QUANTILES

The same method of the median can be extended to a fixed value whatsoever of F – the corresponding values are called quantiles (we can have several different sets of them).

- QUARTILES → 3 values (positions): Q1: F=1/4 (C_{25}) Q2: F=1/2=Median (C_{50}) Q3: F=3/4 (C_{75})
- CENTILES → 99 values (sometimes also non-integer centiles are used)

STATISTICAL RATIO

Statistical ratios are ratios between two quantitative statistical data (with denominator > 0). Statistical ratios are dimensionless, meaning that they are expressed without a unit of measurement; thus, they can be compared even if they refer to different kind of variables with different units of measurement.

- COMPOSITION RATIO = partial value / corresponding total → range between 0 and 1 (can be in %)
- COEXISTENCE RATIO = partial datum A / partial datum B → range between1 and + (can be in % but 100% can be overtaken)
- DERIVATION RATIO = effect datum (dynamic) / causal datum (static)
- SPACE DENSITY = numerical datum / area of observation
- TIME FREQUENCY = numerical datum / time interval of observation

Those ratios can be combined together to form indexes: HDI [human development index] and GPI [global peace index].

GINI CONCENTRATION RATIO

Can only be applied to a **transferable** (=possible to transfer a quantity from a statistical unit to another) set of data which need to be arranged in **increasing order**. Considering the ordered set of observations $X_1, X_2, \dots, X_n$ with $X_1 < X_2 < \dots < X_n$ - the total quantity is T(X)=nM(X)

$P_{(k)} = k/n$ → $X_{(K)}$ represents $p_{(k)}$ of the observations

$X_{(K)}$ $q_k < p_k$ and $q_n=p_n=1$

$q_{(K)} = (X_1 + X_2 + \dots + X_K)/T(X)$ → the value $X_{(K)}$ represents $q_{(k)}$ of the quantity

With this method we can define a series of p-weights (p1,p2,p3,…) and q-weights (q1,q2,q3,…).

- In case of **perfectly equal distribution**, we would have $p_j=q_j$ for every j
- In case of **total concentration**, we would have $q_1=q_2=q_{n-1}=0$ (meaning that a single statistical unit have the total quantity)

Considering these two limit patterns and applying the normalizing pattern, we can derive the concentration ratio:

$$R = \frac{\sum(p_i - q_i)}{\sum p_i} = 1 - \frac{\sum q_i}{\sum p_i}$$

HETEROGENEITY

When dealing with qualitative data it is also possible to measure heterogeneity (which could be considered as an extension of the concept of variability). We can define an **INDEX OF HETEROGENEITY** based on the complementary value of the concentration ratio:

$$H_R = 1 - R = \frac{\sum q_i}{\sum p_i}$$

Considering a quantitative variable with k possible modalities, collect a sample of n values reported in a frequency table:

- **PERFECT HOMOGENEITY** → every statistical unit have the same modality → R=1
- **MAXIMUM HETEROGENEITY** → every modality has the same frequency → R=0

VARIABILITY

The tendency of a statistical variable to assume different values in different units. It can be measured in different ways. A suitable measure of variability (W) should:

- **W>0** (variability has no sign)
- If all the data are equal, W=0 (no variability at all)
- Moving positive quantity of variables from a smaller observation to a larger one, W value should never increase.
- A measure of variability should use the most possible quantity of information coming from the observed data.

MEASURES OF VARIABILITY

RANGE OF DATA = largest observation – smallest observation

MEAN DEVIATION = mean distance of the observations from a central value *k*

- Absolute difference → minimum reached using the median as the central value.
  - MEAN ABSOLUTE DEVIATION (simple) → $MAD(X) = \frac{\sum_{j=1}^{n} |x_j - Me(X)|}{n}$
- Squared difference → minimum reached using the arithmetic mean as the central value.
  - **STANDARD DEVIATION** (quadratic) → $S(X) = \sqrt{\sum_{i=1}^{m} [x_j - M(X)]^2 \cdot \frac{n_i}{n}} = \sqrt{\sum_{i=1}^{m} [x_j - M(X)]^2 \cdot f_i}$

VARIANCE

$$V(X) = \sum_{i=1}^{m} [x_j - M(X)]^2 \cdot f_i$$

Applying a linear transformation Z= a + bX to V(x) → $V(Z)= b^2V(X)$

DEVIATION [Dev(X)]

- It is the numerator of the variance.
- decomposition property when data are grouped.

$$Dev(X) = n \cdot V(X) = \sum_{i=1}^{m} [x_j - M(X)]^2 \cdot n_i$$

INDICES OF VARIABILITY

As it is possible to compare only values expressed in the same unit and the abovementioned variability measures are expressed in specific units of measurement, dimensionless measures of variability are needed.

1. Compare the observed variability with a mean value:
   - RELATIVE MAD → RMAD = MAD(X)/Me(X)
   - COEFFICIENT OF VARITION → CV = S(X)/M(X)
2. Compare the observed variability with extreme values (max/min)
   - NORMALIZED STANDARD DEVIATION → S*(X)=S(X)/max(X)
3. Compare two different measures of variability (not common)

HOW TO NORMALIZE (range 0-1) AN INDEX OF VARIABILITY

$$W^* = \frac{W - W_{min}}{W_{max} - W_{min}}$$

BIVARIATE STATISTICAL VARIABLES

X and Y are independent if every event related to the variable X is independent from each event related to Y

ASSOCIATION

The theoretical frequency table [composed of $n^*_{ij} = (n_{h0}\, n_{0i})/n$] reflects the situation of perfect independence, which is the basic reference for checking the degree of association between two variables by comparing the observed data with the theoretical ones through the most used measure of association:

$$\chi^2 = \sum_{h=1}^{r}\sum_{j=1}^{c}\frac{(n_{hj} - n^*_{hj})^2}{n^*_{hj}}$$

Normalized:

$$\chi^* = \sqrt{\frac{\chi^2}{max(\chi^2)}} = \sqrt{\frac{\chi^2}{n \cdot [\min(r,c) - 1]}}$$

Being χ^2 focused on frequencies, its calculation is possible for every kind of variable.

LINEAR REGRESSION

Given two quantitative variables X (independent) and Y (dependent) we can collect a set of bivariate data which can be represented in a cartesian diagram where each point represents an individual datum, and their totality is called SCATTER OF OBSERVATIONS.

The simple kind of curve is the straight-line $Y = a + bX$ with a = value at 0 and b = slope.

- $b > 0$ → the variables have a **direct relationship** (increasing line)
- $b < 0$ → the variables have an **inverse relationship** (decreasing line)

COVARIANCE

It is a measure of linear association (its value depends on the unit of measurement of both variables)

$$\text{Cov(X,Y)} = \frac{\sum_{i=1}^{n}[(x_i - M(X))\cdot(y_i - M(Y))]}{n} = \frac{\sum_{i=1}^{n} x_i \cdot y_i}{n} - M(X)\cdot M(Y)$$

Properties (linear transformations): Cov (a+b·X, c+d·Y) = b·d· Cov (X,Y)

REGRESSION COEFFICIENTS

(Determination and interpretation)

The criterion to pick the best line among the ones crossing the scatter of observations is to minimize the sum of the differences between observed and theoretical values of Y: $\sum_{i=1}^{n}(y_i - y_i^*)^2$ =min. under such condition:

- the interpolating line represents the linear relationship, and it is called least square regression line
- the values of the coefficients are perfectly determined:

$b = \frac{Cov(X,Y)}{V(X)}$ → represents the average variation of Y when X increases of one unit

$a = M(Y) - \frac{Cov(X,Y)}{V(X)} \cdot M(X) = M(Y) - b \cdot M(X)$ → theoretical average value of Y when X=0

LINEAR CORRELATION COEFFICIENT

It is a semi-normalized index (as it ranges among -1 and 1) giving an information about the degree of linearity of the relationship between the two variables, obtained by dividing covariance by its maximum.

Covariance is maximized when **Y = a + bX** is a perfect model → observed values lie on the line.

$$r = \frac{Cov(X,Y)}{\max Cov(X,Y)} = \frac{Cov(X,Y)}{S(X)\cdot S(Y)}$$

- **r = -1** → the regression line is decreasing, and the linear relation is perfect
- **r = 0** → no linear relation between the two variables
- **r = +1** → the regression line is increasing, and the linear relation is perfect

COUNTINUOUS RANDOM VARIABLES

GAUSSIAN (NORMAL) DISTRIBUTION - N(μ,σ)

Symmetric, unimodal distribution having two parameters: the mean and the standard deviation.

Probability distribution function: $f(y) = \frac{1}{\sigma\sqrt{2\pi}} exp\left[-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2\right]$

Properties:

- if it undergoes a linear transformation the result is also normally distributed
- the linear combination of k mutually independent gaussian r.v. is normally distributed
- standardization N(0,1): $\mathbf{Z} = \frac{Y-\mu}{\sigma} = \frac{1}{\sigma}Y - \frac{m}{\sigma} = \frac{Y}{\sigma} - \frac{m}{\sigma}$

CHI-SQUARED DISTRIBUTION χ_d^2

It is the result of a transformation $W=Z^2$ of a standard normal distribution Z. such variable has a decreasing pdf starting at y=0 which is the modal point. $E(Z^2)= 1$ $V(Z^2)=2$

Summing up a certain number d of independent r.v. having such distribution the result will be a χ^2 distribution with d degrees of freedom - χ_d^2. $E(\chi_d^2)= d$ and $V(\chi_d^2)=2d$

STUDENT'S T-DISTRIBUTION

If Z is a standard normal distribution and W is a chi-squared distribution with *d* degrees of freedom, independent from Z, the following transformed variable is called *t*-distribution with *d* degrees of freedom $T_d = \frac{Z}{\sqrt{\frac{W}{d}}}$

The shape of such distribution is always **symmetric with respect to the value 0**. It has a smaller density in the center and a larger density in the tails compared to *N(0,1)*

- 1 dof → non-finite expected value
- d.o.f > 2 → *E(T) = 0*
- d.o.f.> *3* → ***V(T) =*** $\frac{d}{d-2}$ the variance converges to 1 as the number of degrees of freedom increases
- d.o.f.> 150 → t-distribution is virtually equivalent to a *N(0,1)*

LAW OF LARGE NUMBERS

The frequency of observation of any event converges to its probability as the number of trials increases.

$$E(f_n) = p \qquad V(f_n) = \frac{p(1-p)}{n}$$

VARIANCE AND COVARIANCE

COV (X, Y) = E(XY) – E(X)E(Y)
V (X+Y) = V(X) + V(Y) + 2COV (X, Y)
V (X-Y) = V(X) + V(Y) - 2COV (X, Y)

PROBABILITY

RANDOM EXPERIMENT = trial of any nature whose result is not known before performing the experiment itself.
ELEMENTARY EVENT = any possible result of a random experiment (which can result to be true or false).
SPACE OF OUTCOMES = the whole set of elementary events of a random experiment

- Finite
- Countable = when the set of elementary events is a countable infinity
- continuous = when the set of elementary events is a continuous infinity

Set of elementary events = can be defined through the application of logical operations to elementary events (can be empty or even a single event).

LIMIT EVENTS

- $\varnothing$ = the empty set (IMPOSSIBLE EVENT) which is false for every possible event
- Ω = whole space of outcomes which true for every possible result

BOOLEAN OPERATIONS

$A \cup B$ Union → It is true if at least one of the events is true.
$A \cap B$ Intersection → It is true only if both events are true.
$A - B$ Difference → It is true if A is true and B is false (or A/B)

DE MORGAN LAWS (properties valid for every couple of events) $(A \cup B)^* = A^* \cap B^*$ & $(A \cap B)^* = A^* \cup B^*$

RELATIONSHIPS BETWEEN EVENTS

$A \cap B = \varnothing$ DISJOINT – this concept can be extended to more than two events, but we have to distinguish

- GLOBALLY = the overall intersection is impossible
- MUTUALLY = every couple of events is disjoint (obviously this one implies the former)

$A \cup B \cup C = \Omega$ EXHAUSTIVE – their union is the whole space of outcomes

$A \rightarrow B$ IMPLY - the truth of A implies the truth of B

$A \rightarrow B$ & $B \rightarrow A$ EQUIVALENT

AXIOMS OF PROBABILITY

1. $P(E) \geq 0$ - The probability of an event E is never negative
2. $P(\Omega) = 1$ - The probability of the whole space of outcomes is always equal to 1
3. $A \cap B = \varnothing \rightarrow P(A \cup B) = P(A) + P(B)$ - If the events A and B are disjoint, the probability of their union is equal to the sum of their single probabilities (it holds for a number n (finite and integer) whatsoever of events, if they are mutually disjoint: $A \cap B = \varnothing$, $A \cap C = \varnothing$, $B \cap C = \varnothing$ → $P(A \cup B \cup C) = P(A) + P(B) + P(C)$)

IMPLICATIONS

- If $P(E)=p \rightarrow P(E^*) = 1-p$
- $P(\varnothing) = 0 = P(\Omega^*)$
- $P(A-B) = P(A) - P(A \cap B)$
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Conditional probability

$\mathbf{P(B|A)} = \frac{P(A \cap B)}{P(A)}$ with A = conditioning event and B = final event

Product rule: $P(A \cap B) = P(A) \cdot P(B|A) = P(B) \cdot P(A|B)$

INDEPENDENT EVENTS

$$P(\boldsymbol{B}|\boldsymbol{A}) = \frac{\boldsymbol{P(A \cap B)}}{\boldsymbol{P(A)}} = \mathbf{P(B)} \rightarrow \boldsymbol{P(A \cap B) = P(A) \cdot P(B)}$$

Two events are considered independent when the occurrence of one of them does not modify the probability of the other one. Independence is symmetric.

- If **P(B|A) > P(B),** the events are positively correlated
- If **P(B|A) < P(B),** the events are negatively correlated.

COMBINATORICS

SIMPLE PERMUTATIONS	all elements distinguishable	$P_n = n!$	Number of simple permutations of n elements	✓ Order ✓ All elements X No repetition
	some elements identical	$P_{n;\, n1,\, \ldots,\, nk} = \frac{n!}{n_1! n_2! \ldots n_k!}$	Number of permutations of n elements some of which are identical (categories C of elements)	✓ Order ✓ All elements ✓ Some identical (categories)
PARTIAL PERMUTATIONS	WITHOUT REPETITION	$PP_{n,k} = \frac{n!}{(n-k)!}$	We want to arrange in order a part k of n elements	✓ Order ✓ K elements out of n X repetition
	WITH REPETITION	$PP^{(R)}_{n,k} = n^k$	If we admit that some elements can be admitted more than once	✓ Order ✓ k elements ✓ with repetition
COMBINATIONS		$C_{n,k} = \frac{PP_{n,k}}{k!} = \frac{n!}{k!(n-k)!} = \binom{n}{k}$	We draw a subset of k elements from a population of n elements regardless of the order of extraction	X order ✓ k elements

STATISTICAL INFERENCE

Statistical inference consists in the joint use of explanatory statistics and probability to extend sample results to the whole population. It is composed of two main procedures: parameter estimation and hypothesis testing.

- POPULATION [P_N] = a complete set of statistical units
- SAMPLE = subset of the population which entirely observable
- SAMPLE SIZE [n] = number of sample observations
- SAMPLE SPACE = the set of all possible samples of size n that can be drawn from the population using a specific sample criterion.

Given a sample it is possible to calculate a sample statistic, if we apply the latter to all the samples of the sample space, we'll get the SAMPLE DISTRIBUTION associated to the statistics t.

The object of the **PARAMETER ESTIMATION** is to evaluate the overall value taken by one or more parameters in a population, by analysing a sample of such population. A generic parameter can take values in a specific interval [PARAMETRIC SPACE].

ESTIMATOR = a sample statistic which takes only values belonging to the parametric space

POINT ESTIMATION = number given by an estimator in a specific sample

INDICATORS OF PERFORMANCE OF THE ESTIMATORS = tools coming from the analysis of the sample distribution T.

- BIAS: → $B(T) = E(T) - \eta$
  - $B(T) > 0$ → overestimation
  - $B(T) < 0$ → underestimation
  - $B(T) = 0$ → the estimator is unbiased
- MEAN SQARE ERROR [efficiency]: $MSE(T) = E[T - \eta]^2 = V(T) + [B(T)]^2$ V(T) variance of the sample distribution
  the lower the value of MSE, the wider the efficiency.

SAMPLING THEORY

In statistical inference we need to collect, as far as we can, a representative sample, which should have a statistical behaviour similar to the population. Usually a sample can be considered «a good one» if it is ***totally random*** and ***sufficiently large*** (an inference based on a small sample is not efficient: confidence intervals would be too wide, and statistical test procedures would have a reduced power).

Once defined the population P_N (the population size N may be known or not), chosen the sample criterion C, and fixed the sample size n, we can define the sample space $U\ (P_N, C, n)$ and refer to it when performing the statistical inference.

If we choose a sampling criterion and decide to apply a certain estimator for a specific parameter, we have defined a sampling strategy.

Usually the population parameters to be estimated are the population total T_Y, mean m_Y, and frequency q_X. In the last context, we suppose to work with a population divided in two distinct categories.

When analysing sampling methods, two «new» probabilities are involved:

- the drawing probability of a sample, which is the probability of a specific sample to be drawn
- the selection probability of a statistical unit (the probability of a s.u. to be included in the sample). The selection probability may be defined for more than one unit. We will indicate with p_j the selection probability of the unit u_j.

SIMPLE RANDOM SAMPLING [S.R.S.]

SAMPLING METHOD	EXTRACTIONS	SAMPLE SPACE	SELECTION PROBABILITY OF A SINGLE STATISTICAL UNIT u_i
BERNOULLI SAMPLE WITH REPLACEMENT	**independent**, since the replacement reports the same probabilistic pattern at every extraction	N^n (partial permutations with repetition).	$p_i = 1 - \left(\frac{N-1}{N}\right)^n$
BLOCK SAMPLE WITHOUT REPLACEMENT	**exchangeable**: the probability of selecting u_i and then u_j is the same of the probability of selecting u_j and then u_i.	$\binom{N}{n} = \frac{N!}{n!\,(N-n)!}$ (simple combinations)	$p_i = = \frac{n}{N}$ *(sampling fraction)*

Once calculated the selection probabilities, we have to choose the estimator to be used for defining the sampling strategy. We will try to define unbiased parameter estimators.

For population total T_Y	For population mean m_Y	For population frequency q_Y
$\hat{T}_Y = N \cdot \bar{y}$ "Expansion estimator"	$\hat{\mu}_Y = \bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$ (sample average)	$\hat{\theta}_Y = f = \frac{\sum_{i=1}^{n} x_i}{n}$

The sample average and the sample frequency have all the properties we have already studied when introducing statistical inference.

DIFFERENCE BETWEEN STRATUM AND CLUSTER

STRATUM	Reduced level of variability within the classes	High level of variability between the classes	ALL the classes need to be included in the sample	Units belonging to distinct strata have a different distribution of the variable Y

CLUSTER	Reduced level of variability between the classes	High level of variability within the classes	only some of the classes is included in the sample	Units belonging to distinct strata have approximately the same distribution of Y

STRATIFIED SAMPLING

Sometimes we have the opportunity of classifying the units of a population in k classes *(k>2)*, such that there is a strong differentiation between the mean level of the variable Y among these classes, and a reduced variability within each class. These classes (sub-populations) are called strata and each of them has to be represented in the final sample.

We denote by N_h the number of population units belonging to the h-th stratum, and with n_h the corresponding sample size. Once fixed the sample sizes, a SRS without replacement has to be applied to every class.

When the sampling fraction $f_{s,h} = n_h/N_h$ is the same for every stratum (class), the criterium is called stratified proportional sampling. A single sample unit is denoted by u_{hi} (*h* is the stratum, *i* is the unit).

The sampling probability p_{hi} of a single unit u_{hi} depends on its stratum, and it can be derived by considering the possible combinations of all the other units, divided by the total number of samples:

$$\pi_{hi} = \frac{\binom{N_h-1}{n_h-1}}{\binom{N_h}{n_h}} = \frac{\frac{(N_h-1)!}{(n_h-1)!(N_h-n_h)!}}{\frac{N_h!}{n_h!\cdot(N_h-n_h)!}} = \frac{\frac{(N_h-1)!}{(n_h-1)!}}{\frac{N_h!}{n_h!}} = \frac{n_h}{N_h}$$

Second-order (joint) sampling probabilities

two units u_{hi} and u_{hj} belonging to the same class	$\pi_{hi,\,hj} = \frac{n_h}{N_h} \cdot \frac{n_h-1}{N_h-1}$
two units u_{hi} and u_{lj} belonging respectively to the h-th and the l-th class	$\pi_{hi,\,lj} = \pi_{hi} \cdot \pi_{lj} = \frac{n_h}{N_h} \cdot \frac{n_l}{N_l}$

Specific estimators:

PARAMETER	T_Y (total)	m_Y (mean)	q_X (frequency)
UNBIASED (CENTERED) ESTIMATOR	$\hat{T}_Y = \sum_{h=1}^{k} N_h \cdot \bar{y}_h$	$\hat{M}_Y = \frac{\hat{T}_Y}{N} = \sum_{h=1}^{k} w_h \cdot \bar{y}_h$	$\hat{\theta}_X = \sum_{h=1}^{k} \frac{N_h}{N} f_h = \sum_{h=1}^{k} w_h \cdot f_h$
EXPLANATION	Estimator defined by estimating the total T_{Yh} within each stratum, applying the expansion estimator, and summing for all the strata	$w_h = \frac{N_h}{N}$ represents the numeric weight of the *h*-th stratum	

CLUSTER SAMPLING

If we classify the units of a population in *k* classes *(k>2)*, called clusters, such that almost all the variability lies within the classes and there is only a minimal random differentiation among the classes, we can apply another

sampling method which is called cluster sampling. Each cluster is, unlike a stratum, a representative sample of the population. Only a small number of classes will be included in the sample.

Cluster sampling is essentially a method for selecting a reduced random sample from a high sized population.

There are two different typologies of cluster sampling: one-stage and two-stage cluster sampling.

CLUSTER SAMPLING	PROCEDURE	SAMPLING PROBABILITY of a single generic unit u_{hi}
ONE-STAGE	we have to choose at random a number *g* of clusters, out of the K clusters of the population (g < K), and include in the sample all the units of the selected clusters	equivalent to the probability of selecting its cluster. $\pi_{hi} = \frac{g}{K}$
TWO-STAGE SAME CLUSTER		$\pi_{hi,\,hj} = \frac{g}{K}$
TWO-STAGE DIFFERENT CLUSTERS		$\pi_{hi,\,lj} = \frac{g}{K} \cdot \frac{g-1}{K-1}$

In one-stage cluster sampling, we have to choose at random a number *g* of clusters, out of the K clusters of the population (g < K), and include in the sample all the units of the selected clusters. The sampling probability of a single generic unit u_{hi} is equivalent to the probability of selecting its cluster, and it is:

- $\pi_{hi} = \frac{g}{K}$ (4.12)

If we need to calculate the second-order sampling probabilities, we have to distinguish two situations: the two units may belong to the same cluster, or not. We have, respectively:

$\pi_{hi,\,hj} = \frac{g}{K}$ (same cluster) ; $\pi_{hi,\,lj} = \frac{g}{K} \cdot \frac{g-1}{K-1}$ (different clusters)

The most intuitive estimator of the population total T_Y can be defined applying the expansion estimator to the clusters:

$$\hat{T}_{Y1} = \frac{K}{g}\, t\boldsymbol{Y} = \frac{K}{g}\, n \cdot \bar{y} \qquad (4.14)$$

Example. Suppose that the population has been divided in 20 clusters, and we decide to select 7 clusters.

The estimator will then be: $\hat{T}_Y = \frac{20}{7}\, t\boldsymbol{Y} = \frac{20}{7}\, n \cdot \bar{y}$

However, if we know (or can estimate) the whole population size N, we can apply the «classic» expansion estimator:

- $\hat{T}_{Y1} = \frac{N}{n}\, t\boldsymbol{Y} = \mathrm{N} \cdot \bar{y}$ (4.15)

Since the clusters are essentially similar, with respect to the variable Y, it is not so important to calculate their mean values and frequencies separately, so we can use, as an overall estimator, respectively the overall sample average and the overall sample frequency:

$$\bar{\mu}_{Y1} = \bar{y} = \frac{\sum_{h=1}^{g} \sum_{j=1}^{N_h} y_{hj}}{n} \qquad (4.16)$$

$$\hat{\theta}_{X1} = f = \frac{\sum_{h=1}^{g} \sum_{j=1}^{N_h} x_{hj}}{n} = \frac{number\ of\ \{x_{hj}=1\}}{n} \qquad (4.17)$$

In two-stage cluster sampling, we have to select at random a number g of clusters C_1, C_2, …, C_g, out of the K clusters of the population, and then draw a SRS without replacement within each selected cluster.

The sampling probability of a single generic unit u_{hi} is equivalent to the probability of selecting its cluster, multiplied by the sampling fraction of the cluster:

- $\pi_{hi} = \frac{g}{K} \cdot \frac{n_h}{N_h}$ (4.18)

If we need to calculate the second-order sampling probabilities, we have to distinguish two situations: the two units may belong to the same cluster, or not. We have, respectively:

- $\pi_{hi,\,hj} = \frac{g}{K} \cdot \frac{n_h}{N_h} \cdot \frac{n_h - 1}{N_{h-1}}$ (two units of the same cluster) (4.19)
- $\pi_{hi,\,lj} = \frac{g}{K} \cdot \frac{g-1}{K-1} \cdot \frac{n_h}{N_h} \cdot \frac{n_l}{N_l}$ (two units of different clusters) (4.20)

If we want to estimate the total T_Y, we can apply the expansion estimator twice: within the cluster and between the clusters, having this result:

- $\hat{T}_{Y2} = \frac{K}{g} \sum_{h=1}^{g} \hat{T}_h = \frac{K}{g} \sum_{h=1}^{g} \frac{N_h}{n_h} \cdot \bar{y}_h = \frac{K}{g} \sum_{h=1}^{g} N_h \cdot \bar{y}_h$ (4.21)

Even in two-stage cluster sampling we can estimate the population mean m directly by using the overall sample average (4.16) and the population frequency q directly with the overall sample frequency (4.17). Finally, it is even possible to combine stratified and cluster sampling methods, and to define a structured sampling plan, which consists in applying a stratification on the population, to select a first sample and then to reduce it by applying a two-stage cluster sampling.

For example, if we need to select a sample of American families for an income (or occupational) survey, we can draw a stratified sample, using Federal States as strata, and then select – for each State – a sample of townships (first stage) and finally a sample of families for each township (second stage).