UNIVERSITÉ DU LUXEMBOURG

LUXEMBOURG INSTITUTE OF HEALTH
RESEARCH DEDICATED TO LIFE
Bioinformatics group, Proteome and Genome Research Unit

BIOSTATISTICS

Lecture 4

Sampling and Sampling Distribution

dr. Petr Nazarov

petr.nazarov@lih.lu

6-03-2020

OUTLINE

- **Sampling distribution**
  - population and sample
  - random sample
  - sampling distribution
  - properties of point estimators
  - corrections for finite size of population
  - central limit theorem
  - other sampling methods

POPULATION AND SAMPLE

Parameters

Population parameter
A numerical value used as a summary measure for a population (e.g., the mean μ, variance σ^2, standard deviation σ, *proportion* π)

POPULATION

μ – mean
σ^2 – variance
N – number of elements (usually $N=\infty$)

SAMPLE

$m, \bar{x}$ – mean
s^2 – variance
n – number of elements

Sample statistic
A numerical value used as a summary measure for a sample (e.g., the sample mean m, the sample variance s^2, and the sample standard deviation s)

All existing laboratory
Mus musculus

mice.txt

790 mice from different strains
http://phenome.jax.org

ID	Strain	Sex	Starting age	Ending age	Starting weight	Ending weight	Weight change	Bleeding time	Ionized Ca in blood	Blood pH	Bone mineral density	Lean tissues weight	Fat weight
1	129S1/SvImJ	f	66	116	19.3	20.5	1.062	64	1.2	7.24	0.0605	14.5	4.4
2	129S1/SvImJ	f	66	116	19.1	20.8	1.089	78	1.15	7.27	0.0553	13.9	4.4
3	129S1/SvImJ	f	66	108	17.9	19.8	1.106	90	1.16	7.26	0.0546	13.8	2.9
368	129S1/SvImJ	f	72	114	18.3	21	1.148	65	1.26	7.22	0.0599	15.4	4.2
369	129S1/SvImJ	f	72	115	20.2	21.9	1.084	55	1.23	7.3	0.0623	15.6	4.3
370	129S1/SvImJ	f	72	116	18.8	22.1	1.176		1.21	7.28	0.0626	16.4	4.3
371	129S1/SvImJ	f	72	119	19.4	21.3	1.098	49	1.24	7.24	0.0632	16.6	5.4
372	129S1/SvImJ	f	72	122	18.3	20.1	1.098	73	1.17	7.19	0.0592	16	4.1
4	129S1/SvImJ	f	66	109	17.2	18.9	1.099	41	1.25	7.29	0.0513	14	3.2
5	129S1/SvImJ	f	66	112	19.7	21.3	1.081	129	1.14	7.22	0.0501	16.3	5.2
10	129S1/SvImJ	m	66	112	24.3	24.7	1.016	119	1.13	7.24	0.0533	17.6	6.8
364	129S1/SvImJ	m	72	114	25.3	27.2	1.075	64	1.25	7.27	0.0596	19.3	5.8
365	129S1/SvImJ	m	72	115	21.4	23.9	1.117	48	1.25	7.28	0.0563	17.4	5.7
366	129S1/SvImJ	m	72	118	24.5	26.3	1.073	59	1.25	7.26	0.0609	17.8	7.1
367	129S1/SvImJ	m	72	122	24	26	1.083	69	1.29	7.26	0.0584	19.2	4.6
6	129S1/SvImJ	m	66	116	21.6	23.3	1.079	78	1.15	7.27	0.0497	17.2	5.7
7	129S1/SvImJ	m	66	107	22.7	26.5	1.167	90	1.18	7.28	0.0493	18.7	7
8	129S1/SvImJ	m	66	108	25.4	27.4	1.079	35	1.24	7.26	0.0538	18.9	7.1
9	129S1/SvImJ	m	66	109	24.4	27.5	1.127	43	1.29	7.29	0.0539	19.5	7.1

POPULATION AND SAMPLE

Random Sampling

Simple random sampling

Finite population: a sample selected such that each possible sample of size *n* has the same probability of being selected.

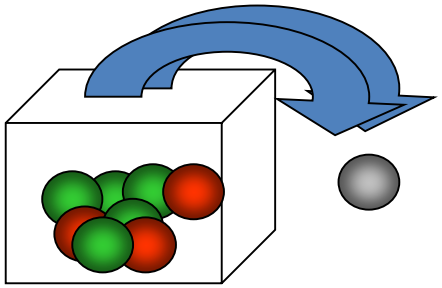
Infinite population: a sample selected such that each element comes from the same population and the elements are selected independently.

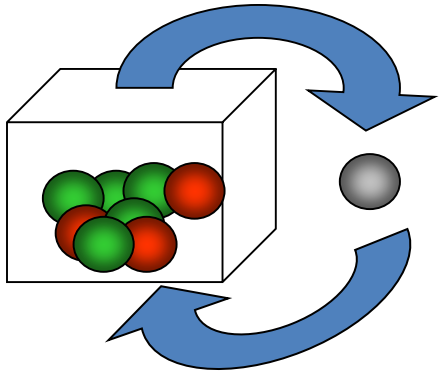
Sampling without replacement

Once an element has been included in the sample, it is removed from the population and cannot be selected a second time.

Sampling with replacement

Once an element has been included in the sample, it is returned to the population. A previously selected element can be selected again and therefore, may appear in the sample more than once.





POPULATION AND SAMPLE

Example: Making a Random Sampling

mice.xls

790 mice from different strains

http://phenome.jax.org

ID	Strain	Sex	Starting age	Ending age	Starting weight	Ending weight	Weight change	Bleeding time	Ionized Ca in blood	Blood pH	Bone mineral density	Lean tissues weight	Fat weight
1	129S1/SvImJ	f	66	116	19.3	20.5	1.062	64	1.2	7.24	0.0605	14.5	4.4
2	129S1/SvImJ	f	66	116	19.1	20.8	1.089	78	1.15	7.27	0.0553	13.9	4.4
3	129S1/SvImJ	f	66	108	17.9	19.8	1.106	90	1.16	7.26	0.0546	13.8	2.9
368	129S1/SvImJ	f	72	114	18.3	21	1.148	65	1.26	7.22	0.0599	15.4	4.2
369	129S1/SvImJ	f	72	115	20.2	21.9	1.084	55	1.23	7.3	0.0623	15.6	4.3
370	129S1/SvImJ	f	72	116	18.8	22.1	1.176		1.21	7.28	0.0626	16.4	4.3
371	129S1/SvImJ	f	72	119	19.4	21.3	1.098	49	1.24	7.24	0.0632	16.6	5.4
372	129S1/SvImJ	f	72	122	18.3	20.1	1.098	73	1.17	7.19	0.0592	16	4.1
4	129S1/SvImJ	f	66	109	17.2	18.9	1.099	41	1.25	7.29	0.0513	14	3.2
5	129S1/SvImJ	f	66	112	19.7	21.3	1.081	129	1.14	7.22	0.0501	16.3	5.2
10	129S1/SvImJ	m	66	112	24.3	24.7	1.016	119	1.13	7.24	0.0533	17.6	6.8
364	129S1/SvImJ	m	72	114	25.3	27.2	1.075	64	1.25	7.27	0.0596	19.3	5.8
365	129S1/SvImJ	m	72	115	21.4	23.9	1.117	48	1.25	7.28	0.0563	17.4	5.7
366	129S1/SvImJ	m	72	118	24.5	26.3	1.073	59	1.25	7.26	0.0609	17.8	7.1
367	129S1/SvImJ	m	72	122	24	26	1.083	69	1.29	7.26	0.0584	19.2	4.6
6	129S1/SvImJ	m	66	116	21.6	23.3	1.079	78	1.15	7.27	0.0497	17.2	5.7
7	129S1/SvImJ	m	66	107	22.7	26.5	1.167	90	1.18	7.28	0.0493	18.7	7
8	129S1/SvImJ	m	66	108	25.4	27.4	1.079	35	1.24	7.26	0.0538	18.9	7.1
9	129S1/SvImJ	m	66	109	24.4	27.5	1.127	43	1.29	7.29	0.0539	19.5	7.1

1. Add a column to the table
2. Fill it with **=RAND()**
3. Sort all the table by this column

4. Assume that these mice is a population with size N=790. Build 5 samples with n=20
5. Calculate m, s for ending weight and p – proportion of males for each sample

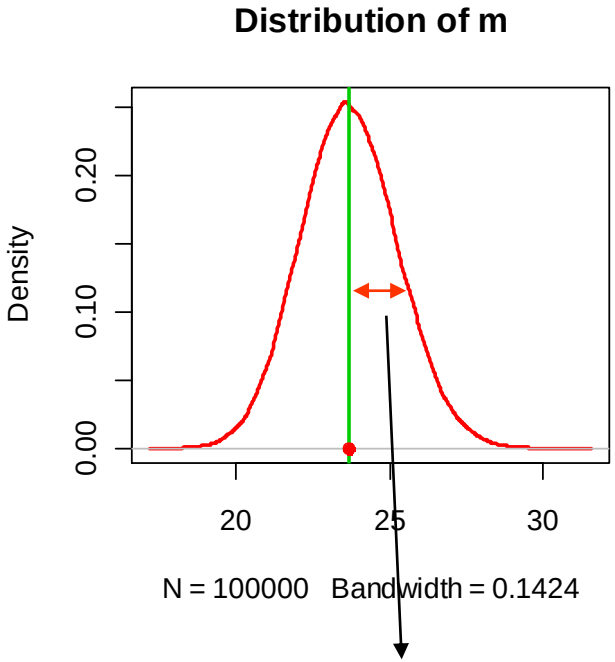
Point estimator

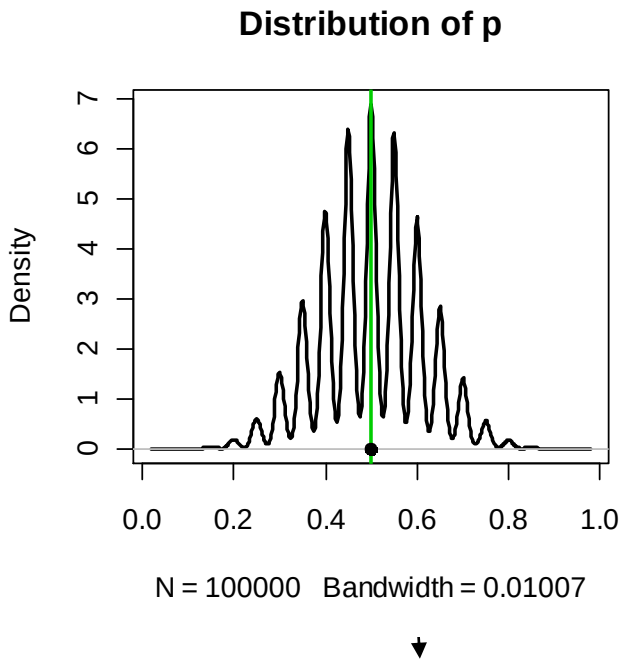
The sample statistic, such as m, s, or p, that provides the point estimation the population parameters μ, σ, π.

SAMPLING DISTRIBUTION

Sampling Distribution

Sampling distribution
A probability distribution consisting of all possible values of a sample statistic.





$$E(m) = \mu$$

$$E(p) = \pi$$

$$\sigma_m = \frac{\sigma}{\sqrt{n}}$$

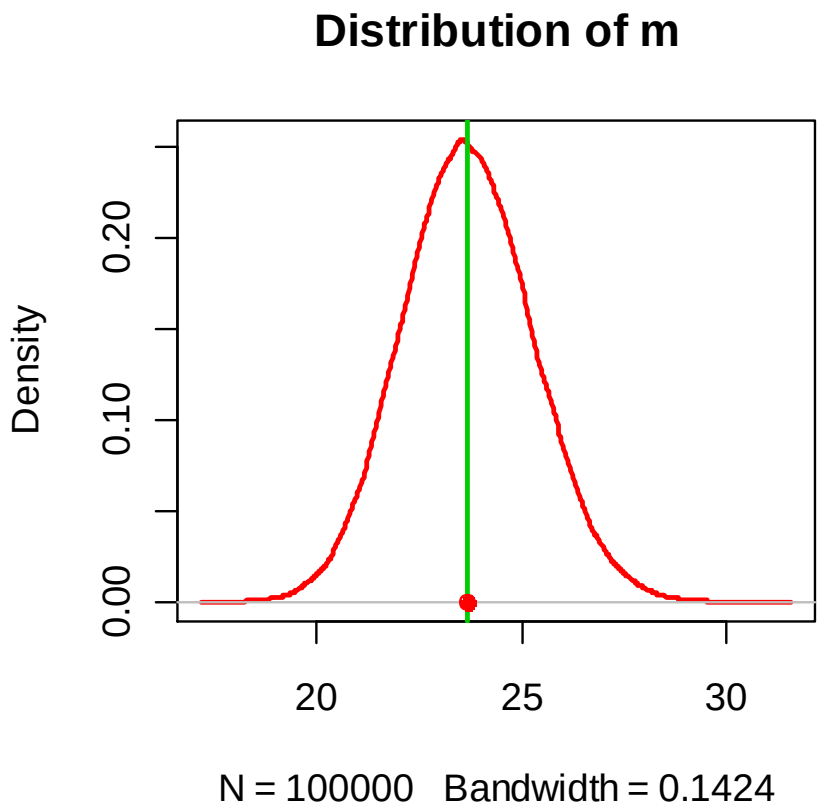
$$\sigma_p = \sqrt{\frac{\pi(1-\pi)}{n}}$$

SAMPLING DISTRIBUTION

Unbiased Point Estimator: mean

Unbiased

A property of a point estimator that is present when the expected value of the point estimator is equal to the population parameter it estimates.

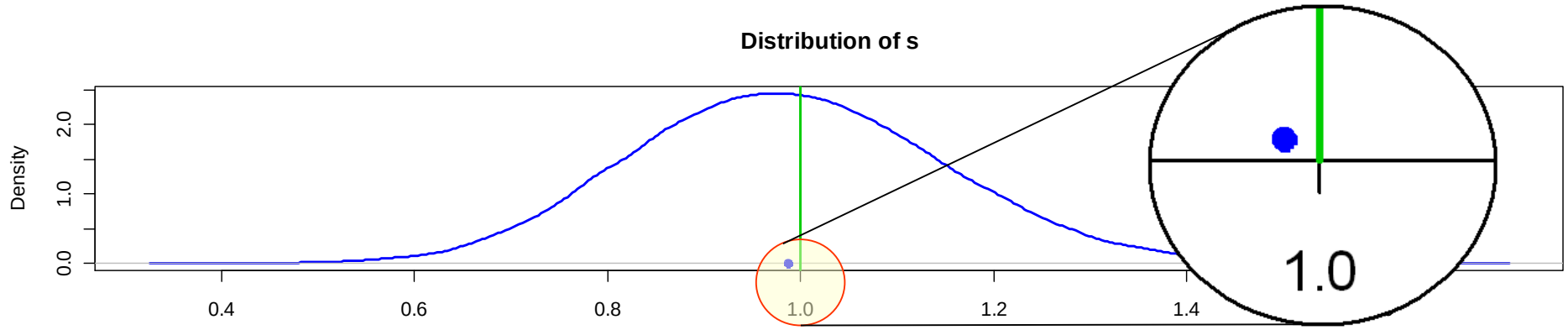


SAMPLING DISTRIBUTION

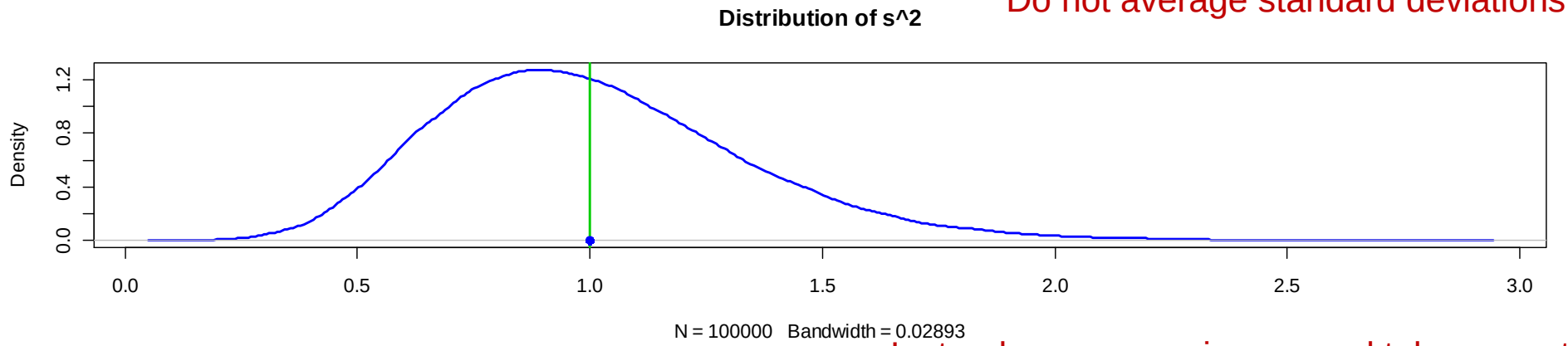
Unbiased Point Estimator: variance

Unbiased

A property of a point estimator that is present when the expected value of the point estimator is equal to the population parameter it estimates.



Do not average standard deviations



Instead average variances and take sq.root

SAMPLING DISTRIBUTION

Correction for Finite Population

Finite population correction factor

The term $\sqrt{\frac{N-n}{N-1}}$ that is used in the formulas for σ_m and σ_p whenever a finite population, rather than an infinite population, is being sampled. The generally accepted rule of thumb is to ignore the finite population correction factor whenever $\frac{n}{N} \leq 0.05$.

Standard deviation of the sample mean *m*

Finite population

$$\sigma_m = \sqrt{\frac{N-n}{N-1}}\left(\frac{\sigma}{\sqrt{n}}\right)$$

Infinite population

$$\sigma_m = \frac{\sigma}{\sqrt{n}}$$

Standard deviation of the sample proportion *p*

$$\sigma_p = \sqrt{\frac{N-n}{N-1}}\sqrt{\frac{\pi(1-\pi)}{n}}$$

$$\sigma_p = \sqrt{\frac{\pi(1-\pi)}{n}}$$

Standard error

The standard deviation of a point estimator.

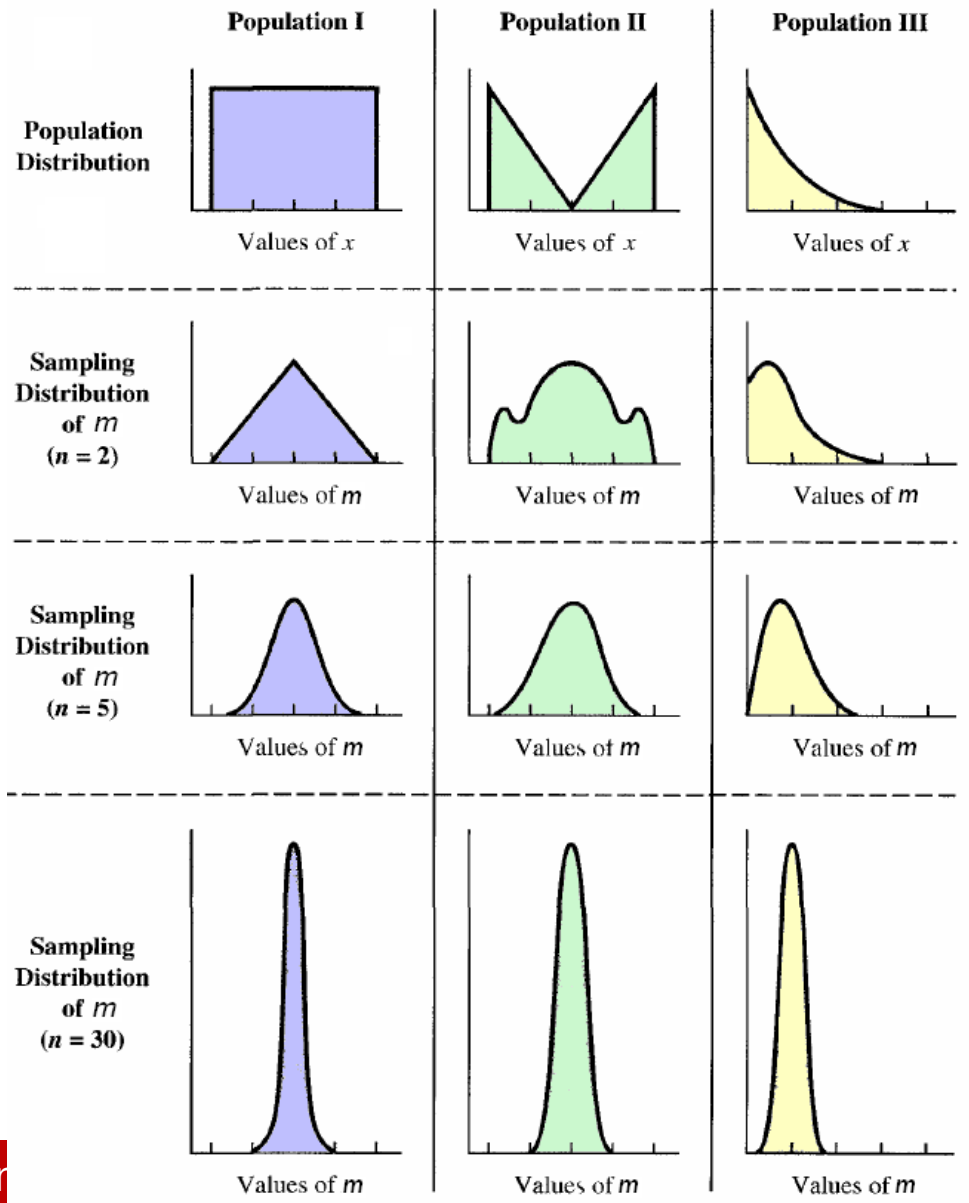
SAMPLING DISTRIBUTION

Central Limit Theorem

Central limit theorem

In selecting simple random sample of size *n* from a population, the ***sampling distribution of the sample mean m can be approximated by a normal distribution*** as the sample size becomes large

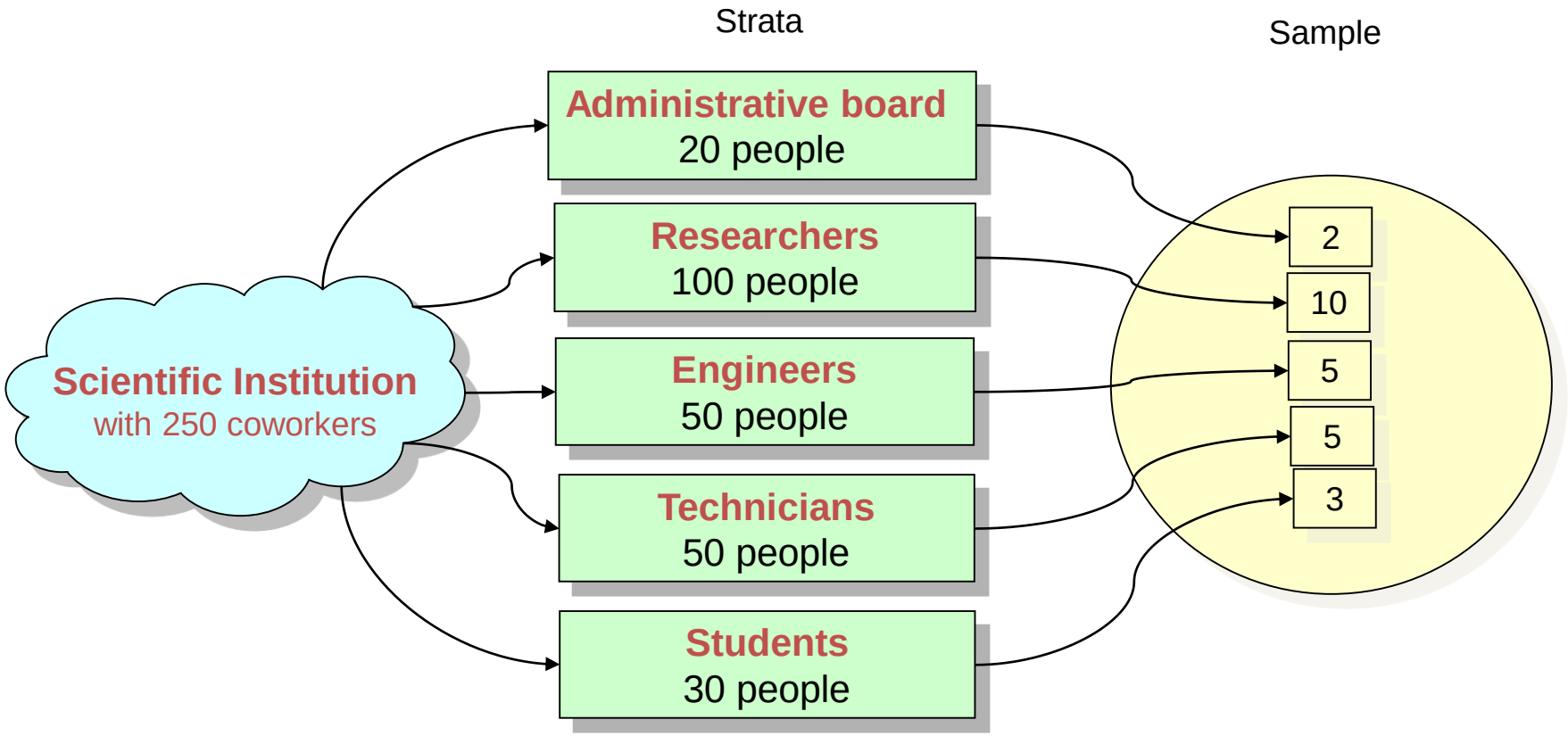
In practice if the sample size is ***n*>30**, the normal distribution is a good approximation for the sample mean for any initial distribution.



SAMPLING METHODS

Stratified Sampling

Stratified random sampling
A probability sampling method in which the population is first divided into strata and a simple random sample is then taken from each stratum.



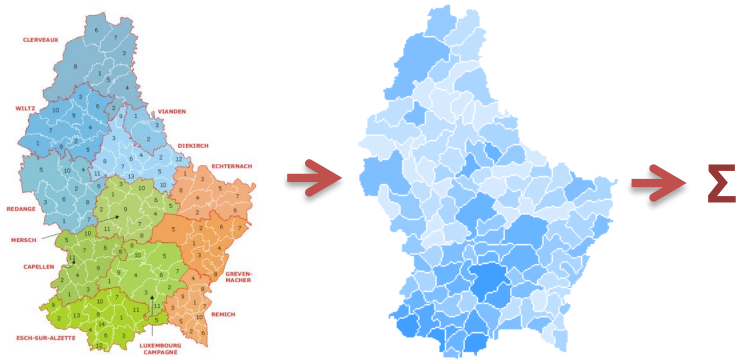
SAMPLING METHODS

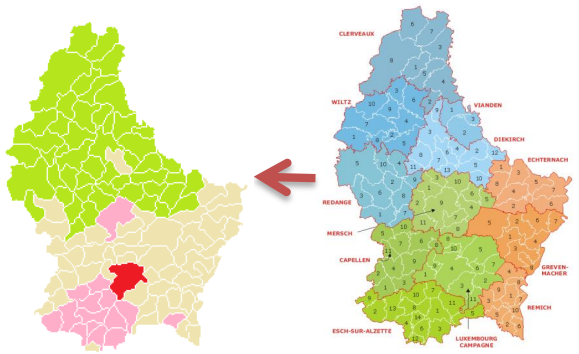
Stratified Sampling

Stratified random sampling
A probability sampling method in which the population is first divided into strata and a simple random sample is then taken from each stratum.

Strategies of Stratified Sampling

- Proportionate allocation
  - Selected size n_i of a sub-sample depends on **population size N_i** of a strata
- Optimum (disproportionate) allocation
  - Selected size n_i of a sub-sample depends on **variance σ_i^2** of a strata





SAMPLING METHODS

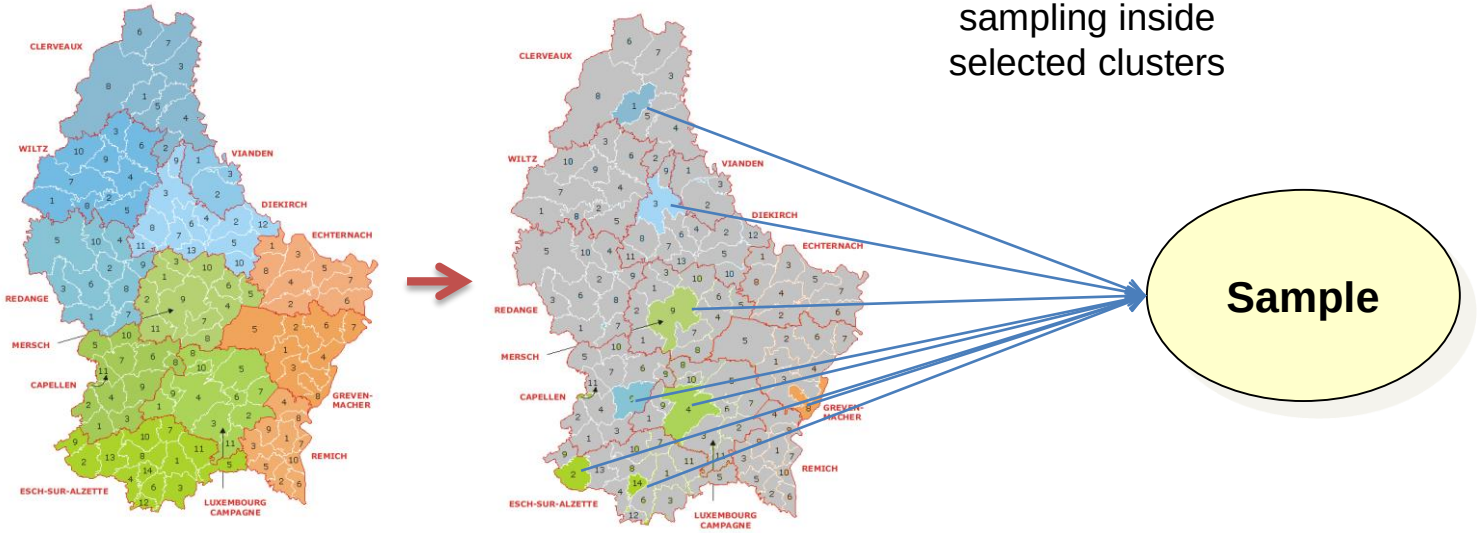
Cluster Sampling

Cluster sampling

A probability sampling method in which the population is first divided into clusters and then a simple random sample of the clusters is taken.

1. Random sampling of sampling based on cost optimization

2. Simple random sampling inside selected clusters

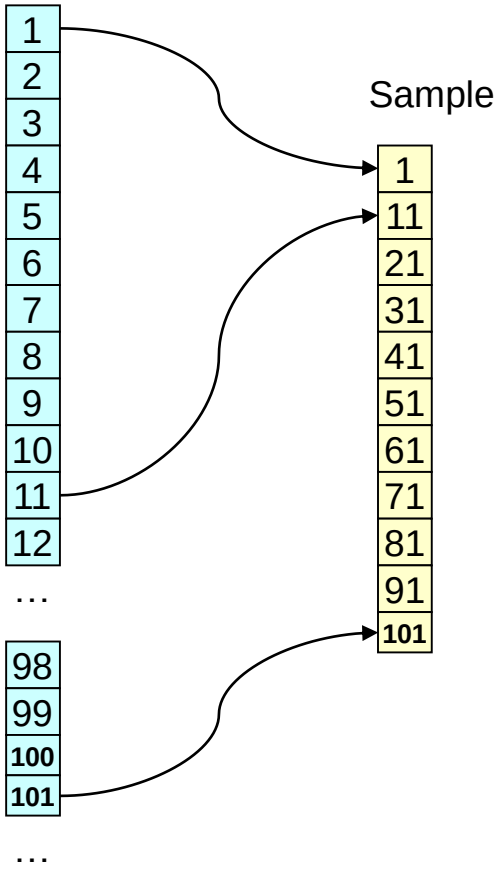


SAMPLING METHODS

Systematic Sample

Systematic sampling

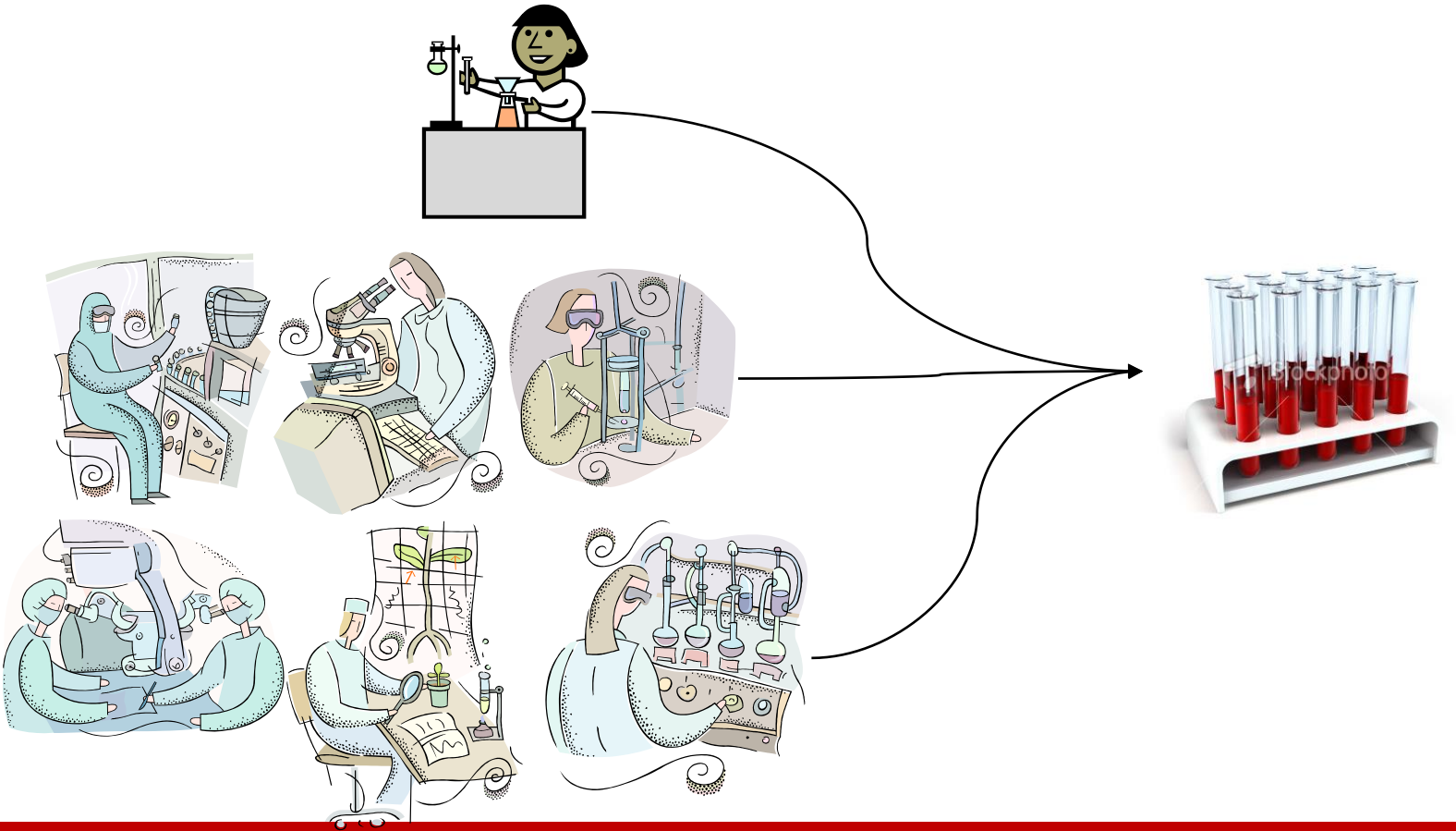
A probability sampling method in which we randomly select one of the first *k* elements and then select every *k*-th element thereafter.



SAMPLING METHODS

Convenience Sampling

Convenience sampling
A nonprobability method of sampling whereby elements are selected for the sample on the basis of convenience.



SAMPLING METHODS

Judgment Sampling

Judgment sampling
A nonprobability method of sampling whereby elements are selected for the sample based on the judgment of the person doing the study.

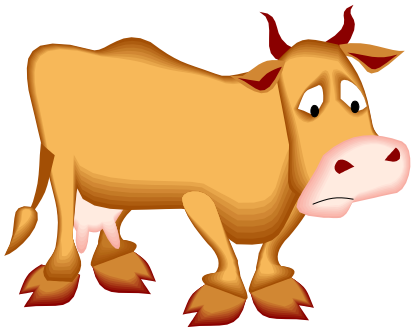
Perform of a selection of most confident or most experienced experts.

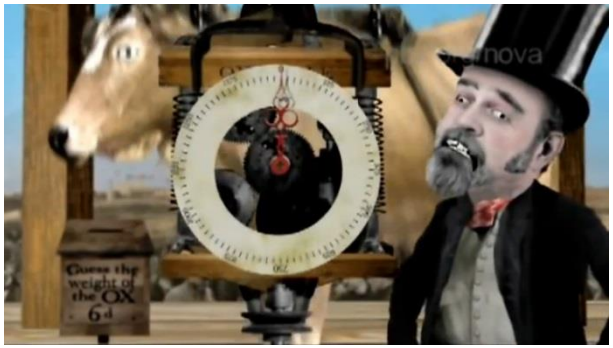
SAMPLING METHODS

The Wisdom of the Crowd

The wisdom of the crowd
is the process of taking into account the collective opinion of a group of individuals rather than a single expert to answer a question. A large group's aggregated answers to questions involving quantity estimation has generally been found to be as good as, and often better than, the answer given by any of the individuals within the group.

The classic wisdom-of-the-crowds finding involves point estimation of a continuous quantity. At a 1906 country fair in Plymouth, eight hundred people participated in a contest to estimate the weight of a slaughtered and dressed ox. Statistician Francis Galton observed that the median guess, 1207 pounds, was accurate within 1% of the true weight of 1198 pounds.



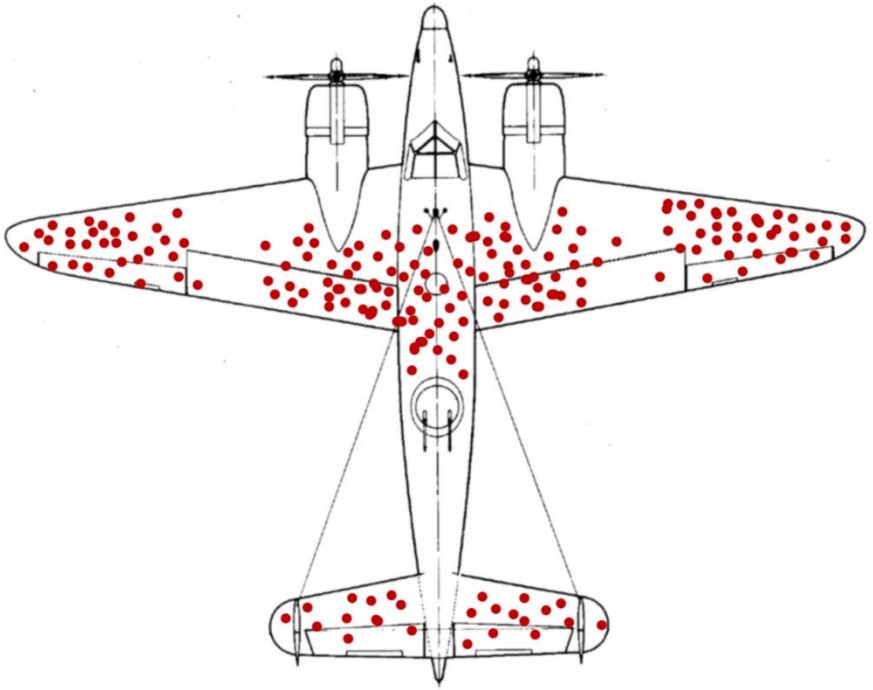


http://www.youtube.com/watch?v=r-FonWBEb0o

AN EXAMPLE

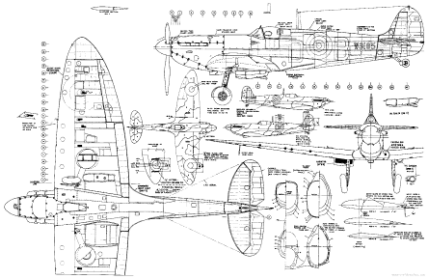
Be Careful with Sampling!!!

‘Spitfire’: damage analysis



Were to put an additional protection?

Another example: paleolithic remains -> lifestyle



QUESTIONS ?

Thank you for your attention

to be continued…