$n = \left(\frac{Z_{1-\alpha/2} + Z_{1-\beta}}{ES}\right)^2$ $ES = \frac{ p_1 - p_0 }{\sqrt{p_0(1-p_0)}}$
dichotomous	two independent samples	$n_i = \{p_1(1-p_1) + p_2(1-p_2)\}\left(\frac{Z}{E}\right)^2$	$n_i = 2\left(\frac{Z_{1-\alpha/2} + Z_{1-\beta}}{ES}\right)^2$ $ES = \frac{ p_1 - p_2 }{\sqrt{p(1-p)}}$

Types of sampling (which sample)

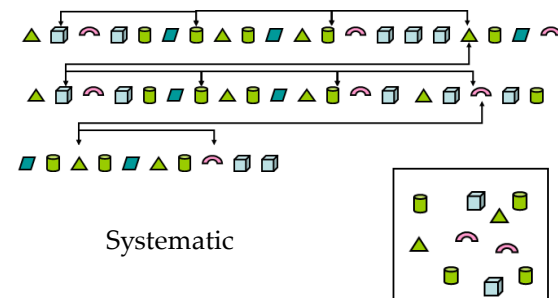
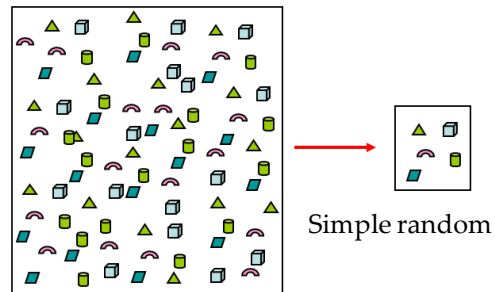
Non probability sampling

Types of sampling (which sample)

Non probability sampling

Probability sampling

- Simple random sampling
- Systematic sampling

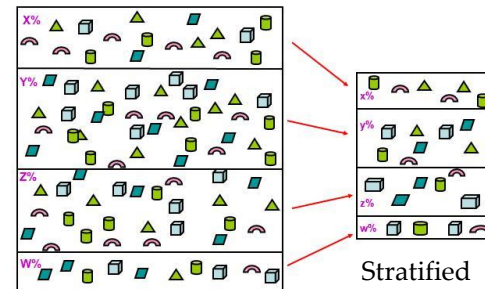


Types of sampling (which sample)

Non probability sampling

Probability sampling

- Simple random sampling
- Systematic sampling
- Stratified sampling

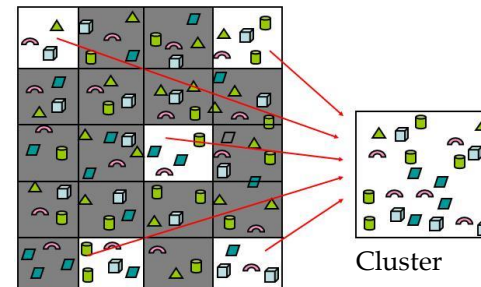


Types of sampling (which sample)

Non probability sampling

Probability sampling

- Simple random sampling
- Systematic sampling
- Stratified sampling
- Cluster sampling
- Mixed sampling



Questionnaires

A set of written questions

To record information in a standard format

Several uses

Basic

Which is the objective and the information needed, who has the information,
method to gather data, make pilot questionnaire

Classification of data

Qualitative (categorical): nominal and ordinal

Quantitative: discrete or continuous

Types of studies

Experimental (clinical trials)

Observational

cross-sectional (prevalence, incidence)

longitudinal: case control, cohort (risk factors)

	Infected	Non-infected	Total
Hypothesized risk factor present	a	b	a+b
Hypothesized risk factor absent	c	d	c+d
Total	a+c	b+d	n =a+b+c+d

Cohort studies: (a+b) and (c+d) are predetermined

Case-control studies: (a+c) and (b+d) are predetermined

Cross-sectional studies, only n can be predetermined

Clinical studies

Target or reference population *vs.* experimental (study) population

Inclusion and exclusion criteria

Internal and external validity

Sample size?

Analysis: similar to cohort studies

Cohort studies

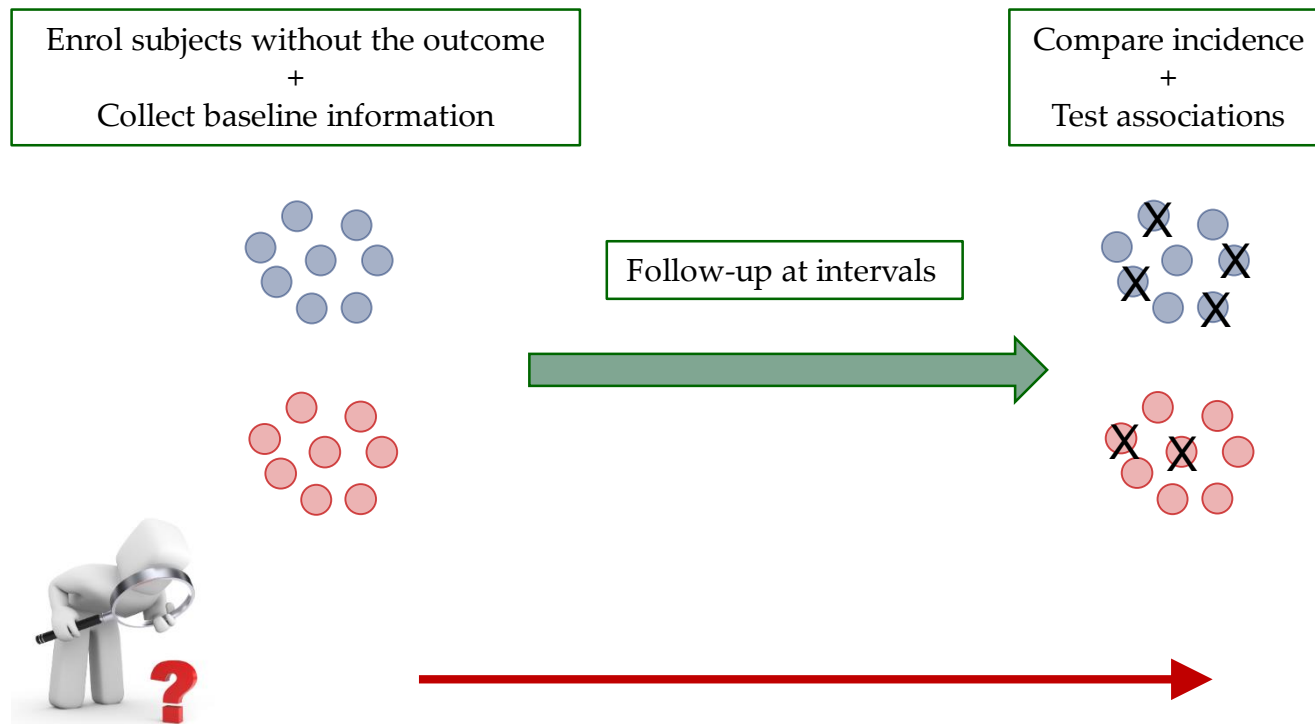
Estimate the magnitude of association between the exposure and the outcome

None of the subjects have the outcome of interest at the beginning,
and time must pass to determine the frequency of developing the outcome

Inclusion criteria

Cohort studies

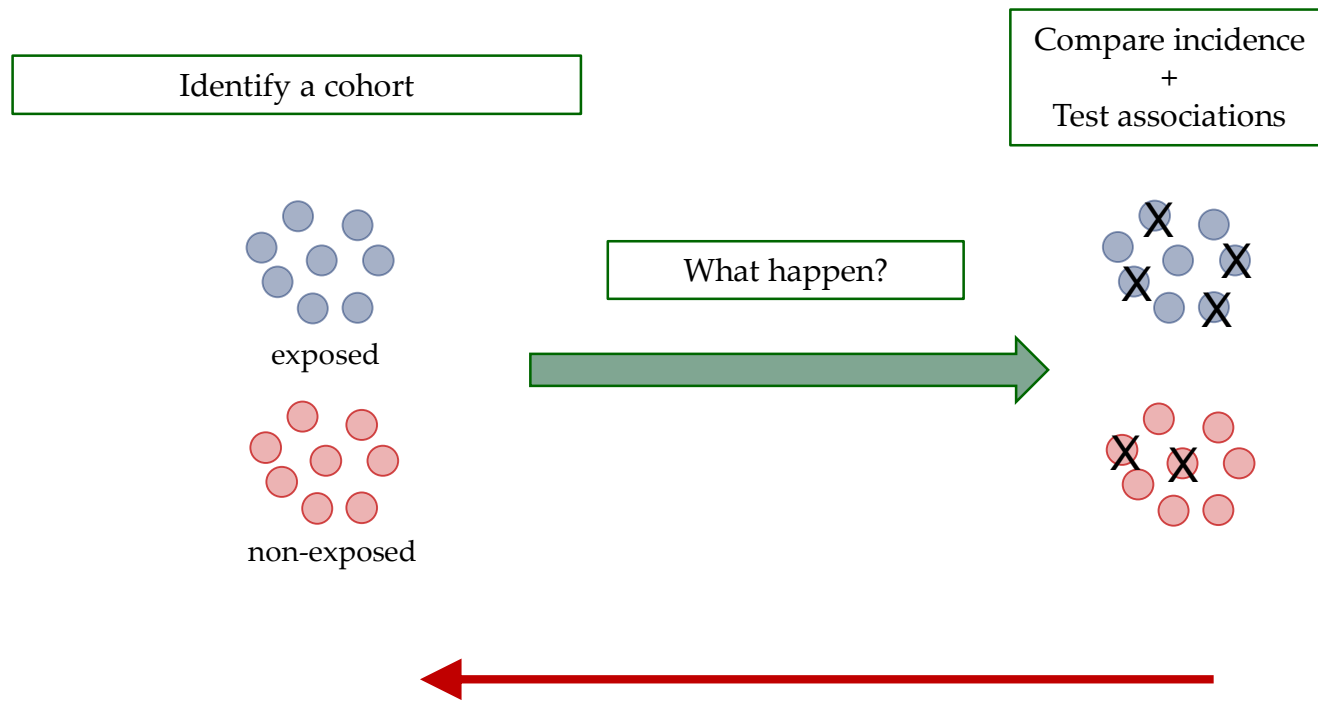
Prospective: baseline data collected before the outcome



Cohort studies

Prospective: baseline data collected before the outcome

Retrospective: collect data after outcomes have occurred



Cohort studies

Key features

- None of the subjects have the outcome at the beginning
- The groups differ in their exposure status
- Measure and compare the incidence of the outcome to determine whether there is an association between exposure and outcome

Cohort studies

Advantages

- indicate the temporal sequence between exposure and outcome
- allow to calculate the incidence of the outcome
- useful for evaluating the effects of rare or unusual exposures
- enables examination of multiple outcomes of a single risk factor

Disadvantages

Prospective

- may have to follow large numbers of subjects for a long time, expensive and time consuming
- not good for rare events

Retrospective

- data of poor quality if record were not designed for the study
- absence of data on potential confounding factors

Case-control studies

Estimate the magnitude of association between the exposure and the outcome

First, identify subjects with the outcome of interest

No measure of incidence  no calculate relative risk

Useful when the outcome is rare or uncommon or little is know about the risk factors

Case-control studies

Definition of cases clear and specific

Source :

- Cases, usually from cross-sectional studies
- Controls, representative of the population and sampled independently

Tips on identifying a study design

1. Is it based on information about individuals or averages in populations?

→ Correlational (Ecologic)

2. Is there just one group? Did all subjects have the disease?

→ Case series

3. Evaluation of outcome and risk factors at the same point in time?

→ Cross-sectional survey

Two or more groups being compared?

4. How were they selected?

- a. find cases and then controls
- b. find a group and follow longitudinally

→ { Case-control
Cohort study } → { Prospective
Retrospective }


5. Assignment of subjects to a intervention and follow to compare outcomes?

→ Clinical trial


Comparison of the advantages and disadvantages of different studies

	<i>Advantages</i>	<i>Disadvantages</i>
Cohort studies	<ol style="list-style-type: none"> 1. Incidence in exposed and unexposed individuals can be calculated 2. Permit flexibility in choosing variables to be systematically recorded 	<ol style="list-style-type: none"> 1. Exposed and unexposed proportions in target population cannot be estimated 2. Large numbers of subjects are required to study rare diseases 3. Potentially long duration for follow-up 4. Relatively expensive to conduct 5. Maintaining follow-up is difficult 6. Control of extraneous variables may be incomplete
Case-control studies	<ol style="list-style-type: none"> 1. Well suited to the study of rare diseases or of those with long incubation periods 2. Relatively quick to mount and conduct 3. Relatively inexpensive 4. Requires comparatively few subjects 5. Existing records occasionally can be used 6. No risk to subjects 7. Allow study of multiple potential causes of a disease 8. Suited to the study of interaction between genotype and environmental factors 	<ol style="list-style-type: none"> 1. Exposed and unexposed proportions in target populations cannot be estimated 2. Rely on recall or records for information on past exposures 3. Validation of information is difficult or sometimes impossible 4. Control of extraneous variables may be incomplete 5. Selection of an appropriate comparison group may be difficult 6. Incidence in exposed and unexposed individuals cannot be estimated
Cross-sectional studies	<ol style="list-style-type: none"> 1. When a random sample of the target population is selected, disease prevalence, and proportions exposed and unexposed in the target population, can be estimated 2. Relatively quick to mount and conduct 3. Relatively inexpensive 4. Current records occasionally can be used 5. No risk to subjects 6. Allow study of multiple potential causes of disease 	<ol style="list-style-type: none"> 1. Unsited to the study of rare diseases 2. Unsited to the study of diseases of short duration 3. Control of extraneous variables may be incomplete 4. Incidence in exposed and unexposed individuals cannot be estimated 5. Temporal sequence of cause and effect cannot necessarily be determined

Bias

- Random error (sampling error or selection bias)  inaccurate estimations
- Systematic error
 - Misclassification
 - Confounding

Bias

- Random error (sampling error or selection bias)  inaccurate estimations
- Systematic error
 - Misclassification
 - Confounding