cepn

CENTRE D'ÉTUDE SUR L'ÉVALUATION DE LA PROTECTION DANS LE DOMAINE NUCLÉAIRE

Young Generation in Radiation Protection Status, current possibilites and challenges Results of an international survey *

** and the opportunity to talk about statistics*

Sylvain Andresz

Junior Researcher at Nuclear Evaluation Protection Centre (CEPN), France

Member of the Leadership Group of the IRPA Young Generation Network

IRPA Congress Refresher Session

The Hague, 4th June 2018

IRPA YOUNG GENERATION NETWORK

Young Radiation Protection Generation Survey [2017 - 2018]

ENVOYER

QUESTIONS RÉPONSES 131

Young Generation in Radiation Protection Survey [2017 - 2018]

First: Welcome!

The Youth Club of the French Society for Radiation Protection (SFRP) and the Rising Generation Group of United Kingdom Society for Radiation Protection (SRP) - under the auspices of the IRPA Young Generation Network - have drafted a survey intended for the young generation in radiation protection.

> Who can participate?
The survey is for "young" radiation protection professionals and scientists (post doc. etc.). "Young" means that you are ≤ 35 years and/or have ≤ 10 years of work in radiation protection.

> Why do I need to participate?
The objective of this survey is to gather elements of information about the young radiation protection generation and also allow them to express their hopes and wishes. The survey is a place where you - as a young RP - can express your view and opinion about your work in radiation protection and foster the future of our generation.

> Expected completion time is < 20 minutes.
> The survey will be closed 31 December 2017 (midnight).
> The survey is of course anonymous. But you can leave your email (somewhere) if you wish being kept posted about the

Why a survey ?

- to gather elements of information about the young generation in radiation protection (YGRP) in the world
- because we would like to connect with the YGRP, to foster it and collect ideas for further development
- because we identified external interest and opportunity

How to do that?

- Survey drafted by Youth Club of French SFPR + Rising Generation Group of United Kingdom SRP
- "young RP generation" = "≤ 35 years and/or ≤ 10 years work in RP"
- Survey put on-line in July 2017 using GoogleForm ("universal")
- Distributed via IRPA network, other networks, personal direct contacts etc.

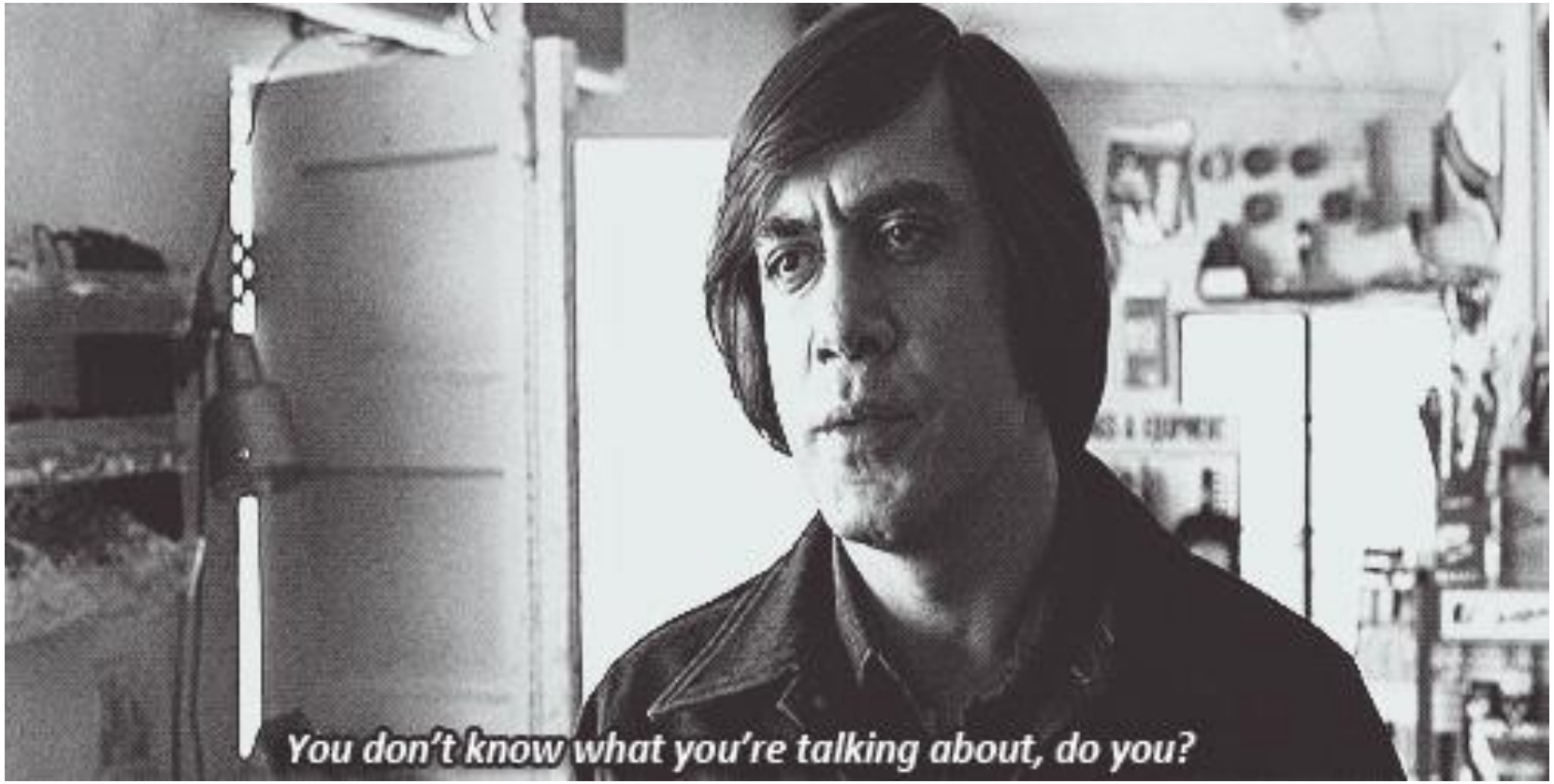
Outline

STEP 1
The opportunity to talk about statistics

STEP 2
The survey

STEP 3
Discussions

What are we talking about ?

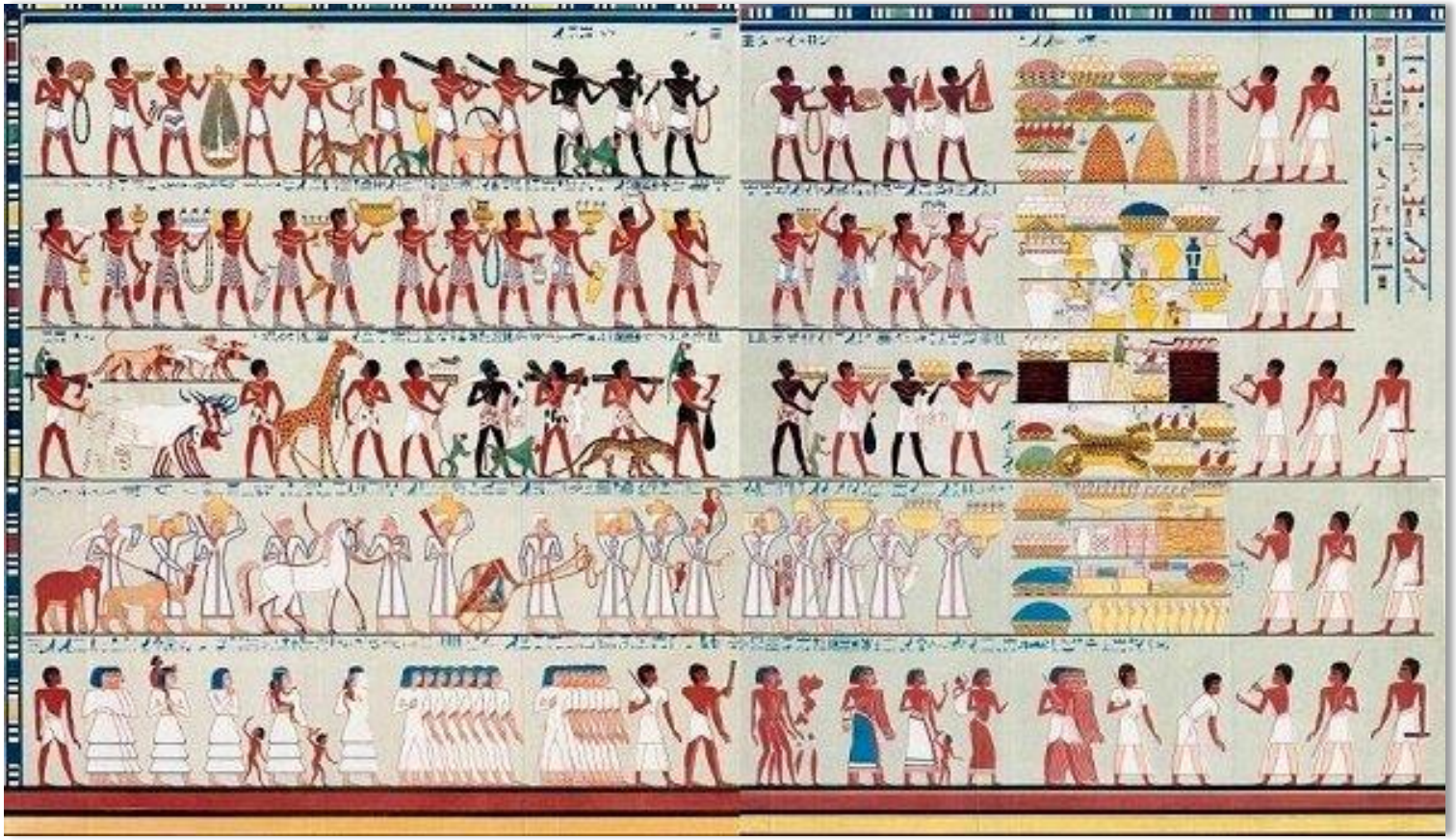


No Country for Old Men, Paramount, 2007

What are we talking about ?

- In English, “survey” can have different meaning :
- a census
- a poll,
  - either by random
  - or by quota
- ⇒ ***Do you now the difference(s) ?***

What are we talking about ?

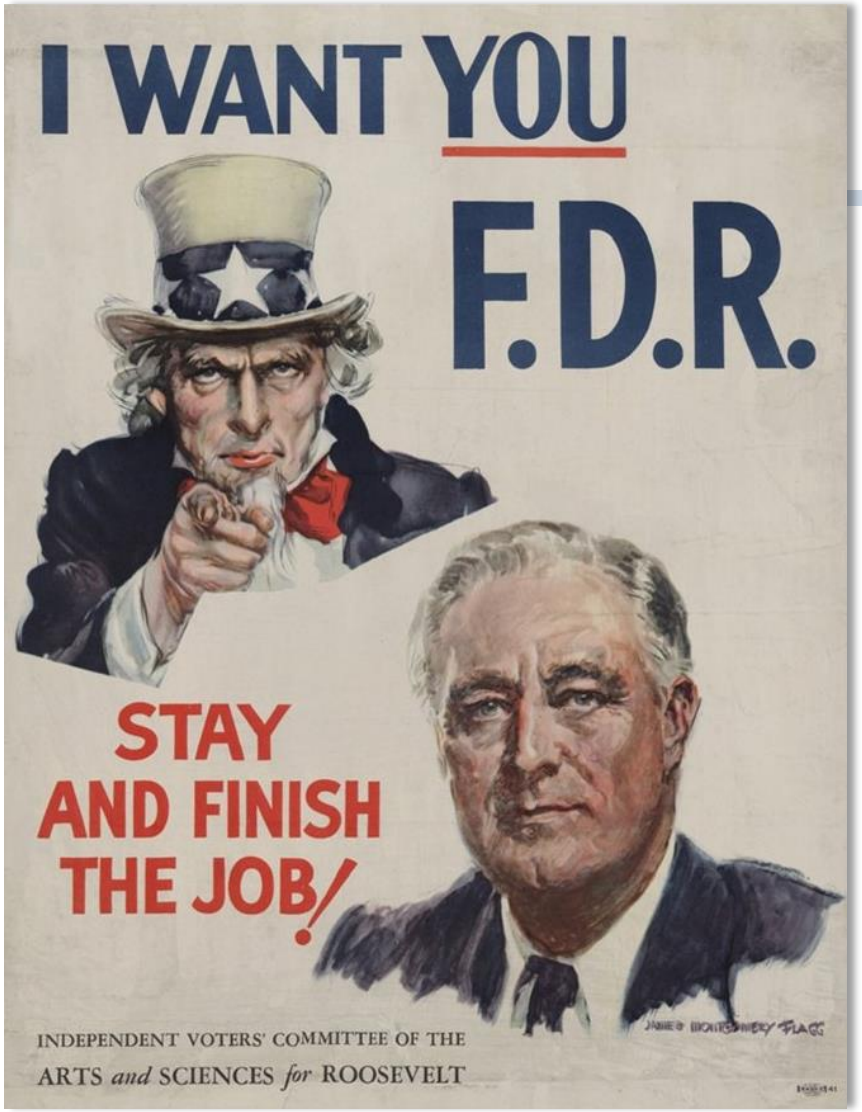


2nd dynasty (2925-2700 BC) – Ancient Egypt is the 1st known civilization where traces of **census** have been found (collection of observations of people and goods). Evolution of agricultural goods (China), goods (Roman Empire), Middle Age (counting the fire) etc.

(Tomb of Vizir Rekhmire, New Kingdom)

1783 – Pierre Simon de Laplace proposed a methodology to perform a **poll** over 1 million inhabitants, **randomly selected** in several regions and to calculate the expected uncertainty. Big debates oppose partisans of census and poll (XIX c.)

(*Portrait de Pierre Simon Marquis de Laplace*, 1838, Paulin Guérin)



3rd November 1936 – G. Gallup predicted the re-election of F. D. Roosevelt using a carefully selected **quota poll** (N = 3,000) when *Literary Digest* Magazine completely failed (N = 2,000,000)

The victory of the quota poll !

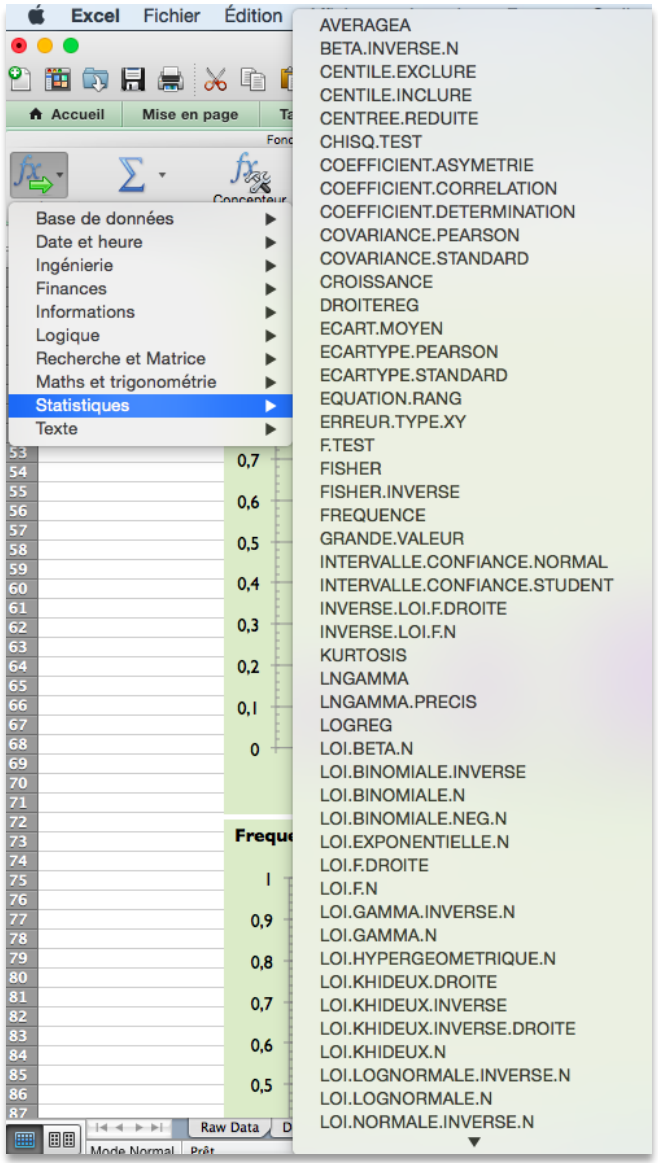
What are we talking about ?

- a **census**: "*a complete evaluation of the population and properties*"
  - *status* (lat.) = *statisticum* (mod. lat.) = statistic = which relates to the State
  - "political arithmetic" in English (until 1798)

- a **poll**: "*a questioning of persons selected at random or by quota to obtain information or opinion to be analysed*"
- The method of quota requires to constitute a representative sample of the population, for which the information we are looking are close to those of the whole population
  - *e.g.* the housewife shopping basket, election opinion poll etc.

- General case : you can perform a census (among all your colleagues) or a poll (among all the workers of your company)
- In our case : census and quota poll were not possible + distribution is driven "by random"

How to present data?



You now have collected your data

- **How to present them?**

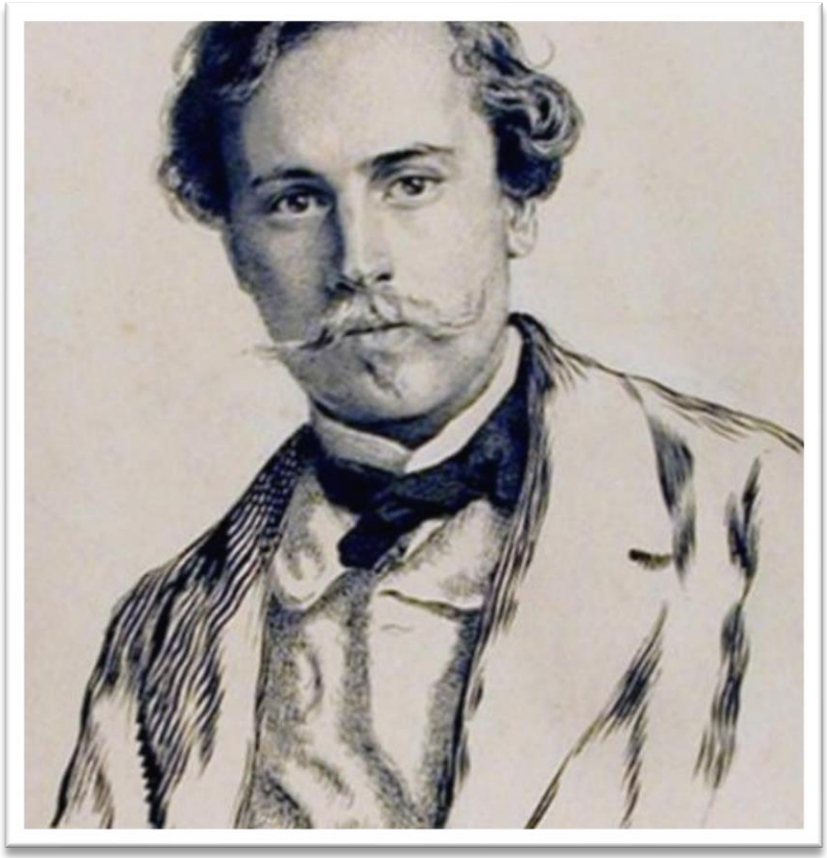
Most software provide a large array of statistical parameters – even for a very small sample !

- **What is the most common statistical parameter?**
- **What does it mean?**
- **Are there other parameters we can use?**

How to present data?

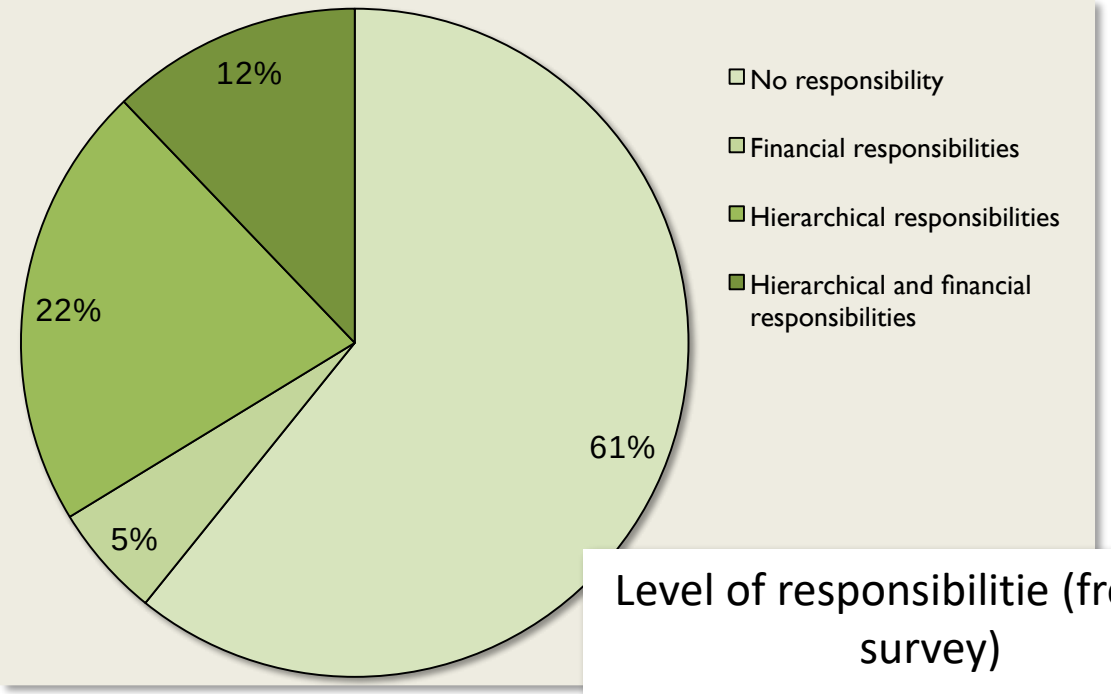
"Statistics are the first inexact sciences"

Jules de Goncourt (1830 – 1870)



How to present data?

- **Discrete data vs. continuous data**
- Discrete data takes specific values, generally non-numerical (yes/no, male/female, cheese/dessert, etc.)
- Pie chart (or cheese chart in France) are good way to present such results. The angle (hence size) is proportional to the % of individuals in the category
- Drawbacks?



Level of responsibilitie (from the survey)

How to present data?
The example of the collected salaries

0	15000	33000	48000
0	15000	33000	48000
0	16000	35000	48000
0	17000	35000	50000
0	17000	35000	50000
0	18000	35000	50000
1	18000	35000	50000
25	18000	35000	51000
35	18500	36000	53000
40	19000	36000	54800
40	19200	36400	55000
40	20000	36650	55000
58	20000	37000	59425
60	20000	37000	60000
65	21400	37000	60000
100	21600	37000	60000
100	22000	38000	60000
160	22000	38000	60803
380	23085	38000	63000
500	23400	38000	64000
800	24000	38000	70000
1000	24000	38000	70000
1000	25000	38000	71000
1000	25000	40000	73000
1200	25000	40000	75000
1500	25000	40000	76000
3000	26500	40000	77000
5000	27000	40000	80000
5500	27000	40000	80000
6000	28000	40000	80000
6000	28000	40500	80000
10000	28000	42000	81000
10000	30000	42000	82700
10000	30000	42400	85000
10000	30000	43000	85000
10000	30000	44000	85000
12000	30000	45000	85000
12000	30000	45000	87000
12000	30000	45000	90000
12500	30000	45000	90000
12793	30000	45000	92000
13600	30000	45000	100000
13750	30000	45000	100000
14000	30000	46000	105000
14400	30000	46000	105000
15000	32000	46000	110000
15000	32000	47200	110000
	32000	48000	125000
	32000	48000	130000
	33000		150000
	33000		160000

Option 0: Presenting just the raw data, n = 198
All information, but no interpretation!

How to present data?
Ranking and presenting data

- **Discrete data vs. continuous data**
- Continuous data are numerical and can take all possible value (eventually limited between upper and lower values)
- If sufficient precision and accuracy is allowed, there will be no two identical salaries!
- By gathering the values in **several classes**, it is possible to come back to the discrete case. In particular, histograms are often used to present the data

How to present data?
Ranking and presenting data

- The classes are chosen under arbitrary (but reasonable) rules: number of intervals, size and upper/lower values

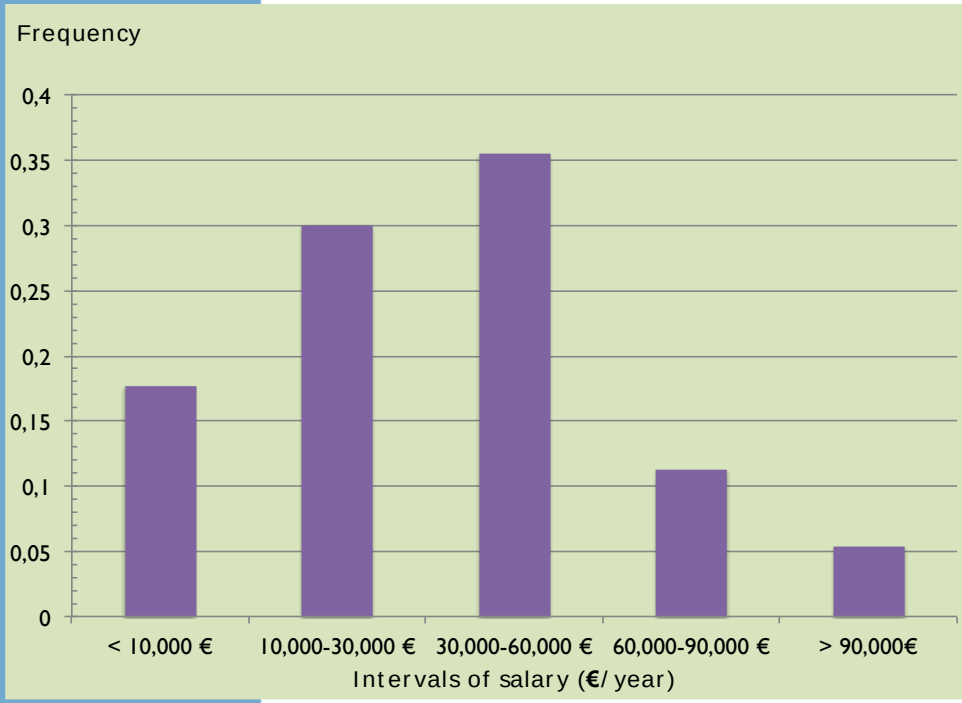
Classe	Number of values	Frequency
< 10,000 €	36	0,177
[10,000; 30,000[€	61	0,300
[30,000; 60,000[€	72	0,354
[60,000; 90,000[€	23	0,113
> 90,000€	11	0,054

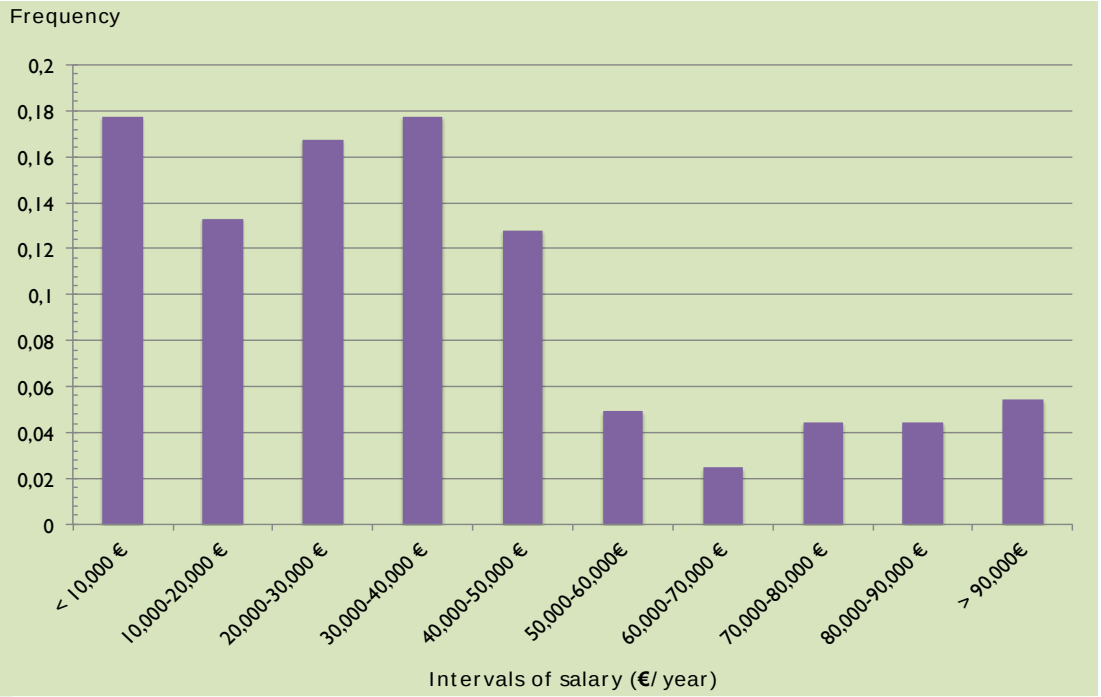
- Or ?

Classe	Number of values	Frequency
< 10,000 €	36	0,177
[10,000; 20,000[€	27	0,133
[20,000; 30,000[€	34	0,167
[30,000; 40,000[€	36	0,177
[40,000; 50,000[€	26	0,128
[50,000; 60,000[€	10	0,049
[60,000; 70,000[€	5	0,024
[70,000; 80,000[€	9	0,044
[80,000; 90,000[€	9	0,044
> 90,000€	11	0,054

How to present data?
Ranking and presenting data

- To be displayed in histograms:

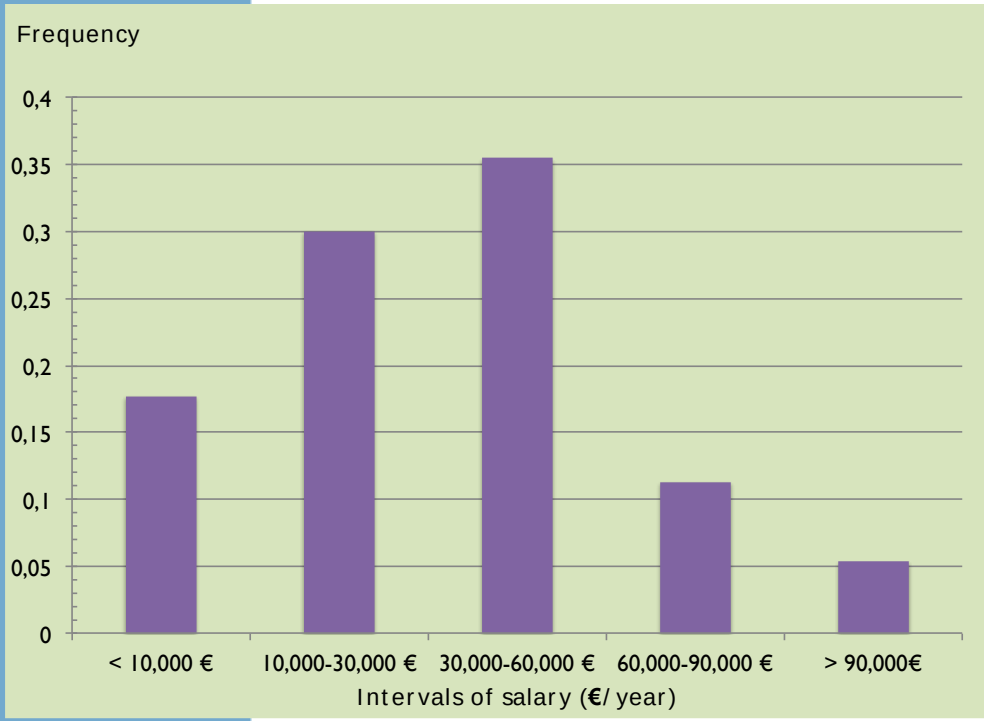


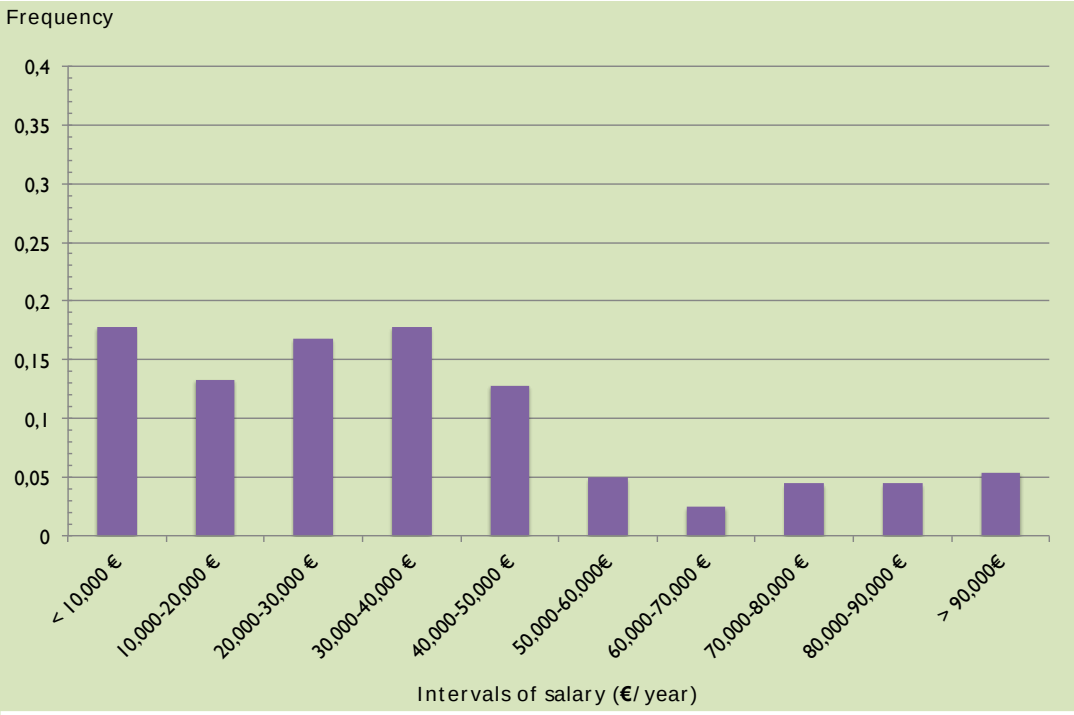


- **⇒ Compare : which one is 'the best'?**
- **⇒ What to conclude with the distribution?**

How to present data?
Ranking and presenting data

- To compare 2 histograms, it is preferable to use the same scale!



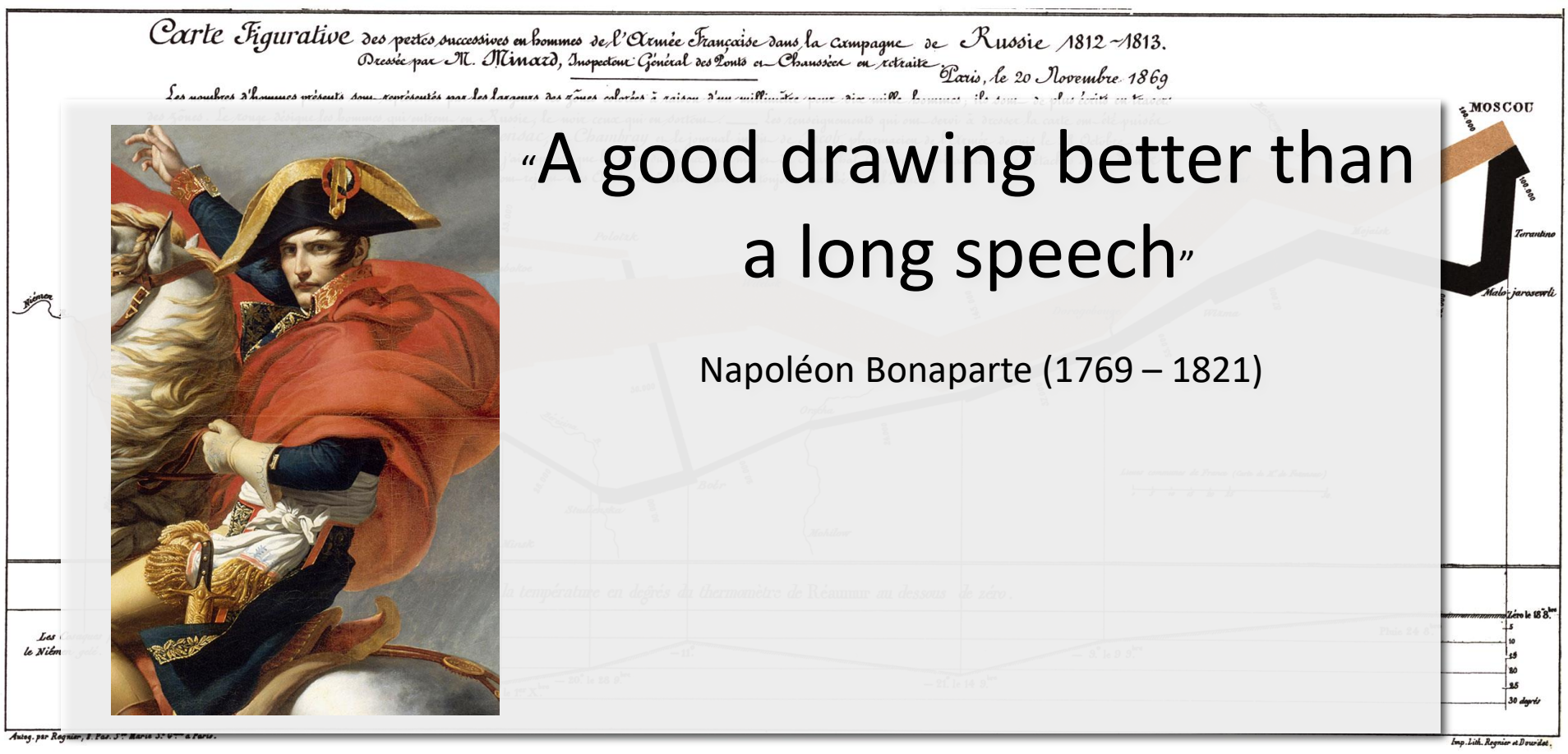


- Suggestion 1 : *do not let Excel choose for you*
- Suggestion 2 : consider the **precision** you would like to provide and the **number of individuals** in each class

How to present data?
Ranking and presenting data

- Histograms relies on arbitrary choices for presenting the data
- Size of the intervals (especially if they are different) make frequencies not *intrinsic* (i.e. belonging to a things by its very nature)
- A possibility is to introduce the *density of frequency* :
    - d = frequency/size of the class
- But this is not the most commonly know value + you still loose information (what about the dispersion *inside* a class ?)
- Providing graphics or not? Being exhaustive or demonstrative?

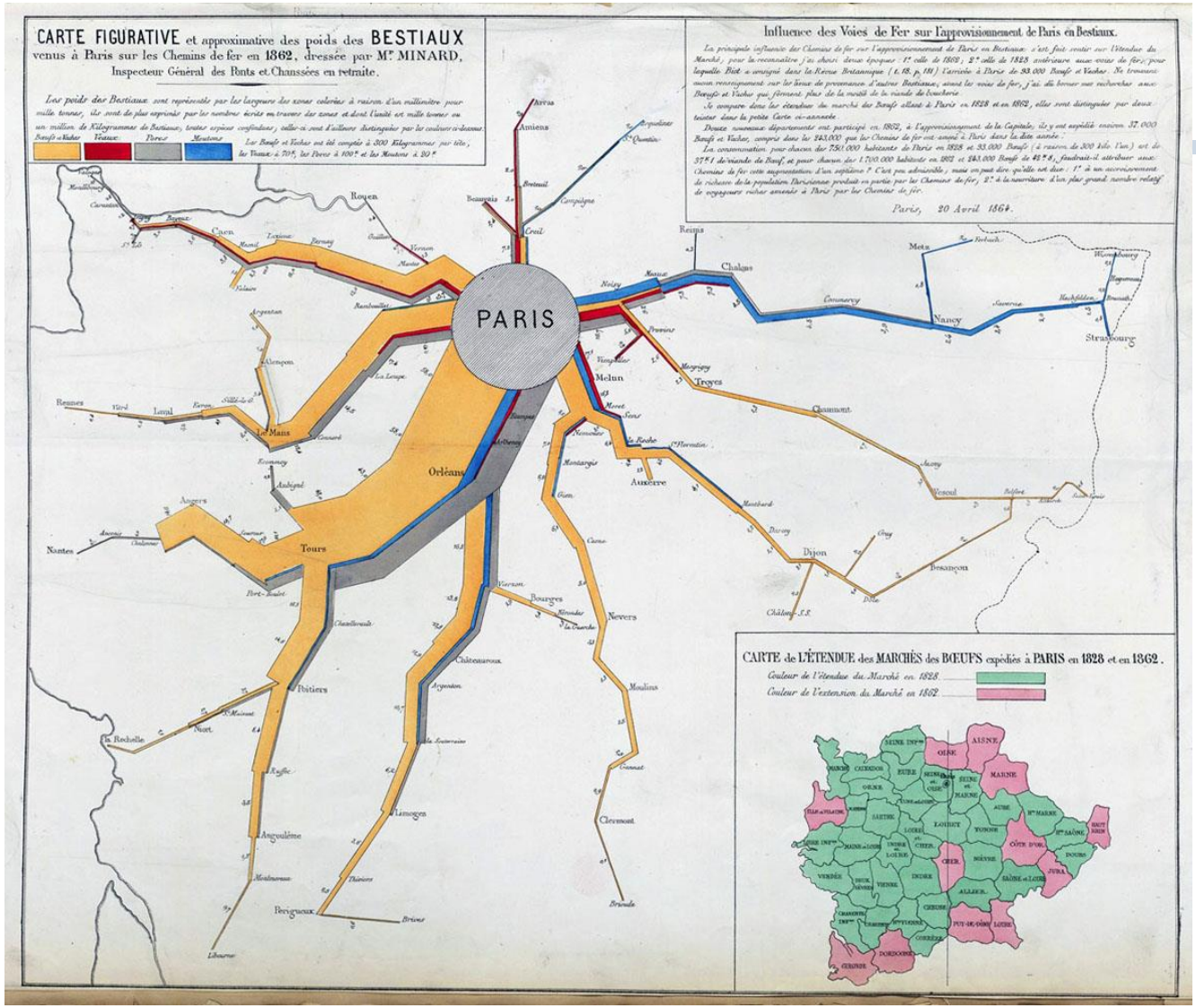
How to present data?
Thinking innovative



"A good drawing better than a long speech"

Napoléon Bonaparte (1769 – 1821)

Flow Chart of the Napoleon Army during the Russian Campaign (C. Minard, 1821)
Way (in orange) and return (in black) ; 1 mm = 6 000 soldiers



Map of the weight (tons) of cattle brought to Paris by train from the regions (C. Minard, 1862)
Beef (yellow), Veal (red), porc (grey) and mutton (blue)

Parameters of centrality and dispersion

- **Can we now summarize all the data x_i with one *parameter*?**
- If we call this unique "a", we can introduce the errors : $e_i = x_i - a$
- Objective is to aggregate all the errors and then minimize
- For GAUSS, the optimum aggregate is the *sum of all squared errors* :

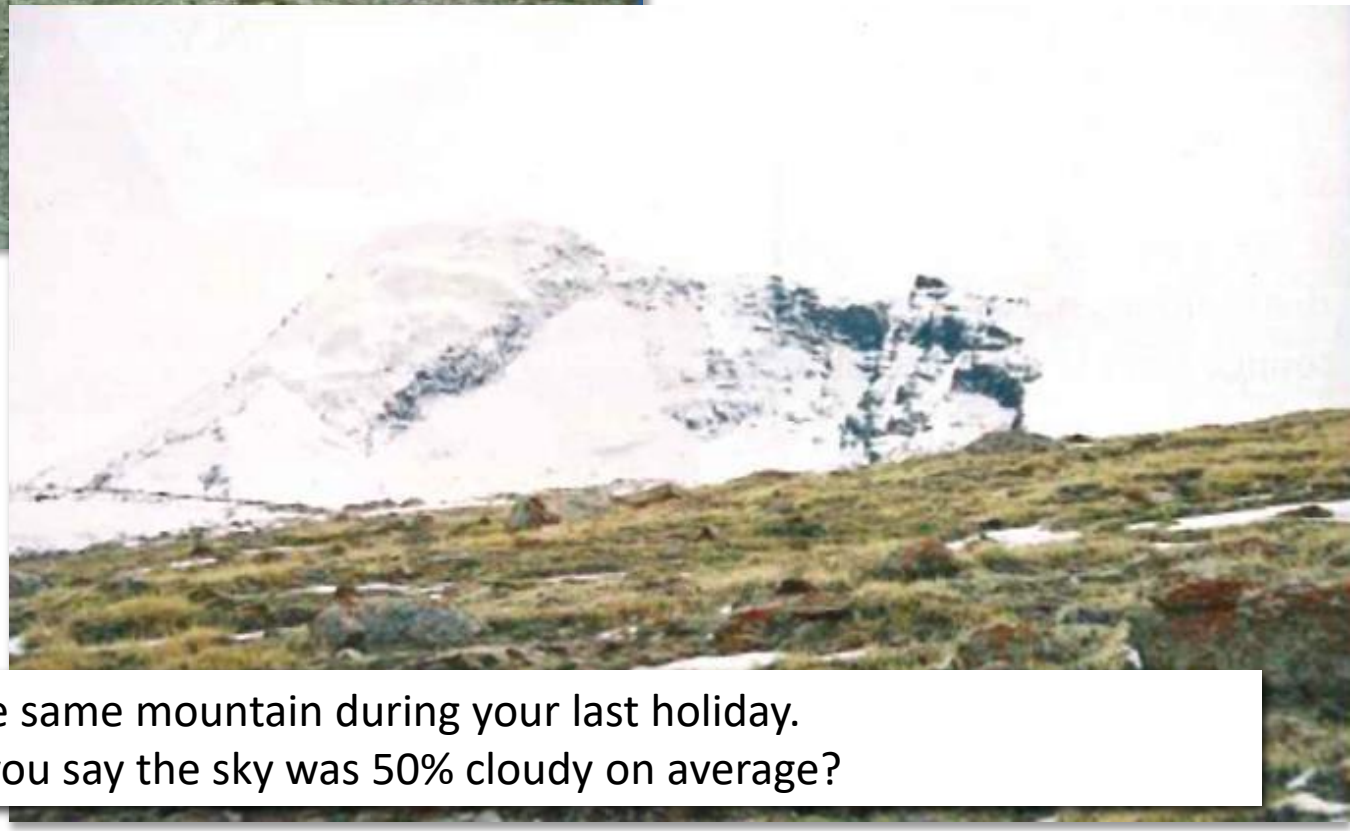
$$ET(a) = \sum_i (x_i - a)^2$$

- This parabola admit one minimum when

$$ET'(a) = \partial(\sum_i x_i^2 - 2ax_i + a^2)/\partial x_i = 0$$

$$\sum_i 2x_i - 2na = 0 \text{ i.e. when } a = 1/n \sum_i x_i$$

- The definition of the arithmetic mean! **AVERAGE(...)**
- In our case m = 42,168 € - but,
  - ⇒ *Provide a very limited view of the data*
  - ⇒ *No indication for evolution or trend*
  - ⇒ *High sensitivity to extreme and odd values:* $(x_i - a)^2$
  - ⇒ *The mean does not (necessary) represents a reality !*



The same mountain during your last holiday.
Can you say the sky was 50% cloudy on average?

Parameters of centrality and dispersion

- The minimum error is then :

$$ET_{min} = \sum_i (x_i - m)^2$$

- You can distribute this minimum error over the number of observations to average it:

$s^2 = 1/n \sum_i (x_i - m)^2$ - which is the variance **VAR.P(...)**

- and then you go back to the initial unit:

$\sigma = \sqrt{s^2}$ - which is the (most famous) standard deviation **STDEV(...)**

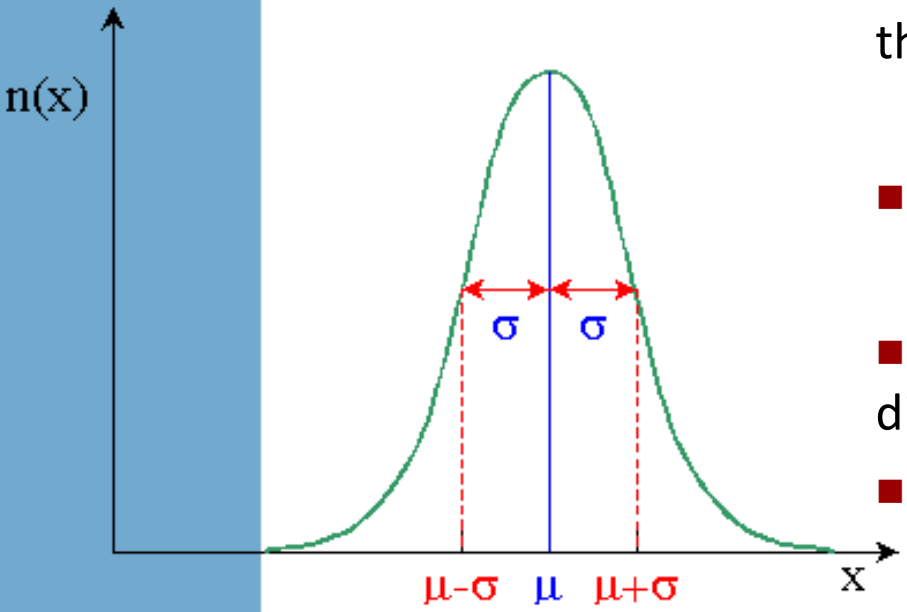
- In our case σ = 30,149 €
- ⇒ What does it mean?

Parameters of centrality and dispersion

- Standard deviation provides a measurement of the **dispersion of the values**
  - *⇒ interpretation and representation remain difficult ("square root of the average sum of the errors")*
  - *⇒ intrinsically linked to the population*
  - *⇒ intrinsically linked to the unit and absolute value*
  - *⇒ A high standard deviation does not mean anything without context:*
    - A: 2 18 2 18 ; $m_A = 10$, $\sigma_A = 9{,}2$
    - B: 20 180 20 180 ; $m_B = 100$, $\sigma_B = 92{,}4$
    - but the coefficient of variation is the same : $\sigma_A / m_A = \sigma_B / m_B$
  - *⇒ and the data are not necessarily evenly spread around the mean!*

Parameters of centrality and dispersion

- GAUSS (again) proved that every complex phenomenon resulting from a large number of independent causes can be described by a 'normal' distribution (or a Gaussian, or Laplace-Gauss, or etc. distribution)
- Only in this case, standard deviation takes real significance because you can calculate *a priori* all the intervals of frequency only with m and σ



- For example, it is possible to calculate that 68% of the observations are contained in:

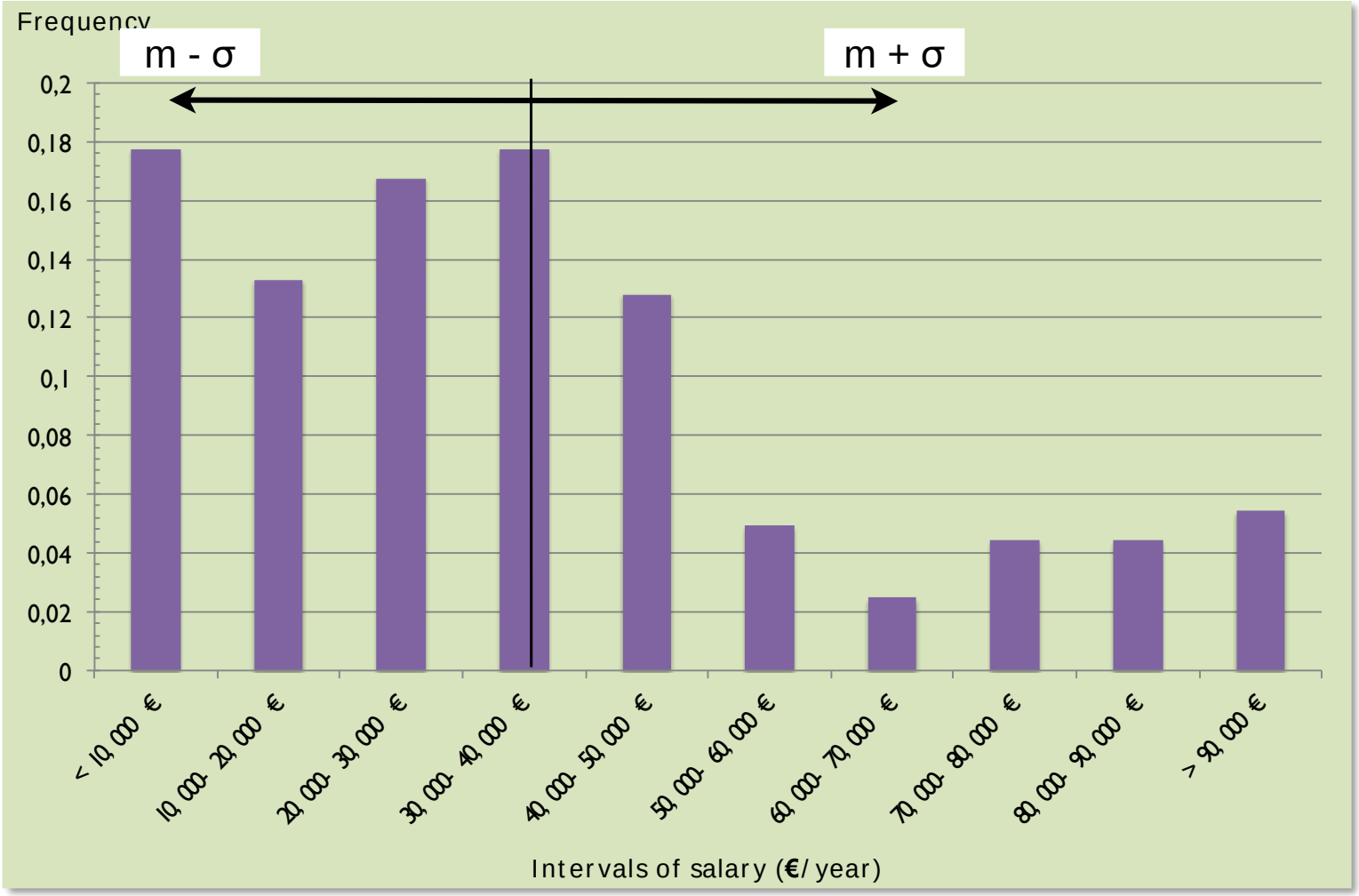
$$I_{68\%} = [\mathbf{m} - \boldsymbol{\sigma} \,;\, \mathbf{m} + \boldsymbol{\sigma}]$$

- And 95% are contained in:

$$I_{95\%} = [\mathbf{m} - 1.96\boldsymbol{\sigma} \,;\, \mathbf{m} + 1.96\boldsymbol{\sigma}]$$

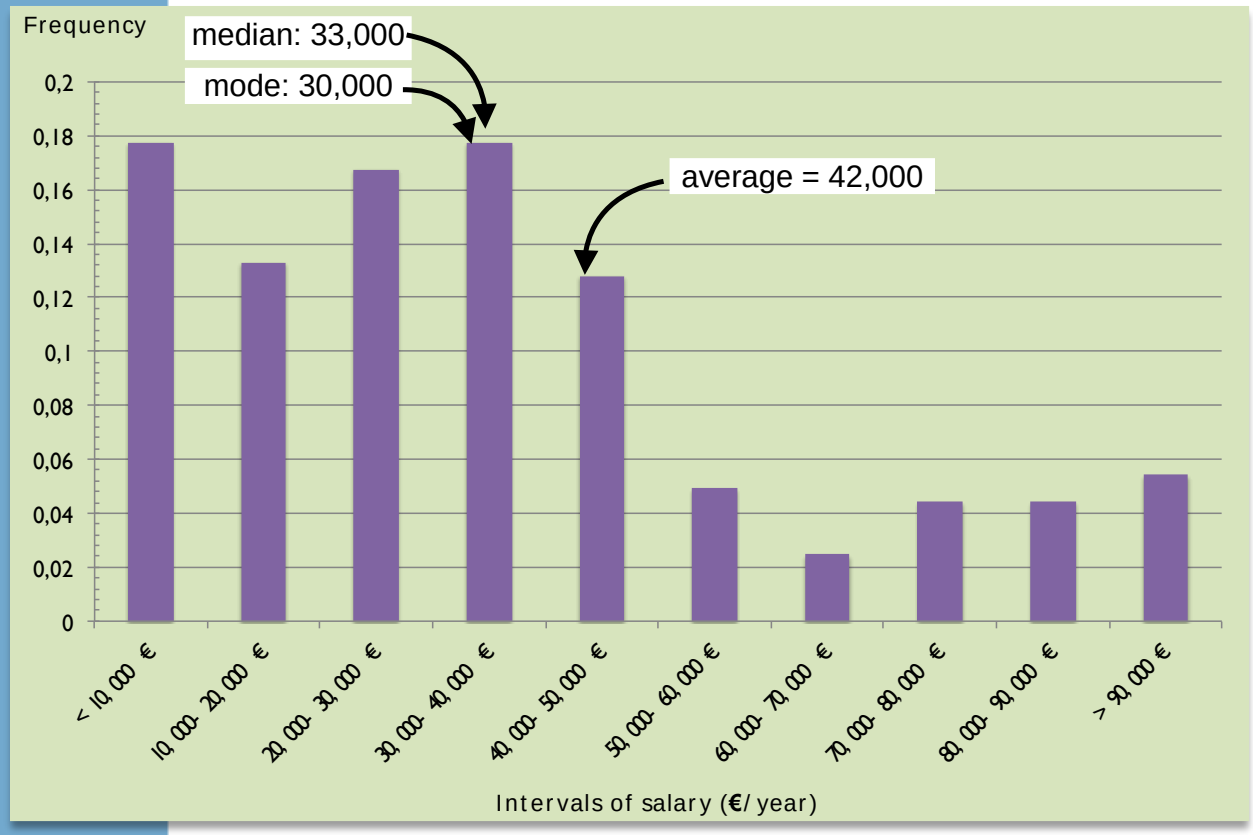
- ⇒ in our case, are we following a normal distribution?
- This is generally admitted (and rarely demonstrate)

Parameters of centrality and dispersion



Parameters of centrality and dispersion

- **Other parameters are possible !**
- the **mode**: the value that appears the most often in a set of data (the value *à la mode*)



- the **median**: the value separating the higher half of the population from the lower population
- generally; the median is the value with the rank number (n + 1)/2, when all values are ranked

- the **spread**: max – min

Parameters of centrality and dispersion

- **σ can be used for discrete data also**
- If you observe a result P from a sample of size n from your population, you can prove their is a confidence at 95% that the "real" value is in:

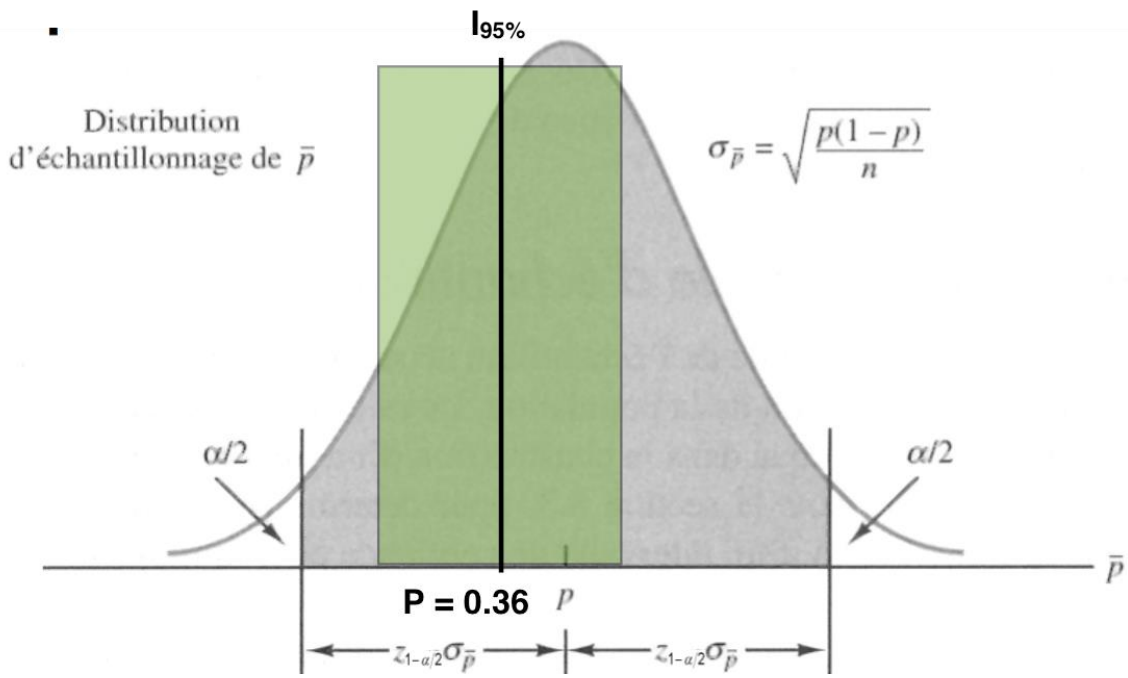
$$[\mathbf{P} - 1.96\boldsymbol{\sigma}/\sqrt{n}\ ;\ \mathbf{P} + 1.96\boldsymbol{\sigma}/\sqrt{n}]$$

$$\text{hence } [\mathbf{P} - 1.96\sqrt{(P(1-P)/n)}\ ;\ \mathbf{P} + 1.96\sqrt{(P(1-P)/n)}]$$

  - *⇒ Contrary to the intuition, the size of the population does not influence the size of the interval (I is the same whatever your sample come from: population of The Hague, population of The Netherlands etc.)*
  - *⇒ But interval size is inversely proportional to $\sqrt{n}$ ⇒ if you want to diminish it by 2, you need 4 × n*
  - *⇒ And if you want to calculate P for a sub-sample : n ↘, so I ↗*
  - *⇒ At a certain point (e.g. n < 30) it is even not licit to calculate a confidence interval*

Parameters of centrality and dispersion

- Ex. 36% of the respondents of the survey are female
- So "*36% of the young radiation protection workers are female ; the confidence interval is [0.30 ; 0.42]*"
- ⇒ But are we following a normal distribution?



Rounding

- There is no definitive recipe for rounding:
  - Too much and you loose information
  - Too little and this suggest high precision (you cannot guarantee) and diminish visibility

- <u>Suggestion for σ</u>:
- keep 2 figures, the last one beeing rounded (classical rounding)

σ	rounded
35.2438	35
2379	2400
0.6694	0.67
0.0023256	0.0023
0.00040401	0.00040
0.89712	0.90

Rounding

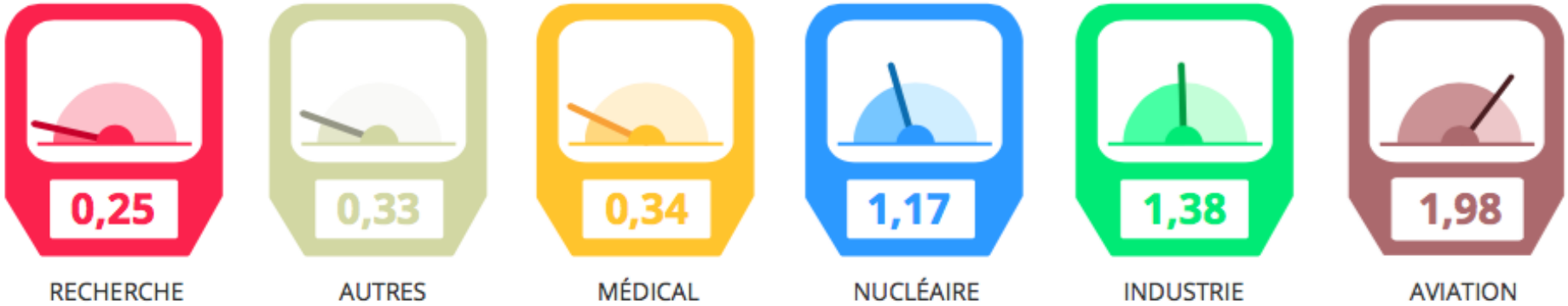
- Suggestion for a result affected by σ:
- keep the same numbers of decimal as for σ:

result	σ	rounded
1864.387	6.8	1864.4
0.68088	0.013	0.681
24.52698	0.25	24.53
2624.57	120	2620
54867.12	2000	54 900

ANALYSE PAR DOMAINE D'ACTIVITÉ

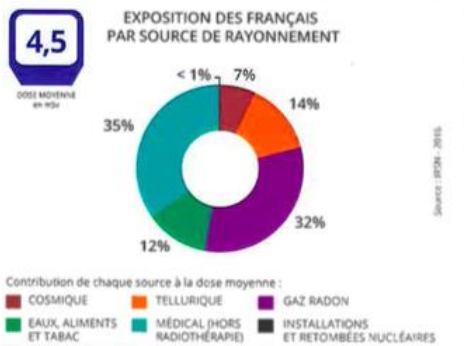
Comme les années précédentes, il existe des disparités fortes entre les différents domaines d'activité. Les personnels navigants de l'aviation, exposés au rayonnement cosmique, présentent les doses individuelles moyennes les plus élevées, suivis par les travailleurs de l'industrie non nucléaire et nucléaire. Dans le domaine médical et vétérinaire et dans celui de la recherche, la dose individuelle moyenne est plus faible.

DOSE INDIVIDUELLE MOYENNE PAR DOMAINE D'ACTIVITÉ (mSv)



Extract from the *Annual Results of Professionnal Exposure to Ionizing Radition in France*, 2015, published by the National Radiation Protection Institute (IRSN)

EXPOSITION DE LA **POPULATION**

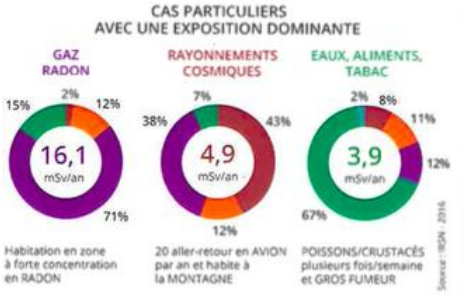


Une exposition majoriraiement d'origine naturelle

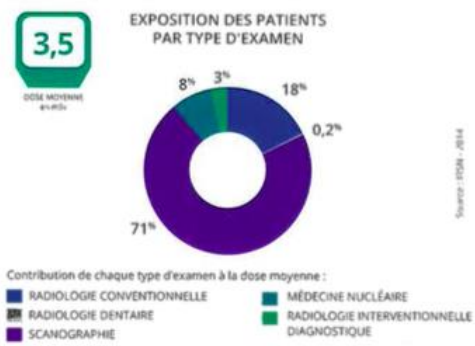
L'exposition moyenne à l'ensemble des sources de rayonnements ionisants est de 4,5 mSv/an, dont 2,9 mSv/an pour celles d'origine naturelle et 1,6 mSv/an pour celles d'origine artificielle (essentiellement médicale).

Une exposition propre à chaque individu

L'exposition peut varier de façon significative, y compris d'une année à l'autre, selon le lieu d'habitation (zone radon, vie en altitude), le nombre et la nature des examens médicaux réalisés ou les habitudes de vie (voyages fréquents en avion, tabac) et alimentaires (fréquence de consommation de poissons et de crustacés).



EXPOSITION DES **PATIENTS**

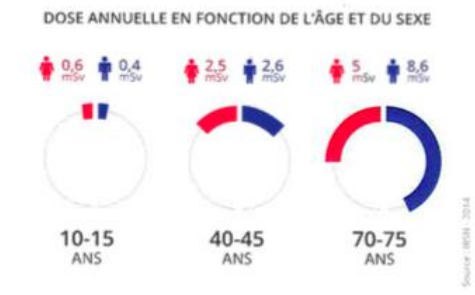


Une exposition non négligeable mais utile

Près de 45 % des Français bénéficient d'au-moins un examen médical à visée diagnostique par an. Chez ces patients, la dose individuelle moyenne s'élève à 3,5 mSv/an, au lieu des 1,6 mSv/an si l'on considère l'ensemble de la population.

Une exposition propre à chaque individu

70 % des personnes bénéficiant d'un examen à visée diagnostique reçoivent une dose inférieure à 1 mSv/an, 20 % entre 1 mSv/an et 10 mSv/an, et 10 % une dose de plus de 10 mSv/an. Par ailleurs, la répartition de ces examens, donc de la dose associée, varie en fonction de l'âge et du sexe.



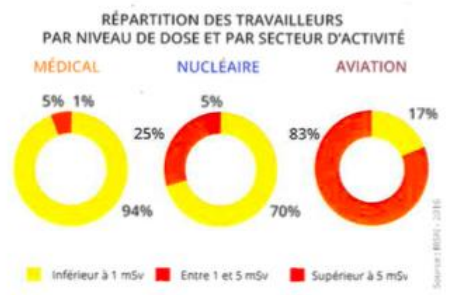
EXPOSITION DES **TRAVAILLEURS**

Une exposition limitée et maîtrisée

En France, environ 386 000 travailleurs sont surveillés pour leur exposition potentielle aux rayonnements ionisants. Parmi eux, seulement 1/4 sont des travailleurs effectivement exposés et leur dose individuelle moyenne est de 0, 95 mSv/an.

Une exposition propre à chaque individu

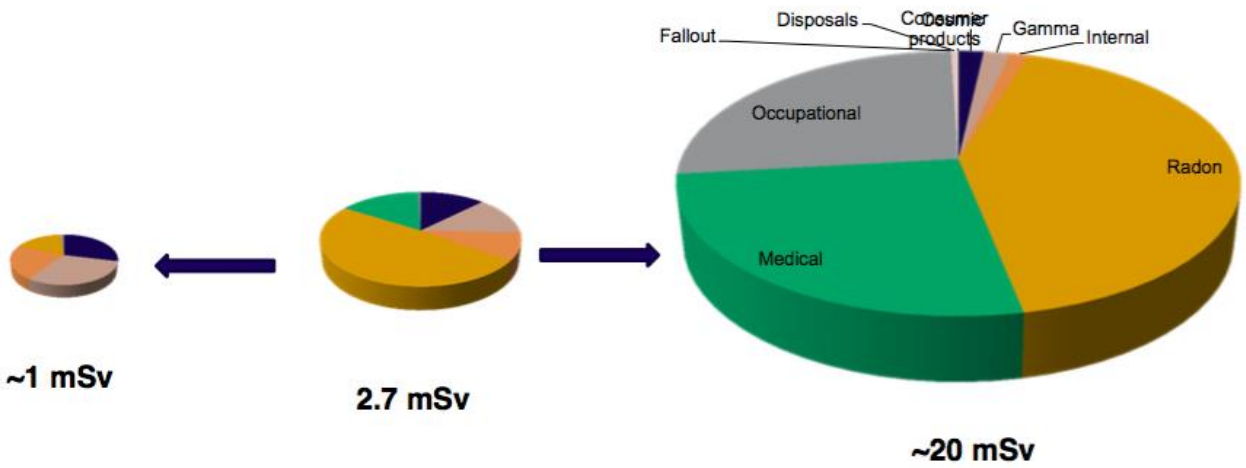
71 % des travailleurs effectivement exposés reçoivent une dose inférieure à 1 mSv/an, 26,5 % entre 1 mSv/an et 5 mSv/an, et 2,5 %, soit 2 600 travailleurs sont exposés au delà de 5 mSv/an. La répartition par classe de dose dépend fortement du domaine d'activité.



Extract from the *Annual Results of Professionnal Exposure to Ionizing Radition in France*, 2016, published by the National Radiation Protection Institute (IRSN)

Public Health England

Uncertainty in average doses can be large



From *Presentation of doses and risks to members of the public: a personal perspective*, T. Cabianca, CRCE, Public Helath England, NEA/CRPPH annual meeting, 11 April 2018.

Action/Task

dose rate × workload × K

	Description	Entité	n° zone	Itération	DED	Nb Pers	tps expo	VTE	Coeff K	Dose	%
préparation	Préparation de chantier (mise en place des protections,…)	CMI	1	1	0,004	4	12	48	0,5	0,096	1,80
	Introduction du matériel modif BU par le TAM	CMI	1	1	0,004	4	6	24	0,5	0,048	0,90
	Suivi de chantier	CMI	1	1	0,004	1	18	18	0,5	0,036	0,68
Travaux sur PTR002 ou 003BU	Déplacement du batardeau PTR001BU sur son support	CMI	1	1	0,007	4	1,15	4,6	0,6	0,019	0,36
	Mise en place du confinement autour du batardeau	PNS	1	1	0,004	3	1,15	3,45	0,6	0,008	0,16
	Soudage des oreilles de levage	CMI	3	1	0,021	2	3,75	7,5	0,9	0,142	2,66
	Ressuage des oreilles	CEPI	3	1	0,021	1	0,75	0,75	0,9	0,014	0,27
	Décontamination fine batardeau + nettoyage + zone de travail	CMI	3	1	0,021	2	0,75	1,5	0,9	0,028	0,53
	Déconfinement sas	PNS	3	1	0,021	4	1,5	6	0,6	0,076	1,42
	Déplacement du ba	CMI	1	1	0,007	4	1,15	4,6	0,6	0,019	0,36
	Mise en place du s	CMI	1	1	0,004	4	4,5	18	0,6	0,043	0,81
	Déplacement du ba modifié	CMI	1	1	0,007	4	1,15	4,6	0,6	0,019	0,36
	Mise en place du c	PNS	3	1	0,021	4	18	72	0,6	0,907	17,02
	Remplacement des 2 joints (statiques et dynamiques) + ajout d'un joint	CMI	3	1	0,021	5	14,5	72,5	0,9	1,370	25,71
	Remplacement des 2 joints (statiques et dynamiques) + ajout d'un joint	CMI	3	1	0,007	5	14,5	72,5	0,9	0,457	8,57
	Remplacement des 2 joints (statiques et dynamiques) + ajout d'un joint	CMI	3	1	0,006	5	14,5	72,5	0,9	0,411	7,71
	Remplacement des 2 joints (statiques et dynamiques) + ajout d'un joint	CMI	3	1	0,004	5	14,5	72,5	0,9	0,228	4,28
	Assistance habillage déshabillage	PNS	3	1	0,004	3	19	57	0,6	0,137	2,57
	Ressuage	CEPI	3	1	0,021	2	3,75	7,5	0,9	0,142	2,66
	Dépose confinement	PNS	3	1	0,021	4	12	48	0,6	0,605	11,35
	Mise en place du batardeau dans la piscine	CMI	3	1	0,007	4	1,15	4,6	0,6	0,019	0,36
	Démontage du support modifié	CMI	1	1	0,004	4	4,5	18	0,6	0,043	0,81
	Montage de la panoplie d'air comprimé en extrêmité du joint gonflable	CMI	3	1	0,004	1	0,75	0,75	0,6	0,002	0,03
	Mise en place du batardeau dans la trémie	CMI	3	1	0,007	4	1,15	4,6	0,6	0,019	0,36
	Suivi de chantier	CMI	1	1	0,004	2	65	130	0,5	0,260	4,88
epli	Evacuation du matériel de chantier	CMI	1	1	0,004	4	6	24	0,5	0,048	0,90
	Evacuation des déchets	CMI	1	1	0,004	4	3	12	0,5	0,024	0,45
	Retrait des protections de sols	CMI	1	1	0,004	4	3	12	0,5	0,024	0,45
					0,009			863,5		5,330	100,00

Ambient ***average*** equivalent dose rate in the room (1 mesure/month)

Collective dose

Extract from a radiation protection analysis for a maintenance task

Synthesis
Key messages

- **A point of vocabulary**
  - *Statisticus* (lat.) = State
  - Survey vs. census vs. poll

- The clear presentation and the search for the most effective use of a database is a challenge
- How to provide the maximum information, and being simple, short and correct?
- Always a balance! cf. Information Theory (1948); *"source code basically tries to reduce the redundancy in the source and represent it with fewer bit (hence reduce entropy)"*

- **Graphical presentation**
  - Discrete vs. Continuous data
  - Focus on pie charts and histograms (relies on arbitrary choices)

Synthesis
Key messages

- **Statistical parameters**
- Definition of the **average**;
  - It minimize the sum of all the squared errors!
  - The most common tool, but can be deceptive and is generally not sufficient
- **Standard deviation** σcan be calculated
  - Application to continuous data (dispersion of observation)
  - Application to discrete data (confidence interval)
  - Assumption of a normal distribution
- Consider providing **other parameters and elements of variability** σ/m ; or simply the ***mode***, the ***median***, ***max. & min. expected value & extreme value*** etc.
- **This is part of an ALARA analysis** !
- Consider **rounding**

Annexe 1

- Difference **VAR.P(...)** and **VAR.S(...)** in Excel ?
- VAR.P is used if you have all the data of the population (this is rare) ; if you only have a sample, the calculation of

$$1/n \sum_i (x_i - \dot{x})^2$$

- can only be made on n – 1 value because $\dot{x}$ is calculated from the x_i

$$s'^2 = \text{VAR.S} = 1/(n-1) \sum_i (x_i - \dot{x})^2$$

- (the Bessel correction factor of n/n – 1)
- And $\sigma' = \sqrt{s^2}$ **STDEV.S(...)** to be used in general
- Does not change much if n >> 0

Annexe 2

- <u>The standard deviation for proportions</u>
- Proportions are averaged value:
- Consider two possibilities A and B then n_A, n_B, $n = n_A + n_B$ and $P = n_B/n$ and x the variable = 0 (for A) and 1 (for B).
- The average of x

$$m = 1/n \sum_i x_i = 1/n\,(0 + \ldots + 0 + 1 + \ldots 1) = n_B/n = P$$

- Then

$$\sigma^2 = 1/n \sum_i (x_i - P)^2$$
$$= 1/n\,[(0 - P)^2 + \ldots + (0 - P)^2 + (1 - P)^2 + \ldots\ (1 - P)^2]$$
$$= 1/n\,[n_A P^2 + n_B (1 - P)^2]$$
$$= n_A/n\, P^2 + n_B/n\,(1 - P)^2$$
$$= (1 - P)\, P^2 + P(1 - P)^2$$
$$= P(1 - P)(P + 1 - P)$$

- Finnally $\sigma = \sqrt{(P(1 - P)/n)}$