

## Inference & Sample Size (I)



- Assessing (**statistical**) *significance* in a 2X2 Table
- Sampling distribution and confidence intervals
- Sample size based on *precision*

# Assessing *significance* in a 2X2 Table

We focus on the assessment of whether an observed association of D and E in a sample of data reflects a population in which D and E are *truly* associated or may have arisen from the vagaries of *random variation* (by chance).

In the language of hypothesis testing, a suitable *null hypothesis* ( $H_0$ ) is that **D and E are independent**

$$H_0: D \text{ and } E \text{ independent} \leftrightarrow RR = 1 \leftrightarrow OR = 1$$

	<i>D</i>	$\bar{D}$	
<i>E</i>	a	b	a+b
$\bar{E}$	c	d	c+d
	a+c	b+d	n=a+b+c+d

Population-based design:

$H_0: P(D\&E) = P(D) * P(E)$

Expected proportions under independence can be estimated from sample proportions

Exp-based design:

$H_0: P(D|E) = P(D|\bar{E})$

Estimable from the exposure samples. Independence then simplifies to the comparison between two separate population proportions


Disease-based design:

$H_0: P(E|D) = P(E|\bar{D})$


The only probabilities that are estimable are exposure probabilities, conditional on disease status.

	$D$	$\bar{D}$	
$E$	$O_{11} = a$	$O_{12} = b$	$a+b$
$\bar{E}$	$O_{21} = c$	$O_{22} = d$	$c+d$
	$a+c$	$b+d$	$n$

Chi-square test\*



$$\chi^2_1 = \sum_{i=1}^2 \sum_{j=1}^2 \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$



**Expected** frequencies of the cell **under the assumption of independence**

\*Fisher exact test for small sample size

The **p-value** for a  $\chi^2$  test **does not represent the probability** that Relative Risk/Odds Ratio is as far as or further from independence (RR/OR=1).

The **p-value** is not the **power** of the study, that is, the **probability of rejecting the null** when it is actually false (**power: 1-type II error**).

		True state of H <sub>0</sub> (Unknown)	
		H <sub>0</sub> true	H <sub>0</sub> false
Decision (sample data)	Reject H <sub>0</sub>	Type I error*	ok
	Do Not reject H <sub>0</sub>	ok	Type II error**

How should we then interpret the p-values ??

How **surprising** are the observed data **under the assumption** that the null hypothesis is **true**...

One way to *minimize* the use of p-values is to focus on **estimation of effects**, rather than testing null values

The goal of an epidemiological study is the **estimation of effects** rather than mere assessment of the independence or non independence of an exposure and disease outcome.

- How **large** is the effect of prenatal care on the incidence of low birthweight babies?
- How **effective** is a vaccine in preventing a disease?
- How **strong** is the association between elevated dietary fat and heart disease?

We now focus on the **inference** about measures of association, according to the different types of designs.

A key component is determination of the **level of uncertainty** associated with proposed estimators, expressed via calculation of **confidence intervals**.

# Sampling distribution of the relative risk

The ingredients of the Relative Risk are the two conditional probabilities  $P(D|E)$  and  $P(D|\bar{E})$  both of which **can be directly estimated** from either a population based/exposure-based study.

	$D$	$\bar{D}$	
$E$	a	b	a+b
$\bar{E}$	c	d	c+d
	a+c	b+d	n=a+b+c+d

$$\widehat{RR} = \frac{a/(a+b)}{c/(c+d)}$$

$$\frac{\log(\widehat{RR}) - \log(RR)}{\sqrt{\widehat{var}(\log(\widehat{RR}))}} \sim N(0,1)$$

$$\widehat{var}(\log(\widehat{RR})) = \frac{b}{a(a+b)} + \frac{d}{c(c+d)}$$

$$\log(\widehat{RR}) \pm z_{\alpha} \sqrt{\widehat{var}(\log(\widehat{RR}))}$$

	CHD YES	CHD NO	tot
TYPE A	178	1141	1319
TYPE B	79	1486	1565
tot	257	2627	2884

Western Collaborative Group Study (WCGS): population-based study on a group of employed men from 10 Californian companies, regarding the onset of **coronary heart disease** (CHD).

Interest focused on several possible **risk factors** including lifestyle variables and certain **behavioral** characteristics.

Particular attention was paid to behavior type, a binary variable whose two levels are referred to as **Type A** and **Type B**.

**Type A** behavior is characterized by **aggressiveness** and **competitiveness**, whereas **Type B** behavior is considered more **relaxed** and **noncompetitive**.

	CHD YES	CHD NO	tot
TYPE A	a=178	b=1141	1319
TYPE B	c=79	d=1486	1565
tot	257	2627	2884

→ from twice to close to three times the risk for CHD in Type A individuals

$$\widehat{RR} = \frac{178/1319}{79/1565} = 2.67$$

$$\widehat{var}(\log(\widehat{RR})) = 0.017$$

$$0.98 \pm 1.96 * \sqrt{0.017}$$

95% CI on log scale : (0.724, 1.235)

95% CI:  $(e^{0.724}, e^{1.235}) = (2.06, 3.44)$

Fox BH, et al., *Type A behavior and cancer mortality. Theoretical considerations and preliminary data.* Ann N Y Acad Sci. 1987

Rosenman RH: *An Update on the Type A Behavior Pattern and its Relationship to Coronary Artery Disease.* Perspect Lipid Disorders, 1987



# Sampling distribution of the Odds Ratio

	<i>D</i>	<i>D̄</i>	
<i>E</i>	a	b	a+b
<i>Ē</i>	c	d	c+d
	a+c	b+d	n=a+b+c+d

**Independently** from the study design, the sampling distribution of the OR is always the same.

$$OR = \frac{\frac{a}{(a+b)} \big/ \frac{b}{(a+b)}}{\frac{c}{(c+d)} \big/ \frac{d}{(c+d)}} = \frac{a}{(a+b)} \cdot \frac{(a+b)}{b} \cdot \frac{(c+d)}{c} \cdot \frac{(c+d)}{d} = \frac{ad}{bc}$$

Disease-based/Case-control

Population-based/exp-based

$$\frac{\log(\widehat{OR}) - \log(OR)}{\sqrt{\widehat{var}(\log(\widehat{OR}))}} \sim N(0,1)$$

$$\widehat{var}(\log(\widehat{OR})) = \frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}$$

$$\log(\widehat{OR}) \pm z_{\alpha} \sqrt{\widehat{var}(\log(\widehat{OR}))}$$

$$z_{\alpha} = 1 - \frac{\alpha}{2} \text{ percentile std Normal dist}$$

Case-control study

		Pancreatic Cancer		
		Cases	Controls	
Coffee drinking (cups per day)	>1	347	555	902
	0	20	88	108
		367	643	1010

Case-control study of pancreatic cancer and its relationship to various lifestyle habits including consumption of tobacco, alcohol, tea, and coffee.

$$\widehat{OR} = \frac{347 * 88}{555 * 20} = 2.75$$

➡

$$95\% \text{ CI: } (e^{0.508}, e^{1.516}) = (1.66, 4.55)$$

$$\widehat{var}(\log(\widehat{OR})) = 0.066$$

$$1.01 \pm 1.96 * \sqrt{0.066}$$

➡

$$95\% \text{ CI on log scale : } (0.508, 1.516)$$



Substantial and **statistically significant increase** of risk (between 1.6 to 4.5 times) of pancreatic cancer for coffee drinkers compared to abstainers

# Sampling distribution of the excess risk

Estimation of the Excess Risk from population-based or exp based data parallels that of the Relative Risk, since ER also depends solely on  $P(D|E)$  and  $P(D|\bar{E})$ .

	CHD YES	CHD NO	tot
TYPE A	a=178	b=1141	1319
TYPE B	c=79	d=1486	1565
tot	257	2627	2884

$$\widehat{ER} = \frac{178}{1319} - \frac{79}{1565} = 0.08$$

$$0.08 \pm 1.96 * \sqrt{0.00012}$$

$$\widehat{Var}(\widehat{ER}) = 0.00012$$

$$95\% \text{ CI: } (0.06, 0.10)$$

$\widehat{ER} = 0.08$ , i.e. we would expect the CHD to increase by 8% if **all subjects** had Type A behaviour as compared to **all subjects** having Type B behaviour (95% CI: 6%-10%).

$$\widehat{ER} = \frac{a}{a + b} - \frac{c}{c + d}$$

$$\widehat{ER} \sim N(ER, Var(ER))$$

$$\widehat{Var}(\widehat{ER}) = \frac{ab}{(a + b)^3} + \frac{cd}{(c + d)^3}$$

$$\widehat{ER} \pm z_{\alpha} \sqrt{\widehat{var}(\widehat{ER})}$$



We have now quantified **uncertainty** in a study of a particular design\* and **given** sample size(s).

Turning these techniques on their heads, we can determine the **size of the sample(s)** required to achieve a **given level of precision** for a specific design.

Such calculations are referred to as **sample size** planning.



\*we will not treat variability of the estimates related to **different sampling schemes/design** in this course, such as case-cohort and nested case-controls studies.



**Validity:** avoid *systematic* error in the estimates (**bias**)

## Key objectives in a Study Design

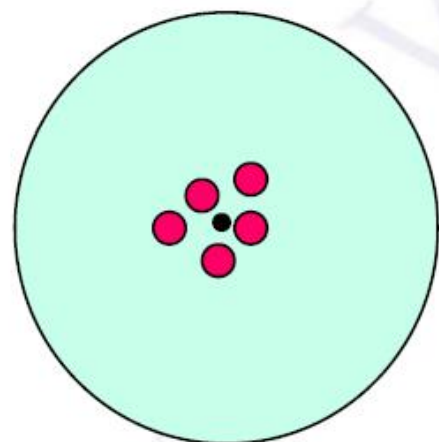
**Internal validity:** are the differences in the outcome caused by exposure or by a systematic error / presence of confounders?

**External validity:** is it possible to generalize what has been observed to other populations?

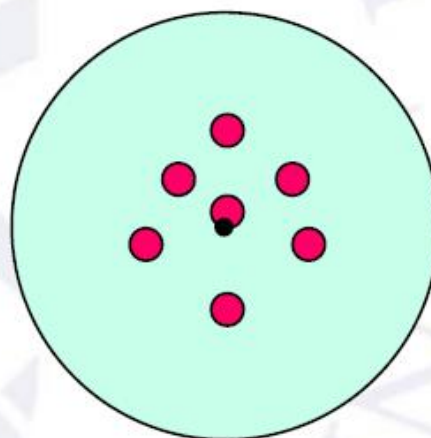
**Accuracy:** given the random variability of the estimates **minimize** the amplitude of the confidence intervals



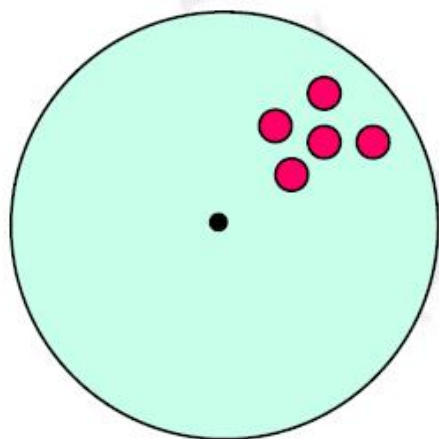
**Sample size** / group balance / use of covariates...



Valid and accurate



Valid but not accurate

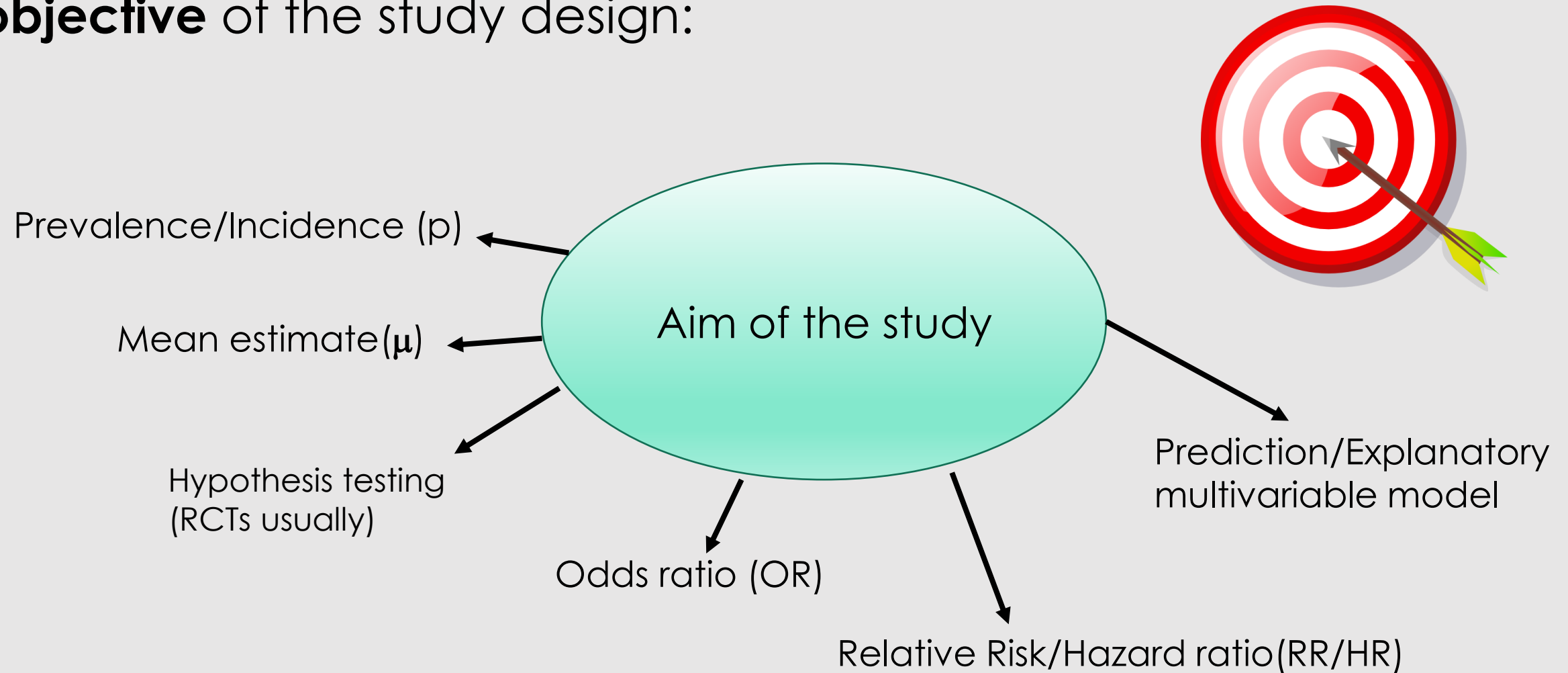


Not valid but accurate

**Validity:** degree of agreement between the estimate and the *true* value

**Accuracy:** degree of variability (dispersion) between estimates and the *true* value

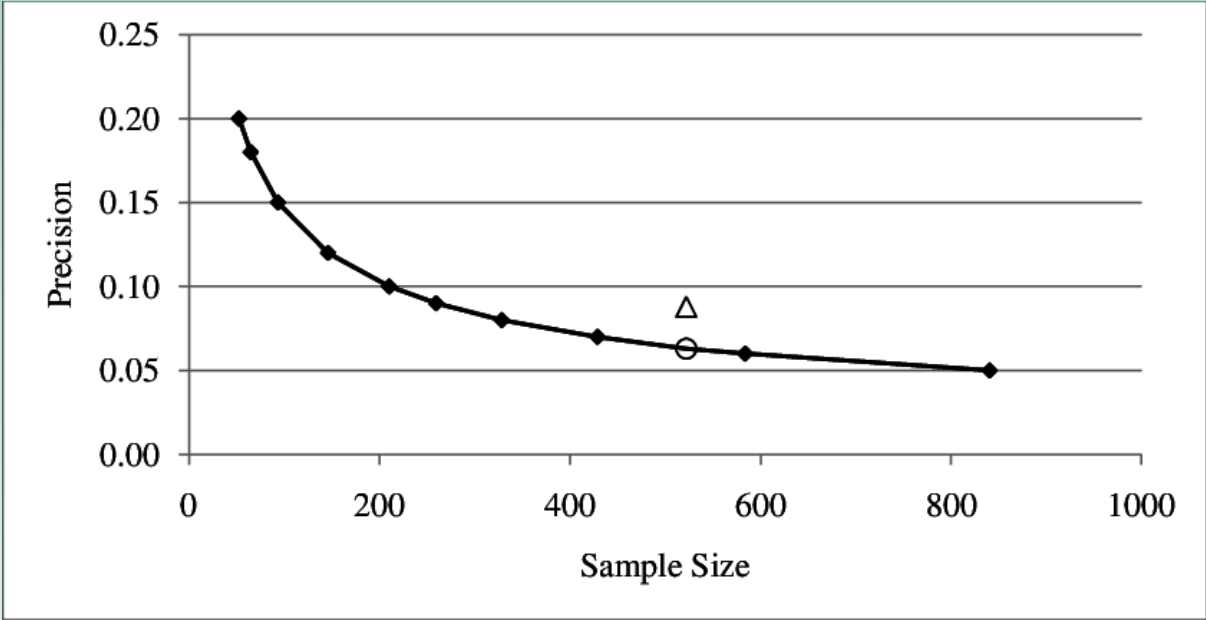
Sample size calculations **depend on the primary objective** of the study design:



# Two basic strategies for sample size



## Precision (confidence intervals)



## Power of the statistical test (*effect size*)

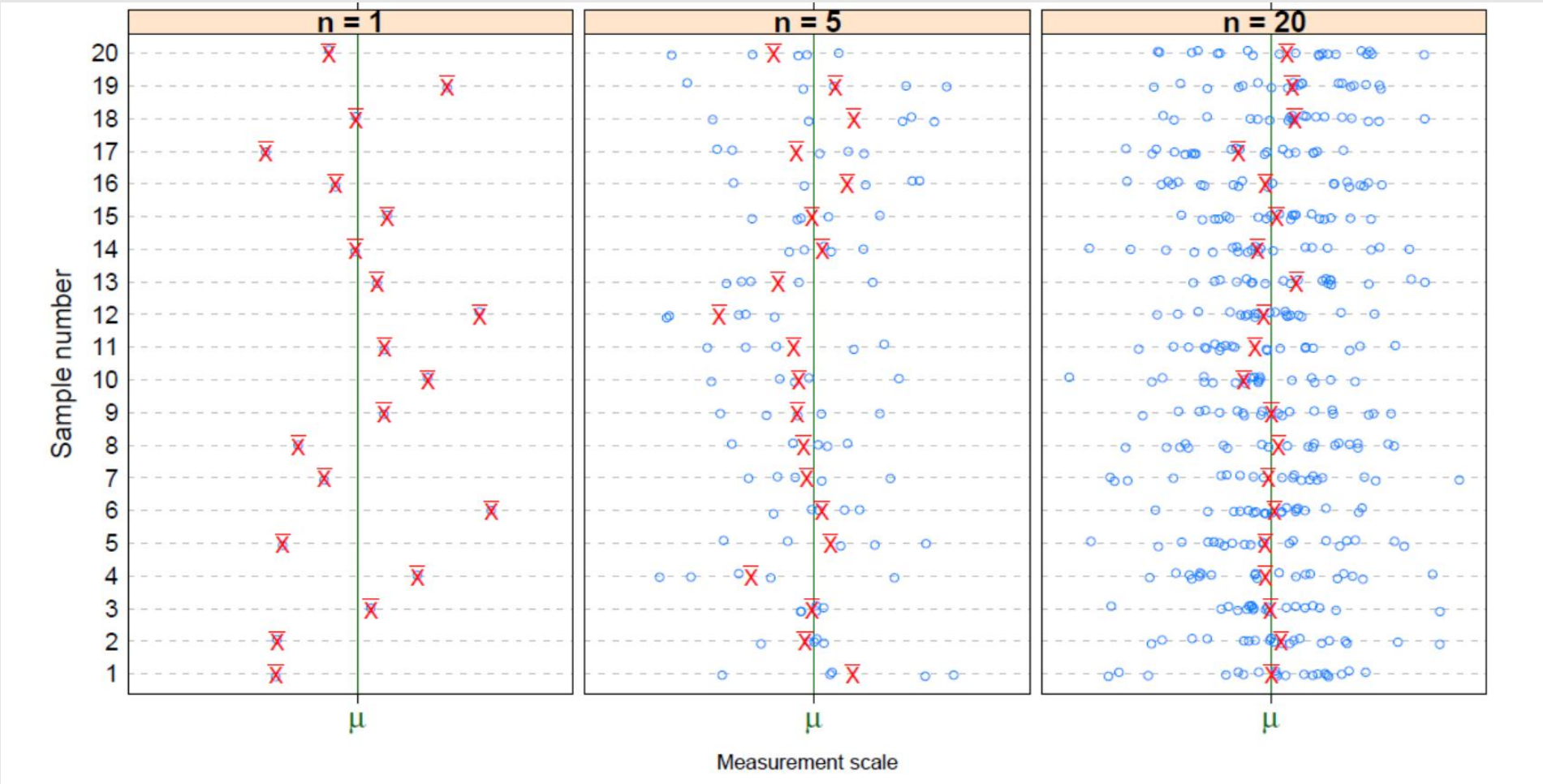
		True state of H <sub>0</sub> (Unknown)	
		H <sub>0</sub> true	H <sub>0</sub> false
Decision (sample data)	Reject H <sub>0</sub>	Type I error*	ok
	Do Not reject H <sub>0</sub>	ok	Type II error**



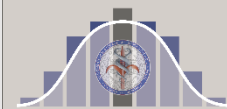
The precision strategy to determine sample size is mainly based on the **width of a desired confidence interval\*** (CI).

**Remind:** CI is a range of values around which a population parameter (e.g., true mean) is likely to lie **in the long run**.

For example, if samples of the same size are drawn **repeatedly** from a population and a 95% CI is calculated around each sample's mean then 95% of these intervals **should contain** the population mean.



\*We adopt for the sample size topic the **frequentist** point of view



For a  $(1 - \alpha) 100\%$  confidence interval, the precision of the interval depends on its **width**. The *narrower* the interval is, the more **precise** the inference is.

Precision analysis for sample size determination is to consider the maximum **half width** of the  $(1 - \alpha) 100\%$  confidence interval of the unknown parameter that **one is willing to accept**.

Note that the maximum half width of the confidence interval is usually referred to as the **maximum error** of an estimate of the unknown parameter.

# Estimate of a mean

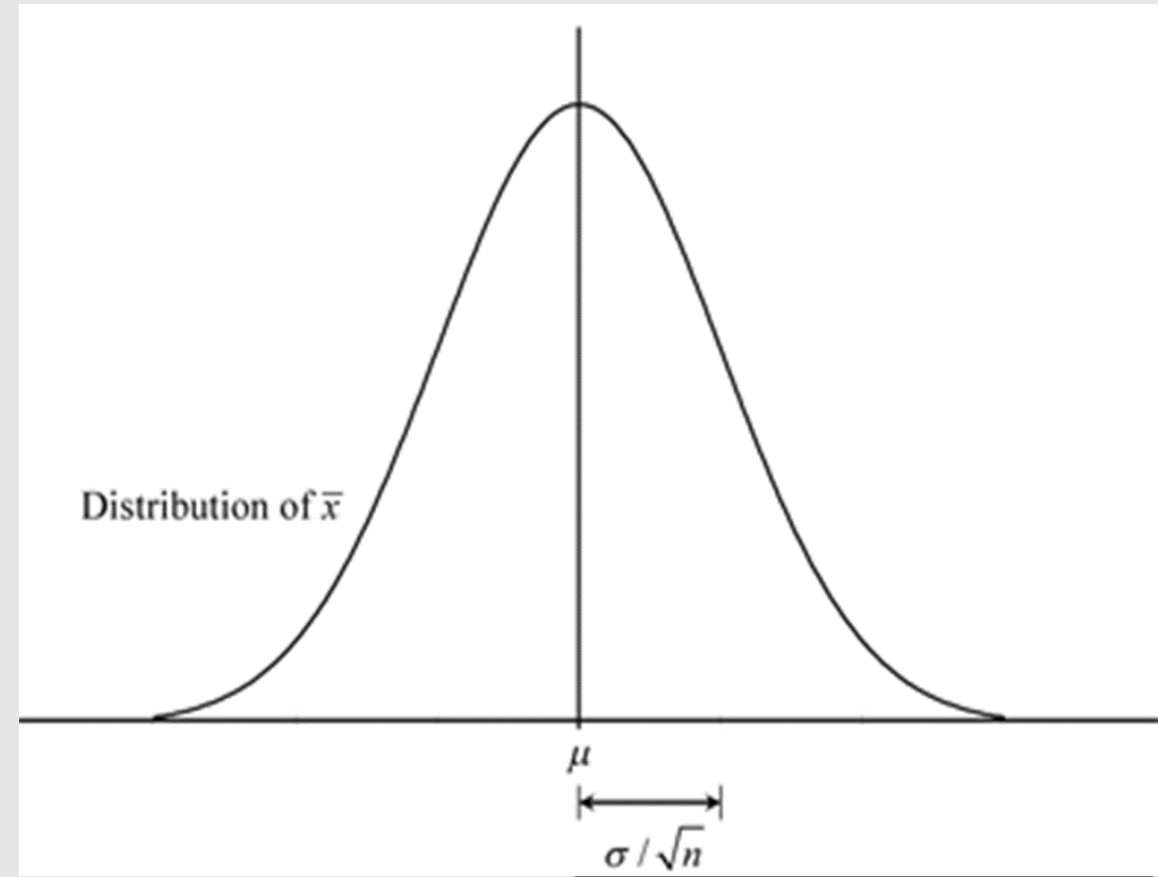
We can use the Central Limit Theorem  $\bar{y} \pm z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}$

$$z_{1-\frac{\alpha}{2}} = 1 - \frac{\alpha}{2} \text{ percentile } N(0,1)$$

The maximum error E in estimating the value of  $\mu$  that one is willing to accept could be then defined as:

$$E = |\bar{y} - \mu| = z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}$$

➔ 
$$n = \frac{z_{1-\frac{\alpha}{2}}^2 * \sigma^2}{E^2}$$



Suppose that we want to estimate the *average price* of tablets of a tranquilizer.

A random sample of pharmacies is selected. The estimate is required to be **within 10 cents** of the true average price with 95% confidence. *Based on a small pilot study*, the **standard deviation** in price can be estimated as 85 cents.

**How many** pharmacies should be randomly selected ?

$$n^* = \left( \frac{Z_{\frac{\alpha}{2}} * \sigma}{d^*} \right)^2 = \left( \frac{1.96 * 0.85}{0.10} \right)^2 = 277.56 \quad \longrightarrow \quad \text{a sample of } \mathbf{278} \text{ pharmacies should be taken.}$$

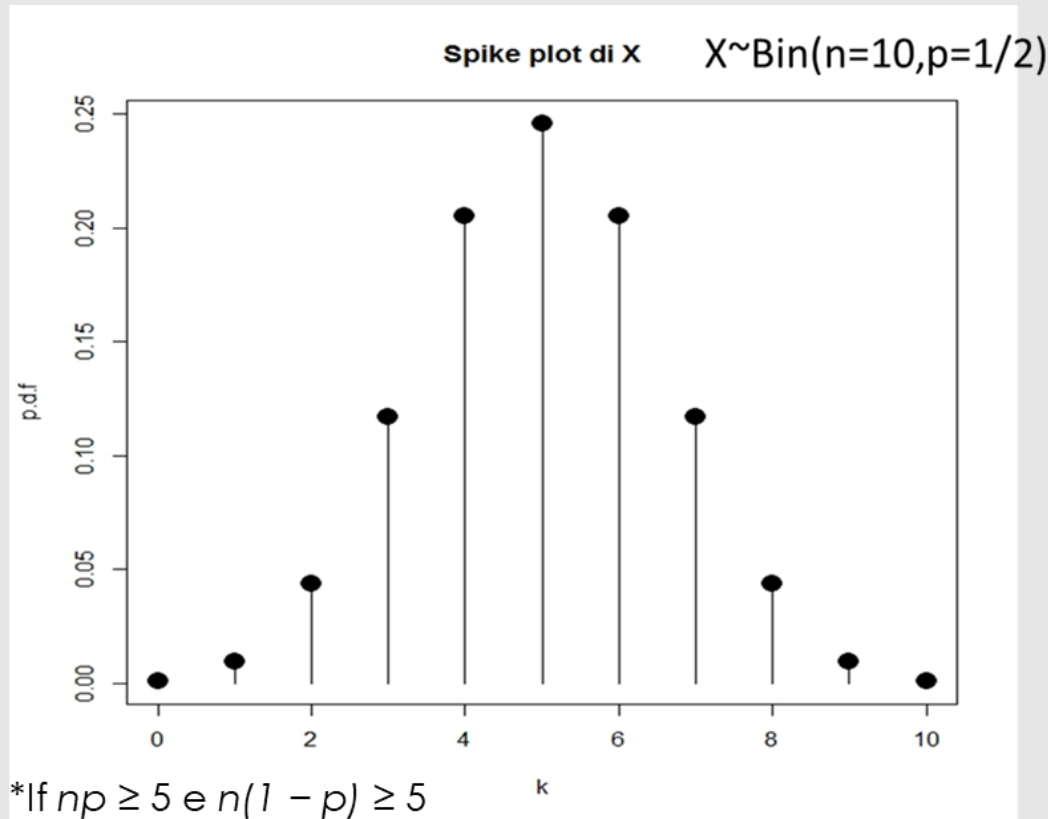
$\sigma^2$  is crucial for determining the optimal sample size. We can use previous studies or pilot studies.

However, it is advisable to **overestimate** the variance (rather than underestimate it) as it is better to use a sample size that is *too high* rather than one that is *too low*.

# Estimate of a proportion

Here too we must set the **precision** that we want for the estimate.

How large the sample has to be to estimate an (unknown) proportion  $p$  with a precision of  $E$  ?



$$Y \approx \text{Bin}(n, p)$$

$$Z = \frac{Y - np}{\sqrt{np(1-p)}} \rightarrow N(0,1)$$

$$n = \frac{z_{1-\frac{\alpha}{2}}^2 * p * (1-p)}{E^2}$$

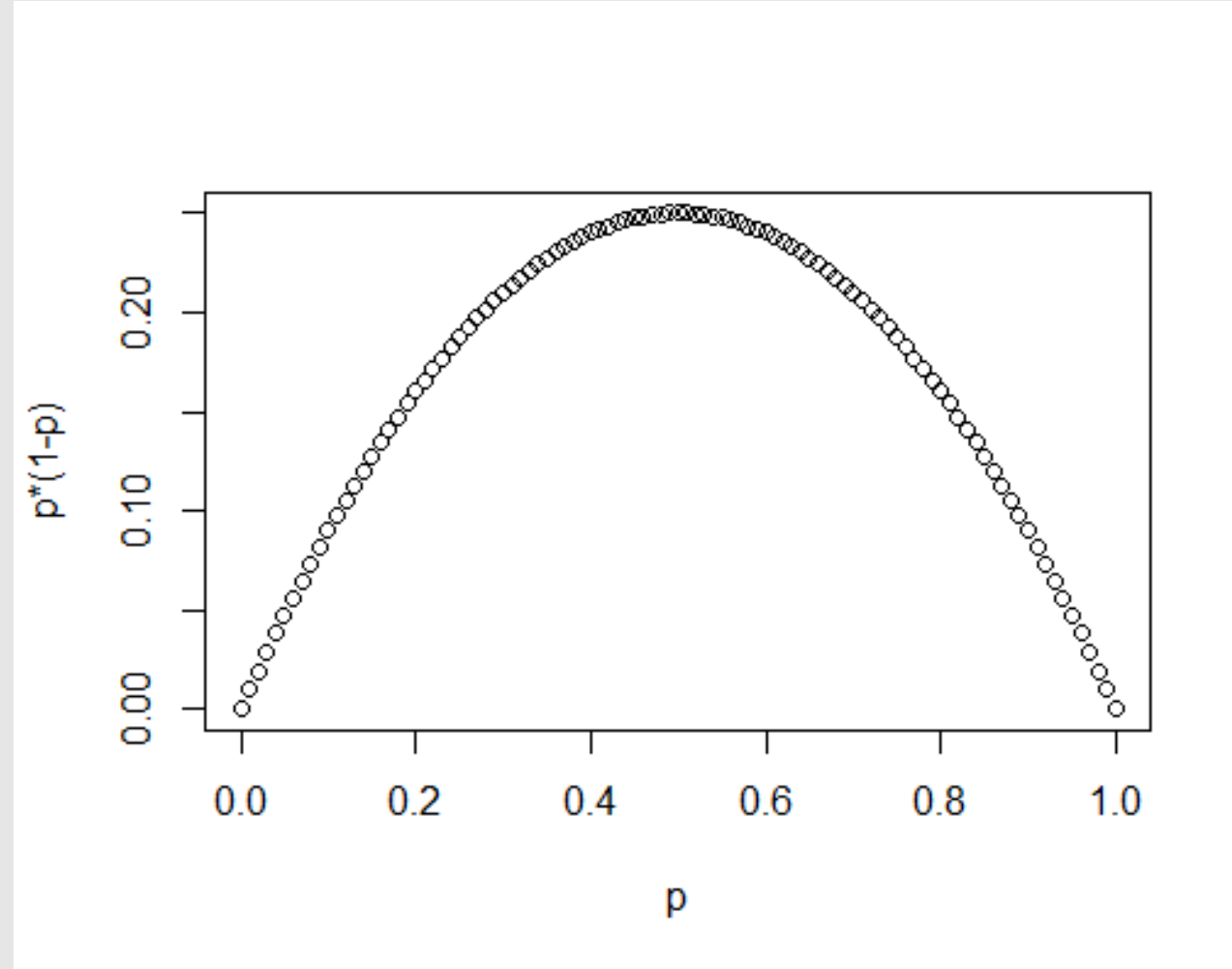
**N.B.** Assuming  $p = 0.5$  for the unknown proportion we will always have the "largest" sample possible

The **precision** of the estimator (and therefore the half width of the interval) depends on the value assumed by  $p$  which is unknown.

It could be appropriate to use (conservative estimate):  $p=0.5$  ;  $p^*(1-p)=0.25$

If we wish to calculate the minimum size required to have an interval for  $p$  that does not exceed the half maximum amplitude  $E^*$  we must look for the minimum value of  $n$  such that:

$$z_{\frac{\alpha}{2}} * \sqrt{\frac{0.25}{n}} \leq E^*$$



$$n^* = 0.25 * \left( \frac{Z_{\frac{\alpha}{2}}}{E^*} \right)^2$$

An epidemiological study is planned to estimate the **prevalence** of subjects with a certain disease and we want the confidence interval at level  $1-\alpha = 0.95$  not to exceed the maximum error ( $E^* = 0.01$ ) we will need a minimum of:

$$n^* = 0.25 * \left( \frac{1.96}{0.01} \right)^2 = 9604$$

at least 9604 subjects from the target population...

## Something more about Sample Size estimation for proportions

**Allowable error  $E^*$**  : should be acceptable to the clinician and decided a priori.

In any protocol or manuscript, it should be stated explicitly along with the basis for its choice.

Conventionally, an absolute allowable error margin of  **$\pm 5\%$**  is chosen, but, as is common in clinical practice, if expected proportion  $p$  is  $<10\%$ , the  $95\%$  confidence boundaries may cross 0, which is impractical.

Hence, for an expected  $p$  between 10 to 90 % ( $p=0.1$  to  $0.8$ )  $E^*$  of  $\pm 5\%$  might be a reasonable choice.

But for **rare** ( $p < 0.1$ ) or **very common** ( $p > 0.8$ ) conditions  $E^*$  should be chosen as a relative value with respect to expected  $p$ .

A common recommendation is to set  $E^* = p/2$  for rare and  $E^* = (1-p)/2$  for very common conditions.



# Estimate of the incidence rate

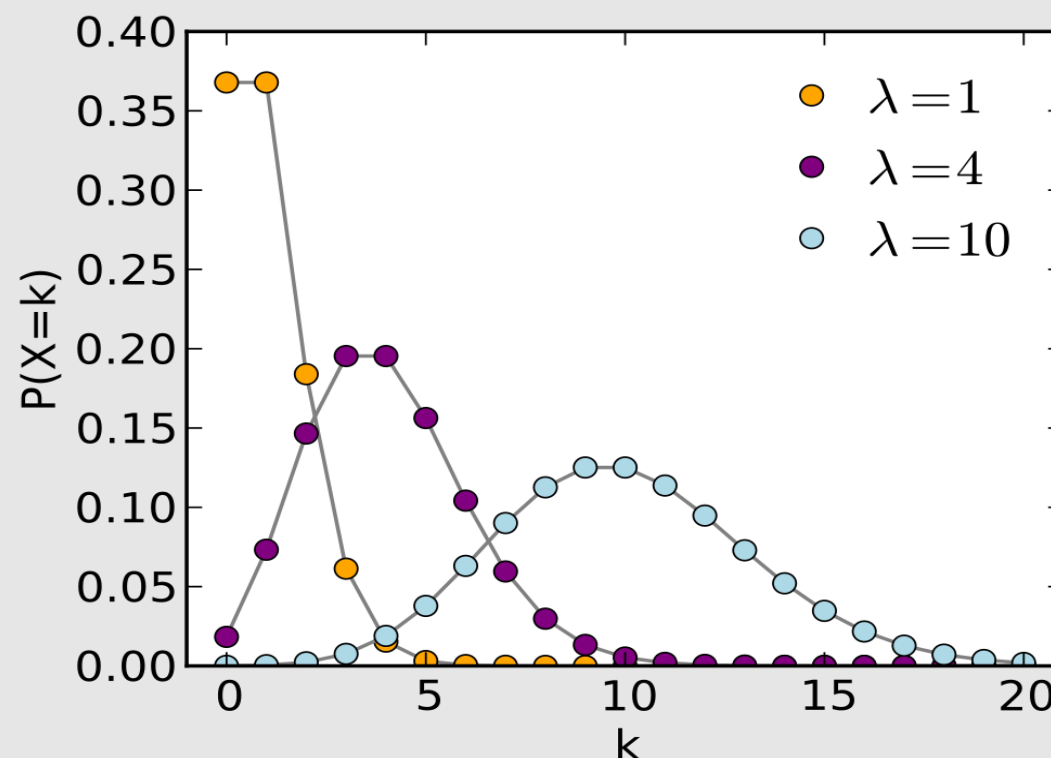
Remind from Block 1:

**Incidence rate** : the occurrence of **new** cases of disease that arise during **person-time** of observation.

$$\frac{\# \text{ New Cases}}{\# \text{ Person} - \text{ Time}}$$

$$P(X = k) = \frac{e^{-\lambda} \lambda^k}{k!}$$

$$E(X) = \text{Var}(X) = \lambda$$



The **incidence rate** could also be viewed as the **mean** (=variance) of a Poisson random variable (**the number of events** in a given time)

A similar approach to proportions (prevalence/cumulative incidence) can be used to estimate the sample size necessary to ensure that the confidence interval for an **incidence rate** is of a *pre-determined* width. The most commonly used method is based on the **normal approximation**.

$$D = \#cases$$

$$95\% CI = IR \pm 1.96 * se(IR)$$

$Y = \text{amount of } \textcolor{red}{p}time$

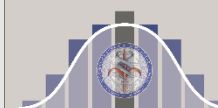
$$\lambda = \frac{D}{Y} = IR$$

Let's say we observed **250** cases of a disease in a population with **10,000 person-years** of follow-up. We want to calculate a 95% confidence interval for the incidence rate.

$$se(IR) = \frac{\sqrt{D}}{Y}$$

The 95% confidence interval for the incidence rate is **(219- 281)** cases per person-year.

For the sample size : we can **fix** the maximum error and then estimate the number of required cases D [**assuming an hypothetical constant rate per p-time**] in this way we obtain the Y required to estimate the rate with the given precision



## Sample size for the incidence rate based on precision

Now, if we want to estimate the sample size required in a study to estimate an incidence rate within a pre-specified precision, let's follow this example: based on data from previously conducted studies, we expect the rate to be about **50 per 10.000 pyrs.**

We want to determine the size of the sample that will be required to estimate the incidence rate in that population within  **$\pm 5$  per 10.000 pyrs.**

We are here imposing that the margin of precision ( $E=5$ ) should be  $E = 1.96 \cdot \text{se}(\text{IR})$ , so we can derive the standard error of the rate as:

$$\text{se.rate} = (5/1.96)$$

$$\text{number.cases} = (50/\text{se.rate})^{**2}$$

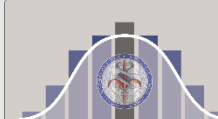
$$\text{number.cases} = 384.16$$

Person-time necessary to observe that number of cases as:

$$\text{person.years} = (\text{number.cases}/50) * 10000$$

$$76832$$

Therefore, we should follow **76832** subjects **for one year** (or **38416** for **two years...etc**) in order to observe **384** events and be able to estimate a 95% confidence interval of the required precision.



## General considerations about sample size

The **feasibility** of a study often rests on whether the projected **number** of accrued patients is **adequate** to address the **scientific aim** of the research.

Accordingly, a **rationale** for the **planned study size** should be provided.

Without the context of a **numeric rationale for the study size**, readers may misinterpret the lack of a *statistically significant* difference in effect as false reassurance of lack of harm, or falsely conclude that there is no benefit when comparing two interventions. Moreover the uncertainty around the estimates (prevalence, incidence... association measures) could be too large to be useful.

Institutional review boards (IRBs) nowadays **require** such statements in a study protocol before the data collection can begin, and **also** if secondary data sources will be used.

<https://arcs.sanita.fvg.it/it/utenti/aziende-sanitarie/comitato-etico-unico-regionale-ceur-fvg-copy/indicazioni-per-i-promotori/>

# General considerations about sample size

Sample size computations could seem somewhat *artificial*.

- available resources are often **constrained**, and sample size calculations are sometimes used to justify the value of a study, **given fixed resources**, as compared with precision assessment driving appropriate fund allocations.
- sample size planning rarely accounts for **all sources of error**, some of which may be a far greater threat than *sampling variability*.

For example, it may be more effective to expend a greater fraction of resources ensuring the **quality of measurement** of exposure and disease than to merely increase the sample size for a study with **inaccurate** data.

It is particularly dangerous to blindly resort to sample size formulas without fully understanding the statistical nuances of a planned design and analysis strategy...

# SUPPLEMENTARY MATERIALS

# Sampling distribution of the attributable risk AR (I)

For a population-based study, AR is derived using sample estimates for  $P(D)$  and  $P(D|\bar{E})$ , or, sample estimates of  $P(E)$  and RR into the (equivalent) formulation:

$$\hat{P}(D) = \frac{a + c}{n}$$

$$\hat{P}(D|\bar{E}) = \frac{c}{c + d}$$

$$\widehat{AR} = \frac{\frac{a + c}{n} - \frac{c}{c + d}}{\frac{a + c}{n}}$$



$$\widehat{AR} = \frac{ad - bc}{(a + c)(c + d)}$$

$$\log(1 - \widehat{AR}) \pm z_{\alpha} \sqrt{\widehat{var}(\log(1 - \widehat{AR}))}$$

$$\widehat{var}(\log(1 - \widehat{AR})) = \frac{b + \widehat{AR}(a + d)}{nc}$$

95% CI on log scale : (−0.74, −0.39)

	CHD YES	CHD NO	tot
TYPE A	a=178	b=1141	1319
TYPE B	c=79	d=1486	1565
tot	257	2627	n=2884

43% of CHD is due to behavior type ... should be treated with **caution** as we have not yet considered the issue of the **causal** nature of the relationship between behavior type and CHD...

$$\widehat{AR} = 0.43$$
$$95\% \text{ CI: } (0.32, 0.52)$$

$$\widehat{var}(\log(1 - \widehat{AR})) = 0.008$$

# Sampling distribution of the AR (II)

For case-control studies an **equivalent** formulation of the Attributable Risk is given as:

Case-control study		Pancreatic Cancer		
		Cases	Controls	
Coffee drinking (cups per day)	>1	347	555	902
	0	20	88	108
		367	643	1010

$$AR = P(E|D) \left(1 - \frac{1}{RR}\right)$$

$$\widehat{AR} = \frac{(ad - bc)}{d(a + c)}$$

**with the rare disease assumption**, an estimate of the OR can be used to approximate RR.

$$\widehat{var} \left( \log(1 - \widehat{AR}) \right) = \frac{a}{c(a + c)} + \frac{b}{d(b + d)}$$

$$\widehat{var} \left( \log(1 - \widehat{AR}) \right) = 0.0571$$

$$\widehat{AR} = 0.60 \quad 95\% \text{ CI: } (0.36, 0.75)$$

Between 36% and 75% of pancreatic cancer cases **may be attributed** to coffee drinking. The implausibility of this large value of AR strongly hints that the **observed association** between coffee and pancreatic cancer may not be **causal !!**