CENTRE DE RECHERCHE PUBLIC

Microarray Center

APPLIED STATISTICS

Lecture 5

Sampling and Sampling Distribution

Petr Nazarov
petr.nazarov@crp-sante.lu

12-11-2009

- **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
- **Interval estimation**
  - interval estimation
  - population mean: σ known
  - population mean: σ unknown
  - Student's distribution
  - estimation the size of a sample
  - population proportion

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.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

Simple random sampling

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

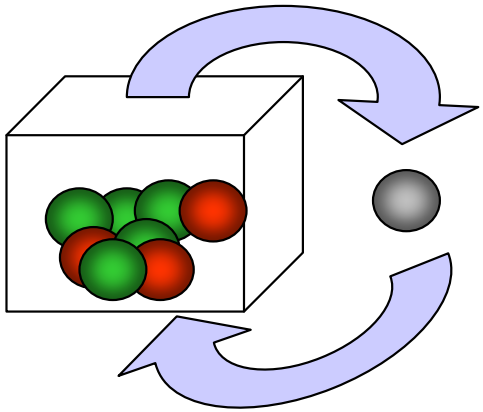
Infinite population: a sample selected such that each element comes from the same population and the elements are selected independently.

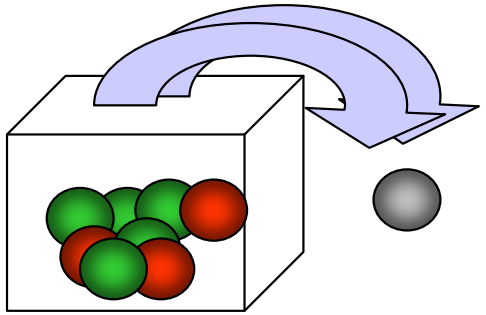
Sampling without replacement

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.





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

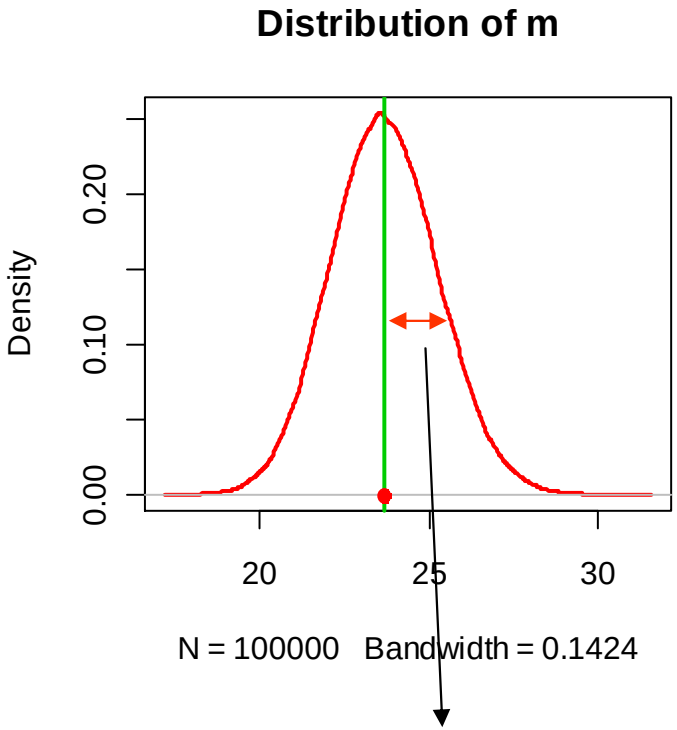
1. Add a column to the table
2. Fill it with **=RAND()**
3. Sort all the table by this column

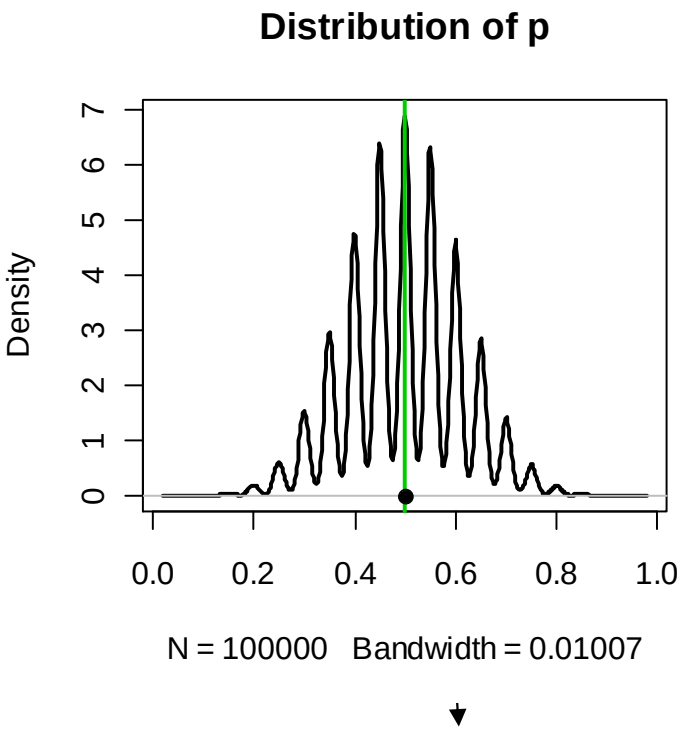
4. Assume that these mice is a population with size N=790. Build 3 samples with n=20
5. Calculate *m*, *s* for ending weight and *p* – proportion of males for each sample

Point estimator

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

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





$$E(m) = \mu$$

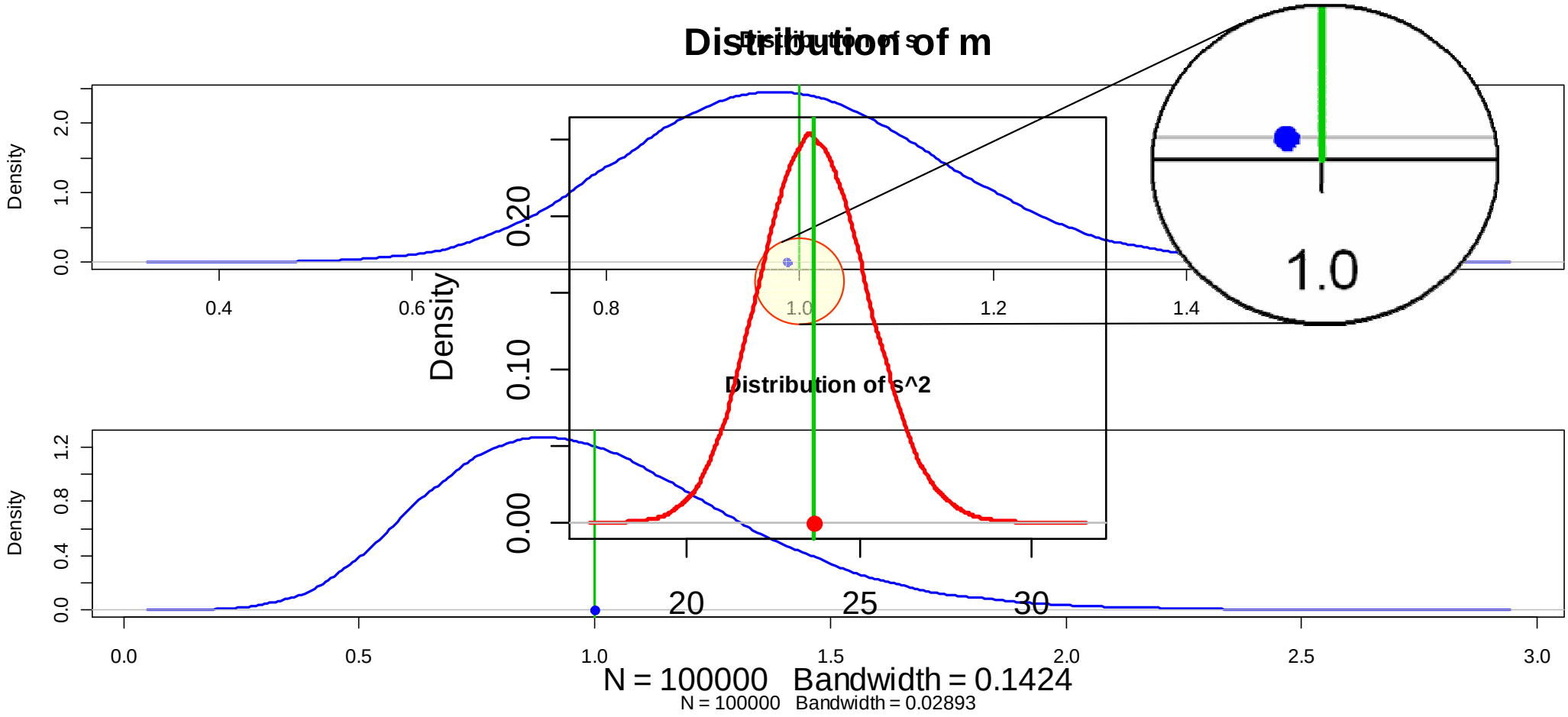
$$E(p) = \pi$$

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

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

Unbiased Point Estimator

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.

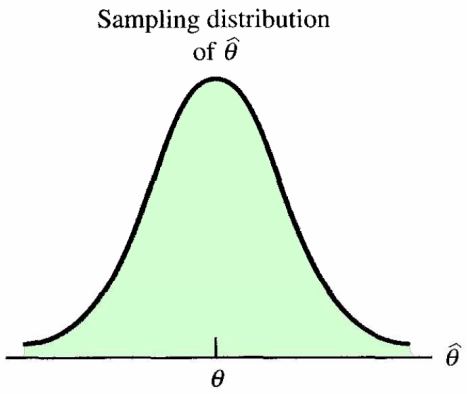


Relative efficiency

Given two unbiased point estimators of the same population parameter, the point estimator with the smaller standard deviation is more efficient.

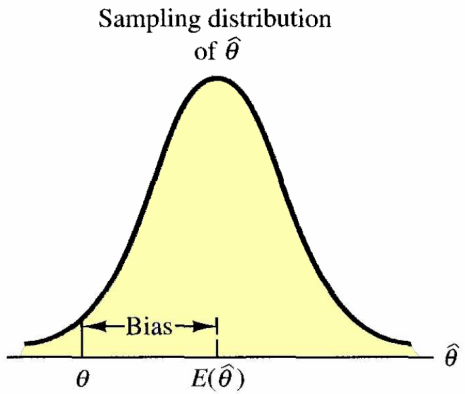
Consistency

A property of a point estimator that is present whenever larger sample sizes tend to provide point estimates closer to the population parameter.



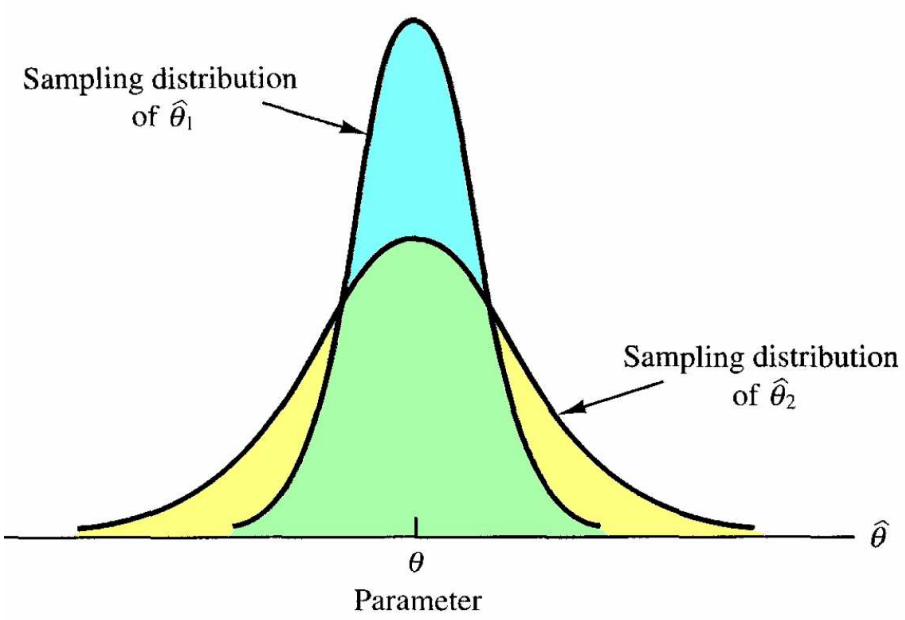
Parameter θ is located at the mean of the sampling distribution; $E(\hat{\theta}) = \theta$

Panel A: Unbiased Estimator



Parameter θ is not located at the mean of the sampling distribution; $E(\hat{\theta}) \neq \theta$

Panel B: Biased Estimator



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 error
The standard deviation of a point estimator.

Standard deviation of the sample proportion *p*

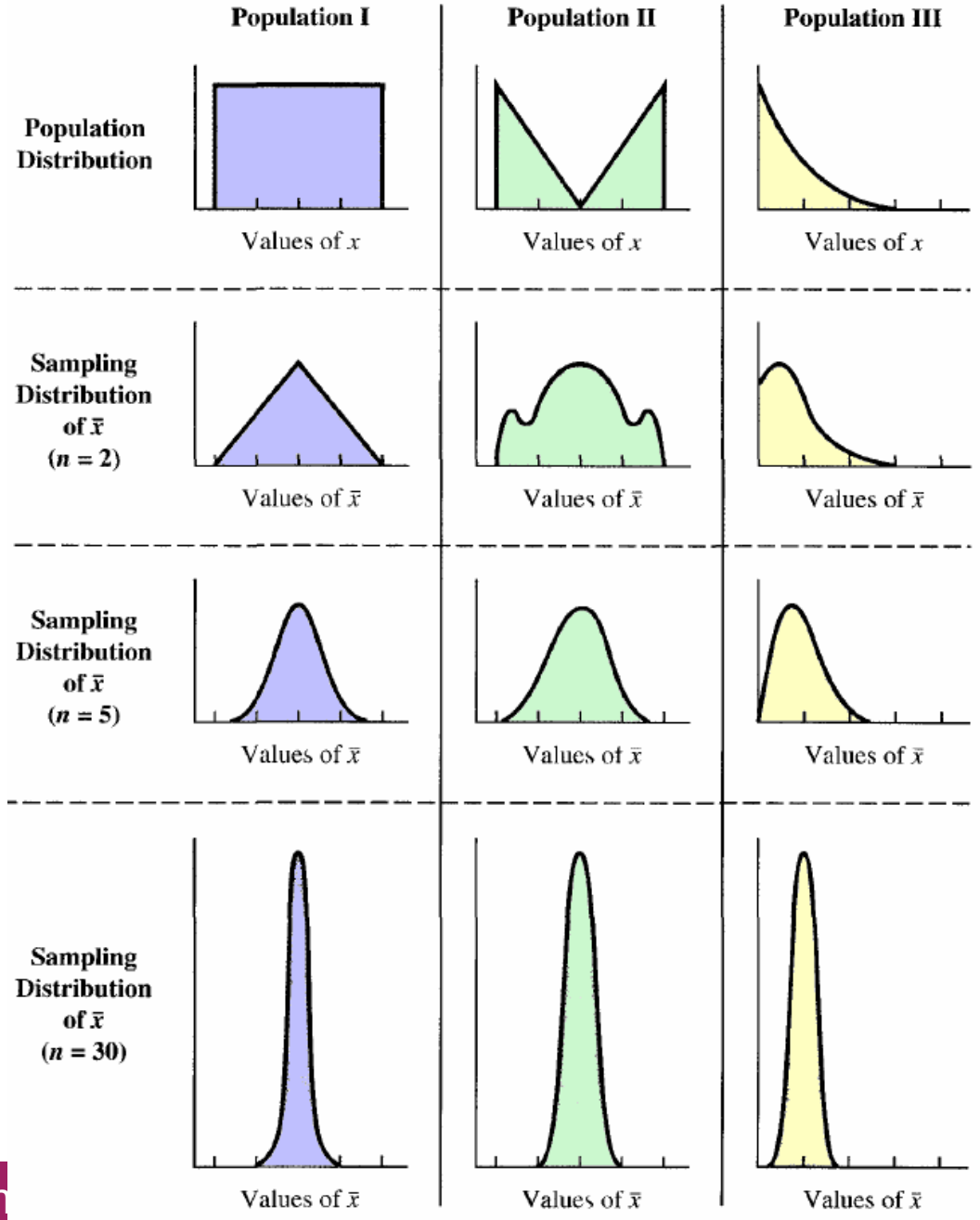
$$\sigma_p = \sqrt{\frac{N-n}{N-1}}\sqrt{\frac{\pi(1-\pi)}{n}}$$

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

Central limit theorem

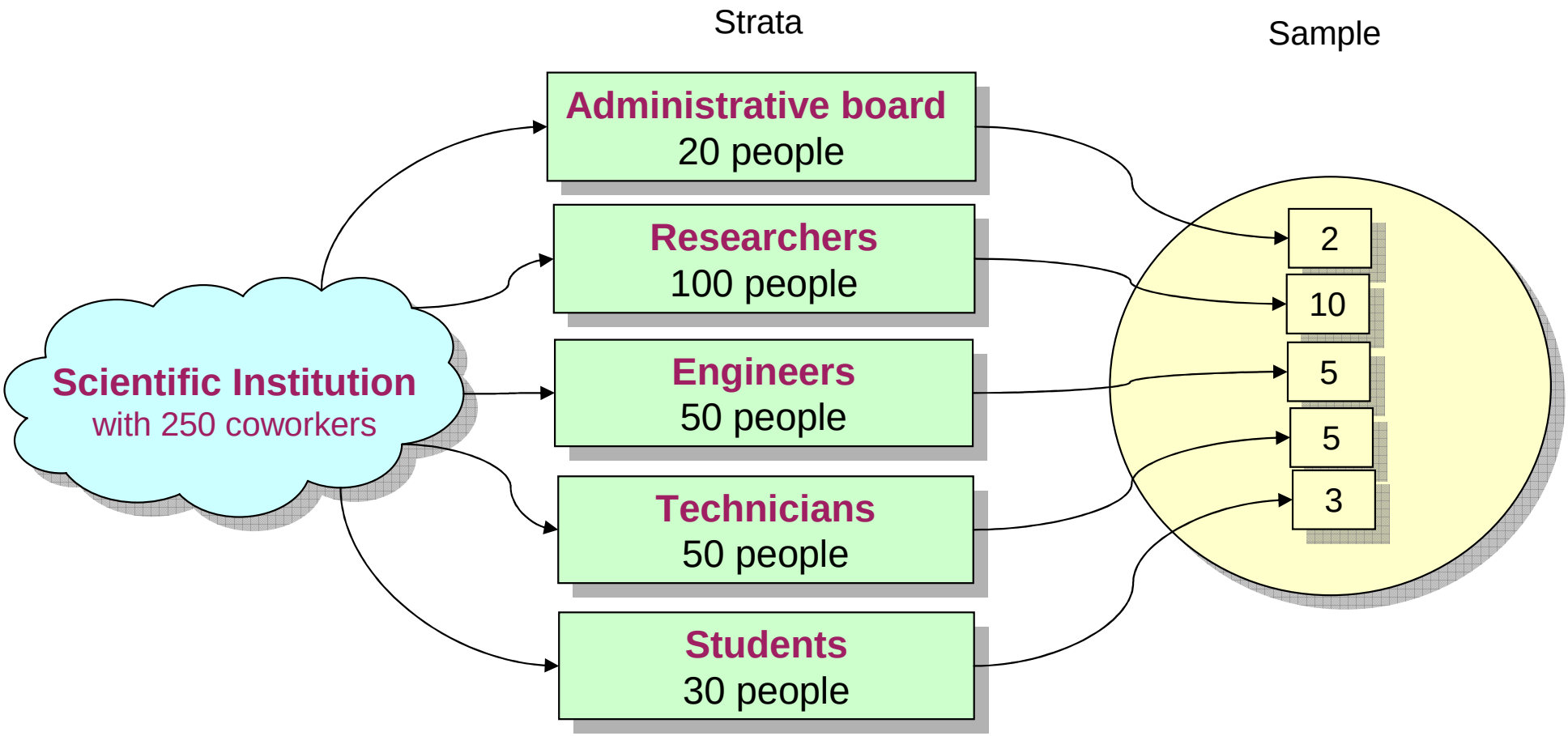
In selecting simple random sample of size *n* from a population, the ***sampling distribution of the sample mean m can be approximated by a normal distribution*** as the sample size becomes large

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



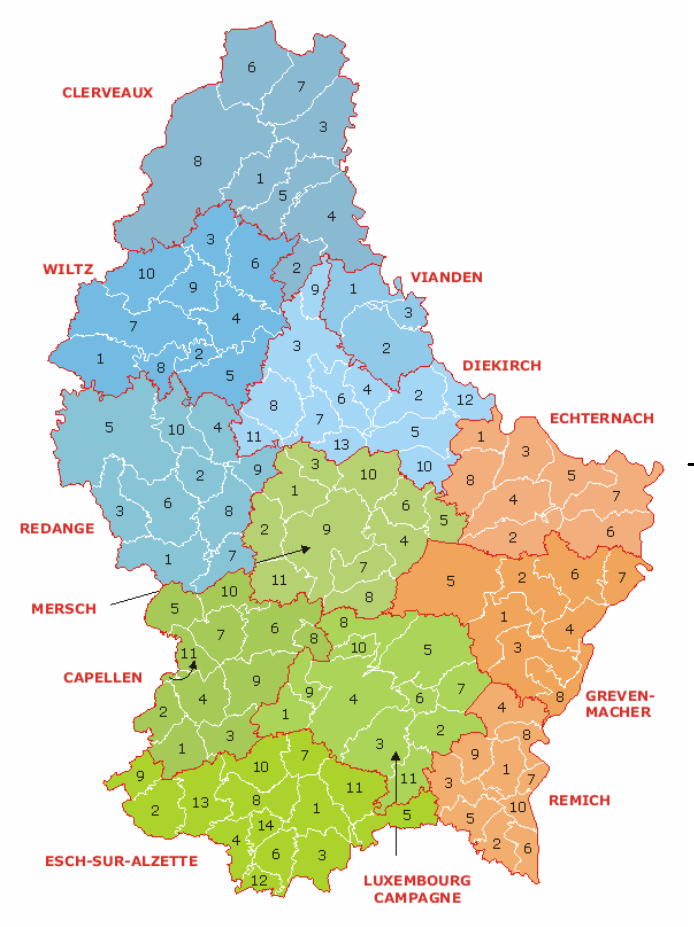
NOTE: here and below $\overline{x}$ will be used together with *m* as a sample mean.

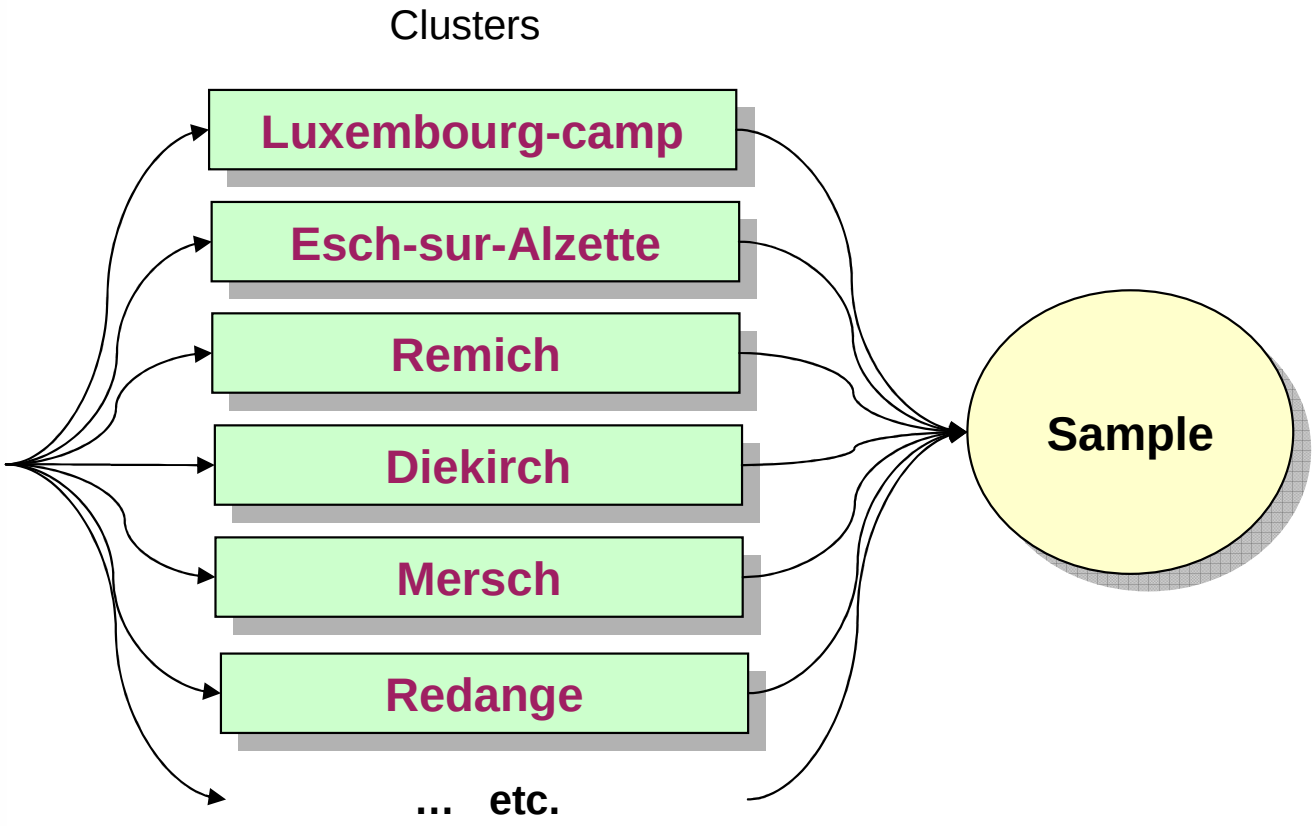
Stratified random sampling
A probability sampling method in which the population is first divided into strata and a simple random sample is then taken from each stratum.



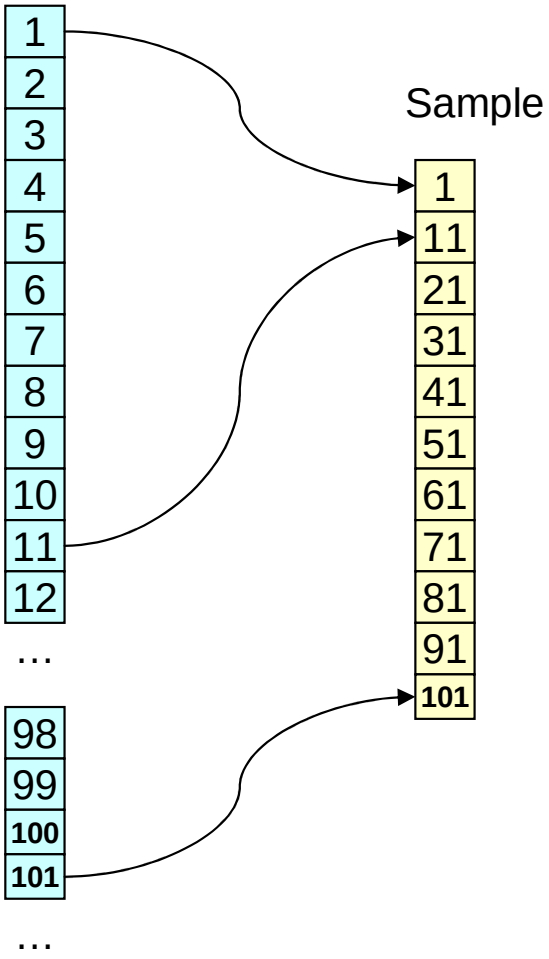
Cluster sampling

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

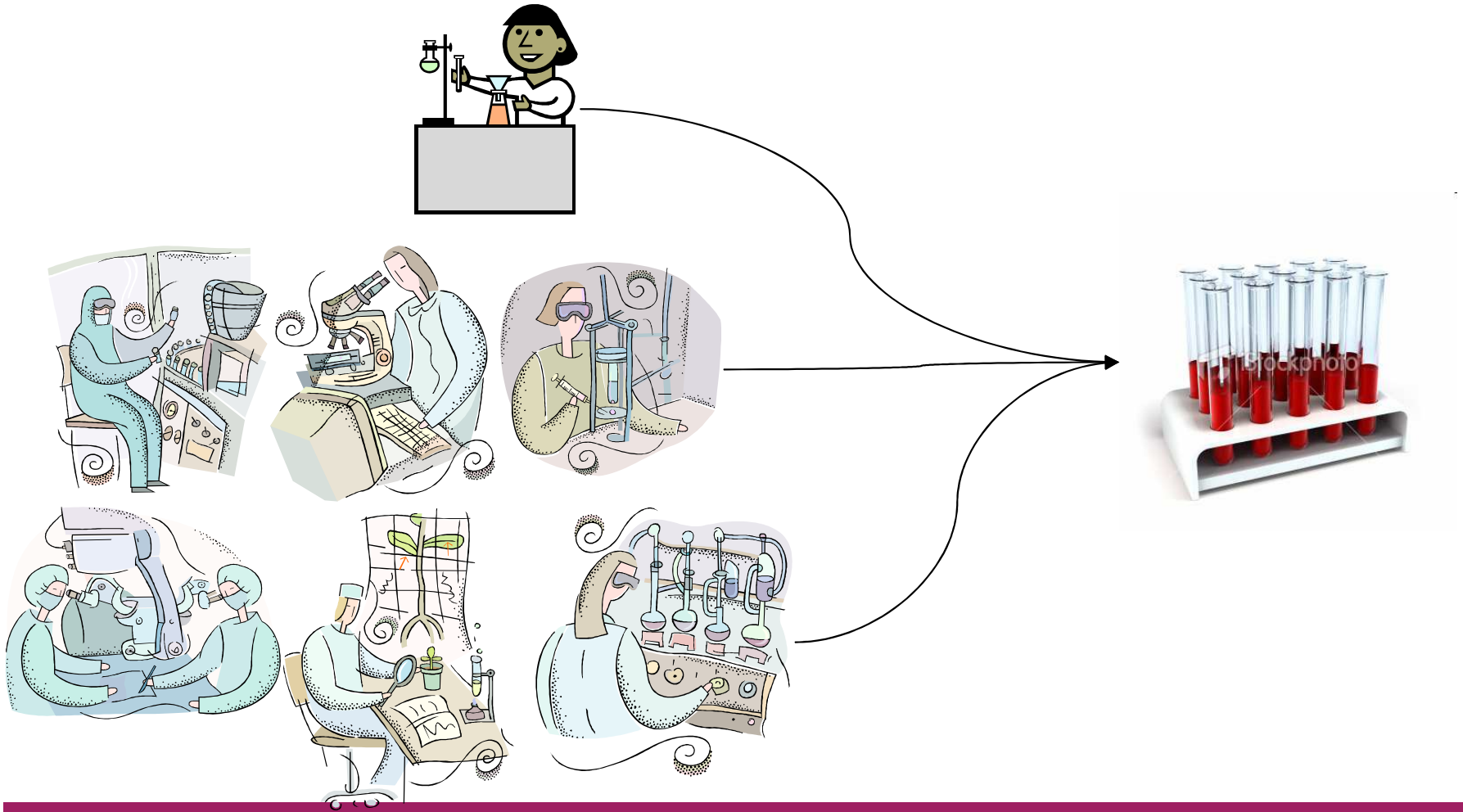




Systematic sampling
A probability sampling method in which we randomly select one of the first *k* elements and then select every *k*-th element thereafter.



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

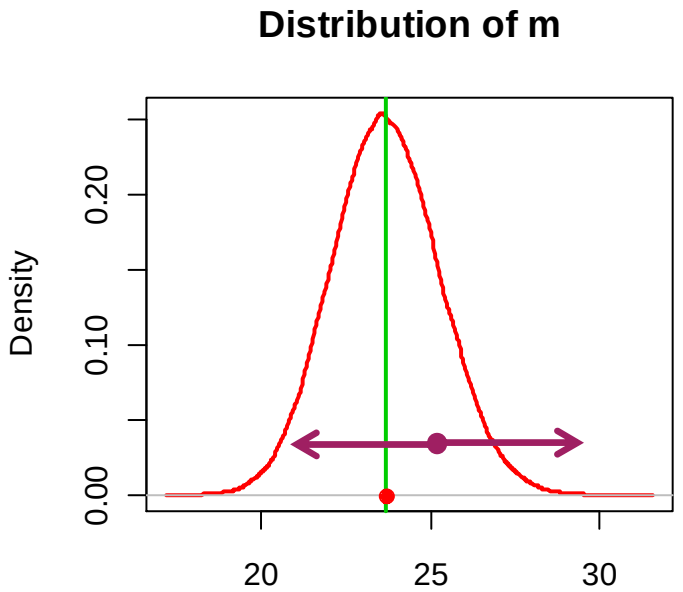


Judgment sampling
A nonprobability method of sampling whereby elements are selected for the sample based on the judgment of the person doing the study.

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

Interval estimate
An estimate of a population parameter that provides an interval believed to contain the value of the parameter. For the interval estimates in this chapter, it has the form: point estimate ± margin of error.

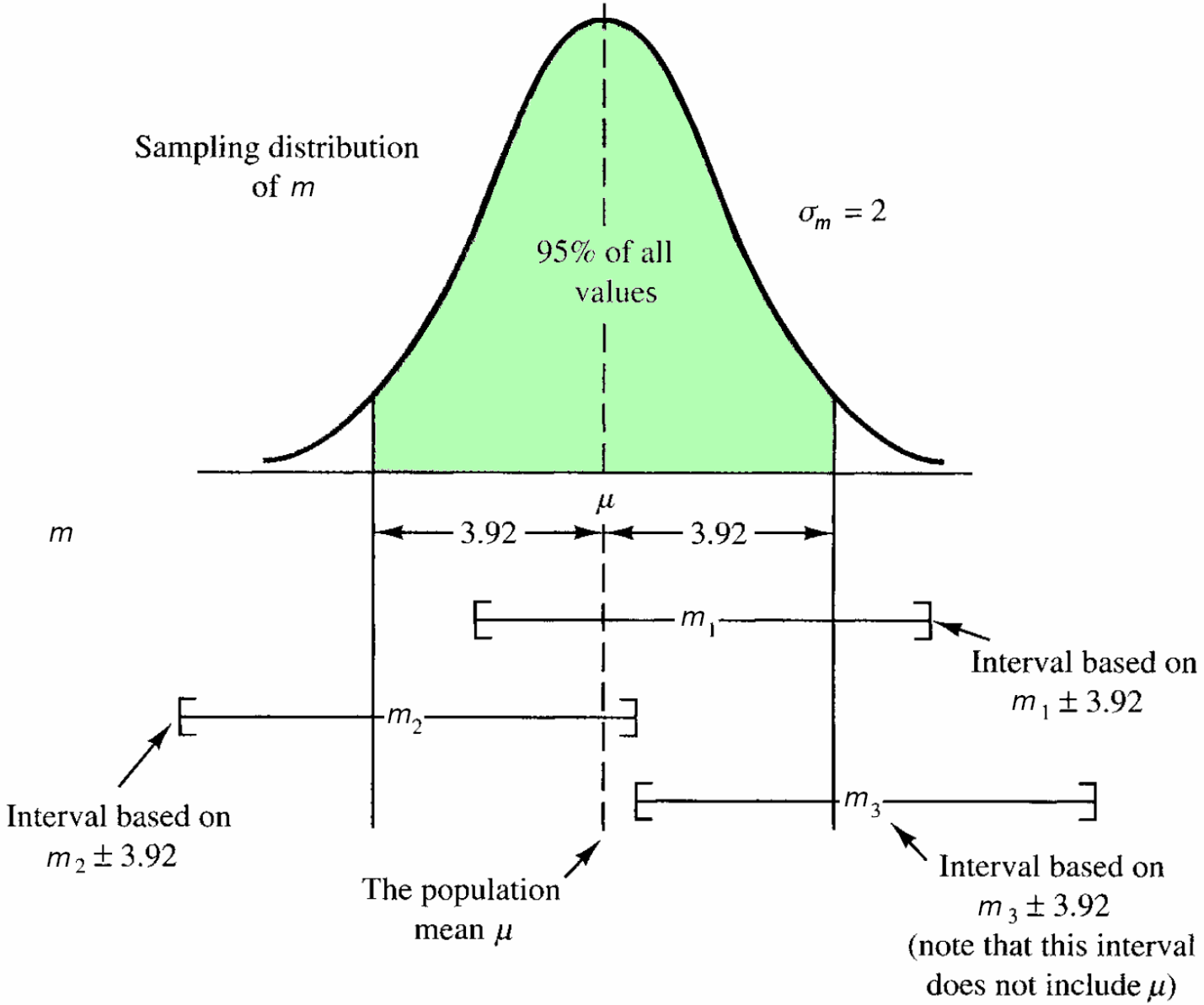
Margin of error
The ± value added to and subtracted from a point estimate in order to develop an interval estimate of a population parameter.



$$\mu = m \pm merror$$

σ known
The condition existing when historical data or other information provides a good value for the population standard deviation prior to taking a sample. The interval estimation procedure uses this known value of o in computing the margin of error.

σ unknown
The condition existing when no good basis exists for estimating the population standard deviation prior to taking the sample. The interval estimation procedure uses the sample standard deviation s in computing the margin of error.


Sampling distribution of m
$\sigma_m = 2$
95% of all values
μ
m
3.92
3.92
m_1
Interval based on $m_1 \pm 3.92$
m_2
Interval based on $m_2 \pm 3.92$
m_3
The population mean μ
Interval based on $m_3 \pm 3.92$ (note that this interval does not include μ)

Confidence level

The confidence associated with an interval estimate. For example, if an interval estimation procedure provides intervals such that 95% of the intervals formed using the procedure will include the population parameter, the interval estimate is said to be constructed at the 95% confidence level.

Confidence interval

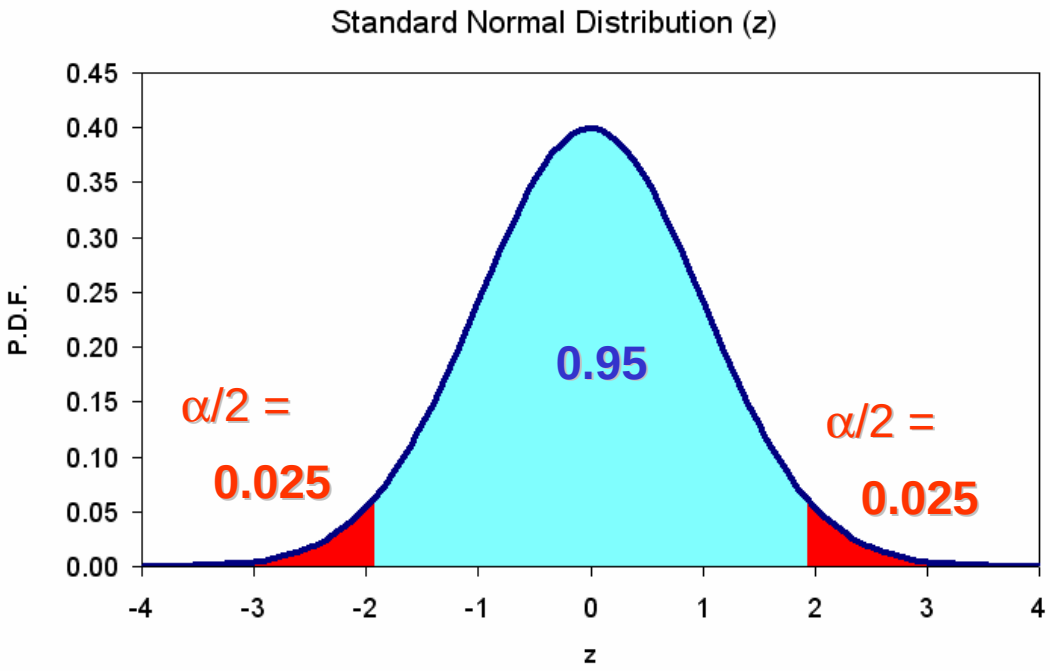
Another name for an interval estimate.

$$\mu = m \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

For 95 % confidence α = 0.05, which means that in each tail we have 0.025. Corresponding $z_{\alpha/2} = 1.96$

In Excel use one of the following functions:

- `= CONFIDENCE(alpha, σ, n)`
- `= -NORMINV(alpha/2,0,1)*σ/SQRT(n)`



Assume that we have a sample of 20 mice and would like to estimate an average size of a mice in population.

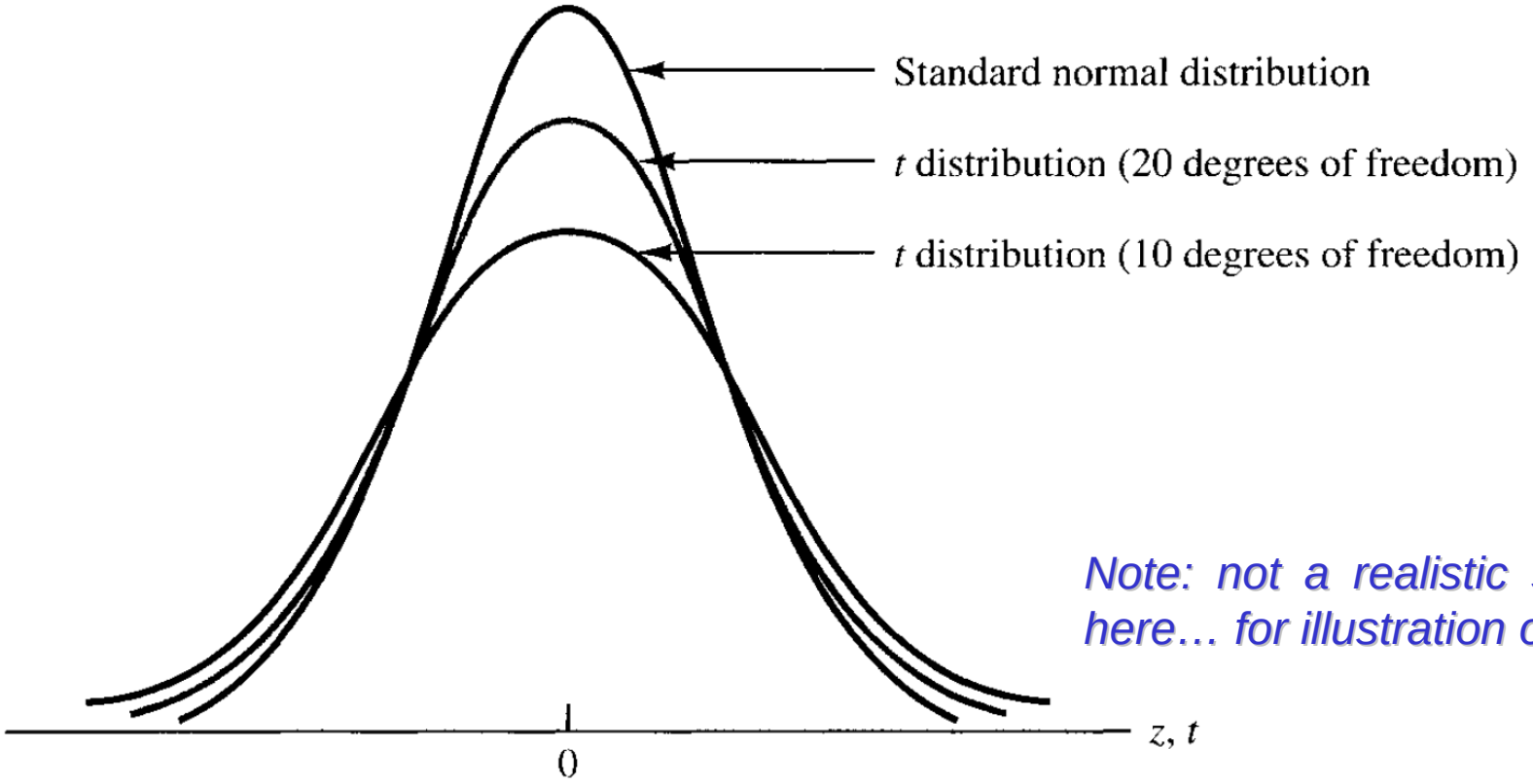
Weight
39.9
19.8
32.4
21
27.5
20.8
21.3
40
10.7
22.6
27
10.8
20.9
14.7
31.4
17.2
11.4
19.1
31.3
14.8

$m = 22.73$

$s = 8.84$

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

As we replace σ → s, we introduce an additional error and this change the distribution from *z* to *t* (Student)



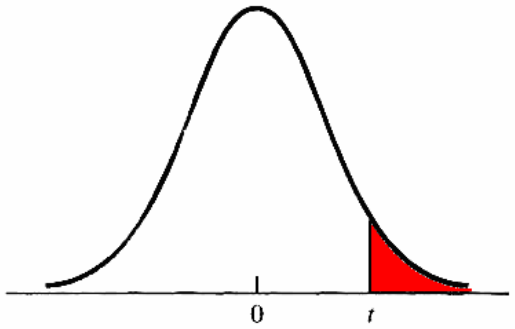
Note: not a realistic scale here… for illustration only

_t_-distribution

A family of probability distributions that can be used to develop an interval estimate of a population mean whenever the population standard deviation σ is unknown and is estimated by the sample standard deviation s.

Degrees of freedom

A parameter of the t-distribution. When the t distribution is used in the computation of an interval estimate of a population mean, the appropriate t distribution has $n - 1$ degrees of freedom, where n is the size of the simple random sample.



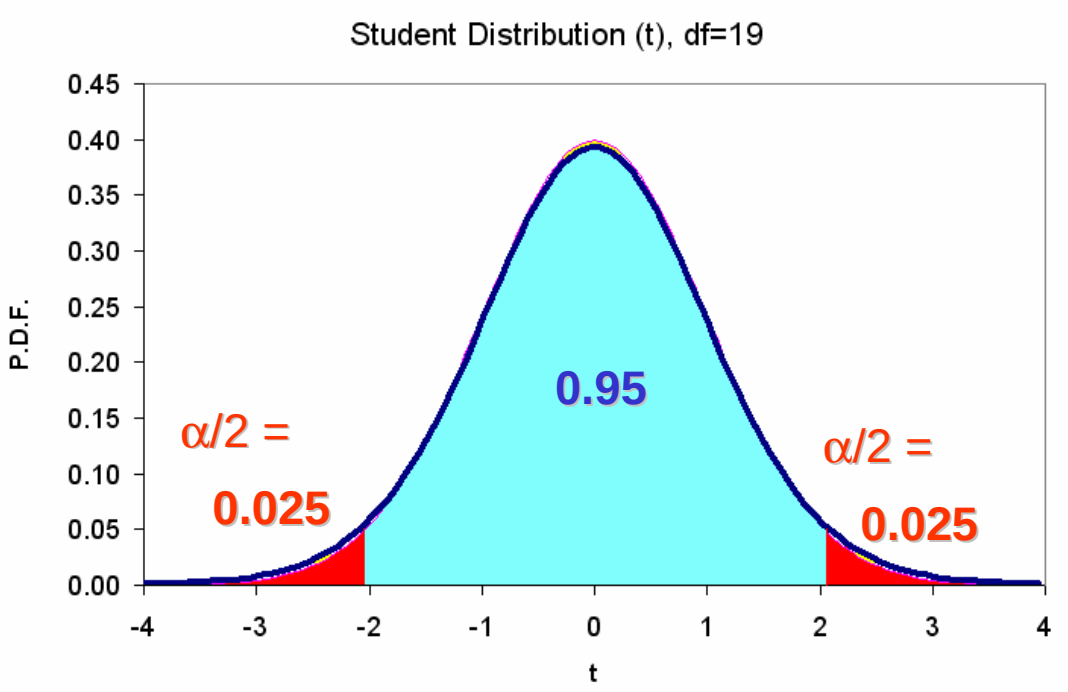
Degrees of Freedom	Area in Upper Tail					
	.20	.10	.05	.025	.01	.005
1	1.376	3.078	6.314	12.706	31.821	63.656
2	1.061	1.886	2.920	4.303	6.965	9.925
3	.978	1.638	2.353	3.182	4.541	5.841
4	.941	1.533	2.132	2.776	3.747	4.604

Population Mean: σ Unknown

Weight
39.9
19.8
32.4
21
27.5
20.8
21.3
40
10.7
22.6
27
10.8
20.9
14.7
31.4
17.2
11.4
19.1
31.3
14.8

m =	22.73
s =	8.84
s(m) =	1.98
t =	2.09
m.e. =	4.14

$$\mu = m \pm t_{\alpha/2}^{(n-1)} \frac{s}{\sqrt{n}}$$



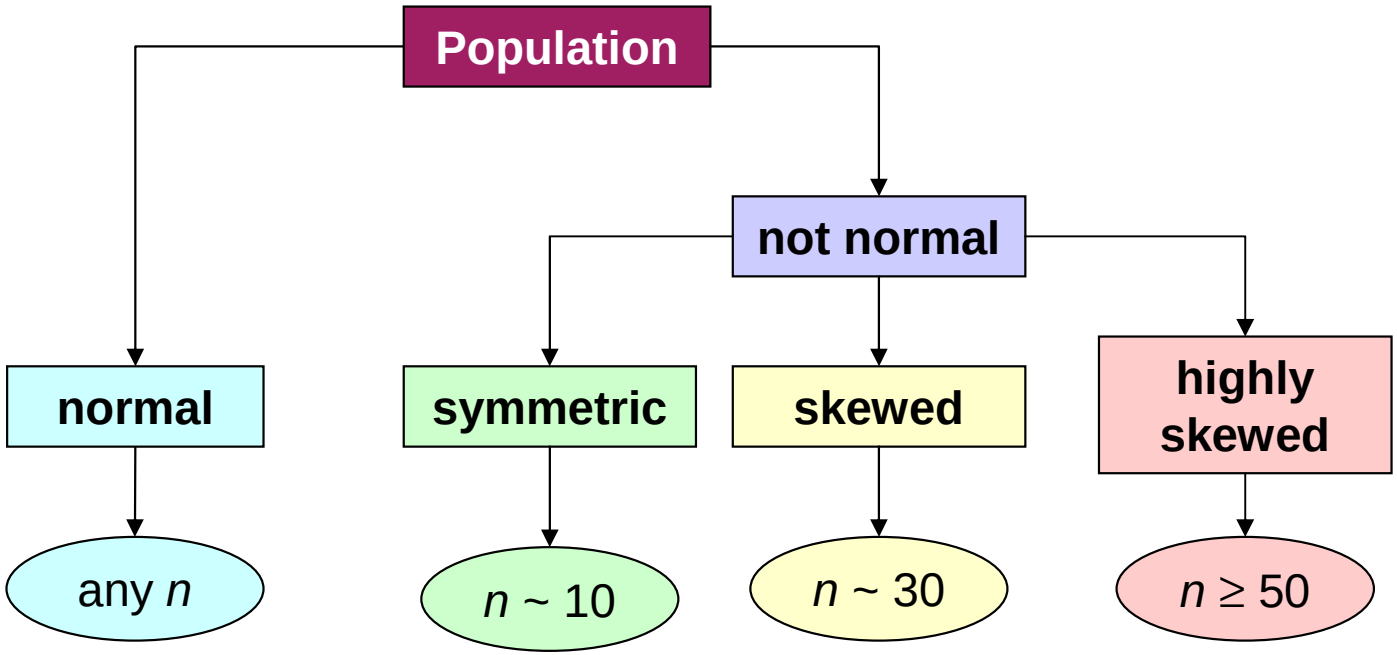
In Excel use:

- `= TINV(alpha, degree-of-freedom)`

Population Mean: Practical Advices

$$\mu = m \pm t_{\alpha/2}^{(n-1)} \frac{s}{\sqrt{n}}$$

Advice 1



Advice 2

if n >100 you can use z-statistics instead of t-statistics (error will be <1.5%)

Determining the Sample Size

Let's focus on another aspect: how to select a proper number of experiments.

$$\mu = m \pm E(n, \sigma)$$
$$E(n, \sigma) = E$$
$$n - ?$$

$$E = z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

$$n = \frac{z_{\alpha/2}^2 \sigma^2}{E^2}$$

$$n = \frac{z_{\alpha/2}^2 \sigma^2}{E^2}$$

$$n = \left(t_{\alpha/2}^{(n_z - 1)}\right)^2 \frac{s^2}{E^2}$$

Practical Work

children.xls

What is an the 95% confidence interval for average weight of the children in a day-care center in kg ?

Estimate the necessary sample size for an error $E \leq 1$ kg

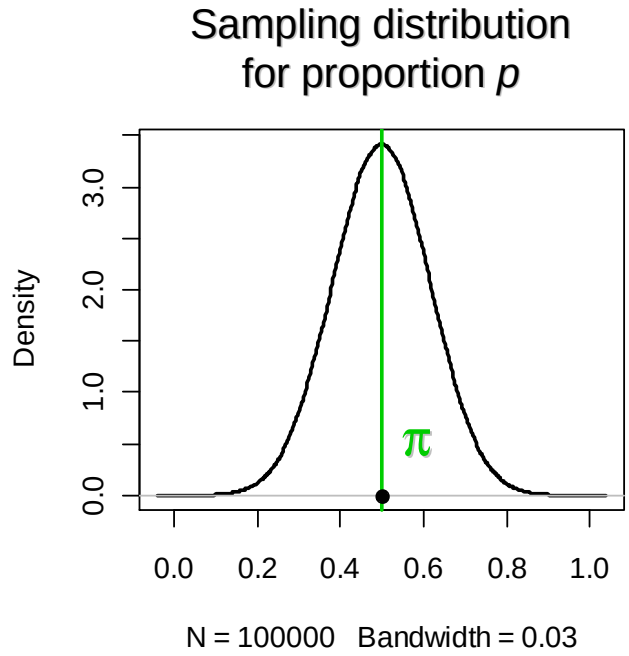
m=	16.7
s=	7.196347
n=	57
t=	2.002465
E=	1.908709
E1=	1
n1=	198.9465

Population Proportion

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

$$\pi = p \pm z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}$$

if $np \geq 5$ and $n(1-p) \geq 5$



Practical Work

pancreatitis.xls

Define a 95% confidence interval for **never-smoking** proportion of people coming to a hospital

n=	270
p(never)=	**0.214815**
sp=	0.024994
E=	**0.048988**

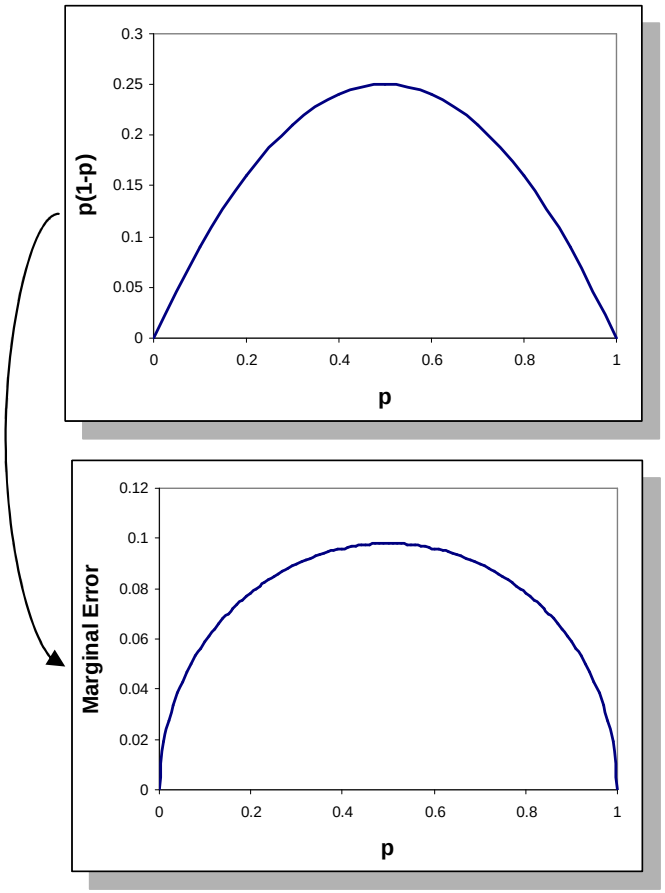
for 95% confidence $z_{0.025} = 1.96$

$\pi = 21.5 \pm 4.9$ %

Population Proportion: Some Practical Aspects

$$\pi = p \pm z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}$$

1. The normal distribution is applicable only when enough data points are observed. The rule of thumb is: $np \geq 5$ and $n(1-p) \geq 5$



2. The maximal marginal error is observed when p=0.5

3. The estimation of the sample size can be obtained:

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

$np \geq 5$ and $n(1-p) \geq 5$

where p is a best guess for π or the result of a preliminary study

Thank you for your attention

to be continued…