ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΑΤΡΩΝ
UNIVERSITY OF PATRAS

ΣΧΟΛΗ ΟΙΚΟΝΟΜΙΚΩΝ ΕΠΙΣΤΗΜΩΝ ΚΑΙ ΔΙΟΙΚΗΣΗΣ ΕΠΙΧΕΙΡΗΣΕΩΝ

ΤΜΗΜΑ ΔΙΟΙΚΗΣΗΣ ΤΟΥΡΙΣΜΟΥ

(πρώην Τμήμα Λογιστικής & Χρηματοοικονομικής – Μεσολόγγι)

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

Μέθοδοι Δειγματοληψίας

ΘΩΜΕΑΣ ΔΗΜΗΤΡΙΟΣ ΑΜ: 15314

ΕΠΟΠΤΕΥΩΝ ΚΑΘΗΓΗΤΗΣ: ΚΑΡΥΩΤΗ ΒΑΣΙΛΙΚΗ

ΜΕΣΟΛΟΓΓΙ 2022

ΠΡΟΛΟΓΟΣ

Όταν κάποιος θέλει να κάνει μια έρευνα για μια ομάδα ανθρώπων, δεν είναι δυνατό να συλλέξει δεδομένα από κάθε άτομο αυτής της ομάδας. Αντίθετα, ο ερευνητής επιλέγει ένα δείγμα. Το δείγμα είναι μια ομάδα ατόμων που θα συμμετάσχουν στην έρευνα.

Για να εξαχθούν έγκυρα συμπεράσματα από τα αποτελέσματα, ο ερευνητής πρέπει να αποφασίσει προσεκτικά πώς να επιλέξει ένα δείγμα που είναι αντιπροσωπευτικό του πληθυσμού. Υπάρχουν δύο τύποι μεθόδων δειγματοληψίας:

Η **δειγματοληψία πιθανοτήτων** περιλαμβάνει τυχαία επιλογή, επιτρέποντας την εξαγωγή ισχυρών στατιστικών συμπερασμάτων για τον πληθυσμό. Η **δειγματοληψία μη πιθανοτήτων** περιλαμβάνει μη τυχαία επιλογή με βάση την ευκολία ή άλλα κριτήρια, επιτρέποντας την εύκολη συλλογή δεδομένων. Πρέπει να είναι ξεκάθαρος ο τρόπος με τον οποίο ο ερευνητής συλλέγει τα δεδομένα.

Πρώτον, είναι απαραίτητο να κατανοήσουμε τη διαφορά μεταξύ ενός πληθυσμού και ενός δείγματος και να προσδιορίσουμε τον πληθυσμό-στόχο της έρευνας. Ο πληθυσμός είναι ολόκληρη η ομάδα για την οποία θέλουμε να εξάγουμε συμπεράσματα. Το δείγμα είναι η συγκεκριμένη ομάδα ατόμων από την οποία θα συλλέξουμε δεδομένα.

Ο πληθυσμός μπορεί να οριστεί ως προς τη γεωγραφική θέση, την ηλικία, το εισόδημα και πολλά άλλα χαρακτηριστικά. Είναι σημαντικό να καθοριστεί προσεκτικά ο πληθυσμός-στόχος σύμφωνα με το σκοπό και τις πρακτικές δυνατότητες του έργου. Εάν ο πληθυσμός είναι πολύ μεγάλος, δημογραφικά μεικτός και γεωγραφικά διασκορπισμένος, μπορεί να είναι δύσκολο να αποκτήσουμε πρόσβαση σε ένα αντιπροσωπευτικό δείγμα.

Στη παρούσα διπλωματική εργασία μελετώνται όλα τα παραπάνω.

ΠΕΡΙΛΗΨΗ

Πλαίσιο δειγματοληψίας

Το πλαίσιο δειγματοληψίας είναι η πραγματική λίστα των ατόμων από τα οποία θα ληφθεί το δείγμα. Στην ιδανική περίπτωση, θα πρέπει να περιλαμβάνει ολόκληρο τον πληθυσμό-στόχο.

Το μέγεθος του δείγματος

Ο αριθμός των ατόμων που πρέπει να συμπεριλάβουμε στο δείγμα εξαρτάται από διάφορους παράγοντες, συμπεριλαμβανομένου του μεγέθους και της μεταβλητότητας του πληθυσμού και του σχεδιασμού της έρευνας. Υπάρχουν διαφορετικοί τρόποι μεγέθους δείγματος και τύποι ανάλογα με το τι θέλουμε να επιτύχουμε στη στατιστική ανάλυση.

Μέθοδοι δειγματοληψίας πιθανοτήτων

Η δειγματοληψία πιθανοτήτων σημαίνει ότι κάθε μέλος του πληθυσμού έχει την ευκαιρία να επιλεγεί. Χρησιμοποιείται κυρίως στην ποσοτική έρευνα. Εάν θέλουμε να παράγουμε αποτελέσματα που είναι αντιπροσωπευτικά για ολόκληρο τον πληθυσμό, οι τεχνικές δειγματοληψίας πιθανοτήτων να είναι η πιο έγκυρη επιλογή.

Υπάρχουν τέσσερις κύριοι τύποι δειγμάτων πιθανοτήτων.

1. Απλή τυχαία δειγματοληψία

Σε ένα απλό τυχαίο δείγμα, κάθε μέλος του πληθυσμού έχει ίσες πιθανότητες να επιλεγεί. Το πλαίσιο δειγματοληψίας πρέπει να περιλαμβάνει ολόκληρο τον πληθυσμό.

Για τη διεξαγωγή αυτού του τύπου δειγματοληψίας, μπορούμε να χρησιμοποιήσουμε εργαλεία όπως γεννήτριες τυχαίων αριθμών ή άλλες τεχνικές που βασίζονται αποκλειστικά στην τύχη.

2. Συστηματική δειγματοληψία

Η συστηματική δειγματοληψία είναι παρόμοια με την απλή τυχαία δειγματοληψία, αλλά συνήθως είναι ελαφρώς ευκολότερη. Κάθε μέλος του πληθυσμού αναφέρεται με έναν αριθμό, αλλά αντί να δημιουργούνται τυχαία αριθμοί, τα άτομα επιλέγονται σε τακτά χρονικά διαστήματα.

3. Στρωματοποιημένη δειγματοληψία

Η στρωματοποιημένη δειγματοληψία περιλαμβάνει τη διαίρεση του πληθυσμού σε υποπληθυσμούς που μπορεί να διαφέρουν κατά σημαντικούς τρόπους. Μας επιτρέπει να εξάγουμε πιο ακριβή συμπεράσματα διασφαλίζοντας ότι κάθε υποομάδα αντιπροσωπεύεται σωστά στο δείγμα.

Για να χρησιμοποιήσουμε αυτήν τη μέθοδο δειγματοληψίας, διαιρούμε τον πληθυσμό σε υποομάδες (που ονομάζονται στρώματα) με βάση το σχετικό χαρακτηριστικό (φύλο, ηλικία, εισόδημα, εργασία).

4. Δειγματοληψία σε ομάδες

Η δειγματοληψία σε ομάδες περιλαμβάνει επίσης τη διαίρεση του πληθυσμού σε υποομάδες, αλλά κάθε υποομάδα πρέπει να έχει παρόμοια χαρακτηριστικά με ολόκληρο το δείγμα. Αντί να κάνουμε δειγματοληψία ατόμων από κάθε υποομάδα, επιλέγουμε τυχαία ολόκληρες υποομάδες.

Εάν είναι πρακτικά δυνατό, μπορούμε να συμπεριλάβουμε κάθε άτομο από κάθε σύμπλεγμα δειγματοληψίας. Εάν τα ίδια τα συμπλέγματα είναι μεγάλα, μπορούμε επίσης να δοκιμάσουμε

άτομα μέσα από κάθε σύμπλεγμα χρησιμοποιώντας μία από τις παραπάνω τεχνικές. Αυτό ονομάζεται δειγματοληψία πολλαπλών σταδίων.

Αυτή η μέθοδος είναι καλή για την αντιμετώπιση μεγάλων και διασκορπισμένων πληθυσμών, αλλά υπάρχει μεγαλύτερος κίνδυνος λάθους στο δείγμα, καθώς θα μπορούσαν να υπάρχουν ουσιαστικές διαφορές μεταξύ των συστάδων. Είναι δύσκολο να εγγυηθούμε ότι οι συστάδες του δείγματος είναι πραγματικά αντιπροσωπευτικές για ολόκληρο τον πληθυσμό.

Μέθοδοι δειγματοληψίας μη πιθανοτήτων

Σε ένα δείγμα μη πιθανοτήτων, τα άτομα επιλέγονται με βάση μη τυχαία κριτήρια και δεν έχει κάθε άτομο πιθανότητα να συμπεριληφθεί.

Αυτός ο τύπος δείγματος είναι ευκολότερος και φθηνότερος στην πρόσβαση, αλλά έχει υψηλότερο κίνδυνο μεροληψίας δειγματοληψίας. Αυτό σημαίνει ότι τα συμπεράσματα που μπορούμε να κάνουμε για τον πληθυσμό είναι πιο αδύναμα από ό,τι με τα δείγματα πιθανοτήτων και τα συμπεράσματά μας μπορεί να είναι πιο περιορισμένα. Εάν χρησιμοποιούμε ένα μη πιθανό δείγμα, θα πρέπει να προσπαθήσουμε να το κάνουμε όσο το δυνατόν πιο αντιπροσωπευτικό του πληθυσμού.

Οι τεχνικές δειγματοληψίας μη πιθανοτήτων χρησιμοποιούνται συχνά σε διερευνητική και ποιοτική έρευνα. Σε αυτούς τους τύπους έρευνας, ο στόχος δεν είναι να δοκιμαστεί μια υπόθεση σχετικά με έναν ευρύ πληθυσμό, αλλά να αναπτυχθεί μια αρχική κατανόηση ενός μικρού ή υπο-ερευνημένου πληθυσμού.

1. Βολική δειγματοληψία

Ένα δείγμα ευκολίας περιλαμβάνει απλώς τα άτομα που τυχαίνει να είναι πιο προσιτά στον ερευνητή.

Αυτός είναι ένας εύκολος και φθηνός τρόπος για να φτάσουμε τα αρχικά της δεδομένα, αλλά δεν υπάρχει τρόπος να διαπιστωθεί εάν το δείγμα είναι αντιπροσωπευτικό του πληθυσμού, επομένως δεν μπορεί να παράγει γενικεύσιμα αποτελέσματα.

2. Δειγματοληψία εθελοντικής απάντησης

Παρόμοια με ένα δείγμα ευκολίας, ένα δείγμα εθελοντικής απόκρισης βασίζεται κυρίως στην ευκολία πρόσβασης. Αντί ο ερευνητής να επιλέξει τους συμμετέχοντες και να επικοινωνήσει απευθείας μαζί τους, οι άνθρωποι προσφέρονται εθελοντικά (π.χ. απαντώντας σε μια δημόσια διαδικτυακή έρευνα).

Τα δείγματα εθελοντικής απόκρισης είναι πάντα τουλάχιστον κάπως προκατειλημμένα, καθώς ορισμένα άτομα είναι εγγενώς πιο πιθανό να προσφερθούν εθελοντικά από άλλα.

3. Σκόπιμη δειγματοληψία

Αυτός ο τύπος δειγματοληψίας, γνωστός και ως δειγματοληψία κρίσης, περιλαμβάνει τον ερευνητή χρησιμοποιώντας την πείρα του για να επιλέξει ένα δείγμα που είναι πιο χρήσιμο για τους σκοπούς της έρευνας.

Συχνά χρησιμοποιείται στην ποιοτική έρευνα, όπου ο ερευνητής θέλει να αποκτήσει λεπτομερείς γνώσεις για ένα συγκεκριμένο φαινόμενο αντί να κάνει στατιστικά συμπεράσματα ή όπου ο πληθυσμός είναι πολύ μικρός και συγκεκριμένος. Ένα αποτελεσματικό στοχευμένο δείγμα πρέπει να έχει σαφή κριτήρια και λογική συμπερίληψης.

4. Δειγματοληψία χιονόμπαλας

Εάν ο πληθυσμός είναι δύσκολο να προσπελαστεί, η δειγματοληψία χιονόμπαλας μπορεί να χρησιμοποιηθεί για τη στρατολόγηση συμμετεχόντων μέσω άλλων συμμετεχόντων. Ο αριθμός των ατόμων στους οποίους έχουμε πρόσβαση σε «χιονόμπαλες» καθώς ερχόμαστε σε επαφή με περισσότερα άτομα.

Λέξεις κλειδιά: μέθοδοι δειγματοληψίας, στατιστική, πιθανότητες

PROLOGUE

When someone wants to make a research about a group of people, it's not possible to collect data from every person in that group. Instead, the researcher selects a sample. The sample is a group of individuals who will participate in the research.

To draw valid conclusions from the results, the researcher has to carefully decide how to select a sample that is representative of the population. There are two types of sampling methods:

Probability sampling involves random selection, allowing to make strong statistical inferences about the population.

Non-probability sampling involves non-random selection based on convenience or other criteria, allowing to easily collect data.

It needs to be clear the way that the researcher collects the data.

First, it is necessary to understand the difference between a population and a sample, and identify the target population of the research.

The **population** is the entire group that you want to draw conclusions about.

The **sample** is the specific group of individuals that you will collect data from.

The population can be defined in terms of geographical location, age, income, and many other characteristics.

It is important to carefully define the target population according to the purpose and practicalities of the project.

If the population is very large, demographically mixed, and geographically dispersed, it might be difficult to gain access to a representative sample.

SUMMARY

Sampling frame

The sampling frame is the actual list of individuals that the sample will be drawn from. Ideally, it should include the entire target population.

Sample size

The number of individuals you should include in your sample depends on various factors, including the size and variability of the population and your research design. There are different sample size calculators and formulas depending on what you want to achieve with statistical analysis.

Probability sampling methods

Probability sampling means that every member of the population has a chance of being selected. It is mainly used in quantitative research. If you want to produce results that are representative of the whole population, probability sampling techniques are the most valid choice.

There are four main types of probability sample.

1. Simple random sampling

In a simple random sample, every member of the population has an equal chance of being selected. Your sampling frame should include the whole population.

To conduct this type of sampling, you can use tools like random number generators or other techniques that are based entirely on chance.

2. Systematic sampling

Systematic sampling is similar to simple random sampling, but it is usually slightly easier to conduct. Every member of the population is listed with a number, but instead of randomly generating numbers, individuals are chosen at regular intervals.

3. Stratified sampling

Stratified sampling involves dividing the population into subpopulations that may differ in important ways. It allows you draw more precise conclusions by ensuring that every subgroup is properly represented in the sample.

To use this sampling method, you divide the population into subgroups (called strata) based on the relevant characteristic (gender, age, income, job).

4. Cluster sampling

Cluster sampling also involves dividing the population into subgroups, but each subgroup should have similar characteristics to the whole sample. Instead of sampling individuals from each subgroup, you randomly select entire subgroups.

If it is practically possible, you might include every individual from each sampled cluster. If the clusters themselves are large, you can also sample individuals from within each cluster using one of the techniques above. This is called multistage sampling.

This method is good for dealing with large and dispersed populations, but there is more risk of error in the sample, as there could be substantial differences between clusters. It's difficult to guarantee that the sampled clusters are really representative of the whole population.

Non-probability sampling methods

In a non-probability sample, individuals are selected based on non-random criteria, and not every individual has a chance of being included.

This type of sample is easier and cheaper to access, but it has a higher risk of sampling bias. That means the inferences you can make about the population are weaker than with probability samples, and your conclusions may be more limited. If you use a non-probability sample, you should still aim to make it as representative of the population as possible.

Non-probability sampling techniques are often used in exploratory and qualitative research. In these types of research, the aim is not to test a hypothesis about a broad population, but to develop an initial understanding of a small or under-researched population.

1. Convenience sampling

A convenience sample simply includes the individuals who happen to be most accessible to the researcher.

This is an easy and inexpensive way to gather initial data, but there is no way to tell if the sample is representative of the population, so it can't produce generalizable results.

2. Voluntary response sampling

Similar to a convenience sample, a voluntary response sample is mainly based on ease of access. Instead of the researcher choosing participants and directly contacting them, people volunteer themselves (e.g. by responding to a public online survey).

Voluntary response samples are always at least somewhat biased, as some people will inherently be more likely to volunteer than others.

3. Purposive sampling

This type of sampling, also known as judgement sampling, involves the researcher using their expertise to select a sample that is most useful to the purposes of the research.

It is often used in qualitative research, where the researcher wants to gain detailed knowledge about a specific phenomenon rather than make statistical inferences, or where the population is very small and specific. An effective purposive sample must have clear criteria and rationale for inclusion.

4. Snowball sampling

If the population is hard to access, snowball sampling can be used to recruit participants via other participants. The number of people you have access to "snowballs" as you get in contact with more people.

Key words: sampling methods, statistics, probabilities

Περιεχόμενα

ΚΕΦΑΛΑΙΟ 1

Εισαγωγή

Η συγκέντρωση στατιστικών δεδομένων είναι μια από τις σημαντικότερες ενέργειες που πρέπει να κάνει ο ερευνητής όταν θέλει να μελετήσει στατιστικά ένα φαινόμενο. Πριν ξεκινήσει η στατιστική έρευνα οφείλουν, οι ερευνητές, να ορίσουν με ακρίβεια το σύνολο που θα μελετήσουν, δηλαδή, τον στατιστικό πληθυσμό, καθώς και τις στατιστικές μονάδες που θα απαρτίζουν τον πληθυσμό. Στατιστική μονάδα μπορεί να θεωρηθεί ένα αντικείμενο, ένα άτομο, ένα νοικοκυριό κ.α. Δύο είναι οι μέθοδοι συγκέντρωσης στατιστικών στοιχείων:

- οι εξαντλητικές έρευνες (απογραφή), και
- οι δειγματοληπτικές έρευνες.

Δειγματοληψία είναι η απογραφή ορισμένων συγκεκριμένων χαρακτηριστικών ενός τμήματος του πληθυσμού. Το τμήμα του πληθυσμού που απογράφεται ονομάζεται δείγμα. Σκοπός των δειγματοληπτικών ερευνών είναι να προσδιορίσουμε όσο γίνεται ακριβέστερα ιδιότητες του πληθυσμού, μελετώντας απογραφικά τα στοιχεία του δείγματος. Γενικά, η δειγματοληψία θεωρείται επιτυχής όταν η επιλογή του δείγματος παράγει αποτελέσματα, δείκτες και μετρήσεις που είναι όσο το δυνατόν ακριβέστερα, δηλαδή βρίσκονται όσο πιο κοντά στις αντίστοιχες παραμέτρους του ευρύτερου συνόλου, δηλαδή του πληθυσμού. Η συνέπεια της επέκτασης των συμπερασμάτων που προέρχονται από τη μελέτη των χαρακτηριστικών του δείγματος, σε ολόκληρο τον πληθυσμό, εξαρτάται από τη μέθοδο δειγματοληψίας που εφαρμόζουμε, καθώς από τη ποιότητα του δείγματος εξαρτάται κατά πολύ η σημαντικότητα των εκτιμήσεων. Τέλος, οι εκτιμήσεις των δειγματοληψιών δεν δίνουν ακριβείς τιμές αλλά προσεγγίσεις για το σύνολο του πληθυσμού. (Thompson, S. K. 2012)

ΚΕΦΑΛΑΙΟ 2

Βασική στατιστική θεωρία

2.1 Βασικές έννοιες

Το σύνολο των παρατηρήσεων που συνδέονται με το φαινόμενο που θέλουμε να μελετήσουμε ονομάζεται **πληθυσμός** (population).

Σε αντίθεση με τον πληθυσμό που ασχολείται με όλες τις δυνατές παρατηρήσεις, το **δείγμα** (sample) περιέχει μόνο παρατηρήσεις κάποιες από τις. Είναι δηλαδή ένα υποσύνολο του πληθυσμού.

Υποθέτουμε ότι ο πληθυσμός για τον οποίο ενδιαφερόμαστε να βγάλουμε συμπεράσματα με τη βοήθεια ενός δείγματος είναι πεπερασμένος, και έστω ότι είναι το πλήθος των μελών του πληθυσμού. Το πλήθος αυτό ονομάζεται και **μέγεθος του πληθυσμού** (population size) και συμβολίζεται με ***N***. Ανάλογα, συμβολίζουμε και ονομάζουμε **μέγεθος δείγματος** (sample size) το πλήθος των μονάδων του πληθυσμού που επιλέγονται για το δείγμα, το οποίο συμβολίζεται με ***n*** **.**

2.2 Μεταβλητές και παράμετροι

Πληθυσμιακές ποσότητες

Χαρακτηριστικό ενός πληθυσμού (population characteristic) ονομάζεται το ερώτημα για τον πληθυσμό, το οποίο όμως ενδιαφέρει να εξετάσουμε. Το χαρακτηριστικό του πληθυσμού που μελετάμε, μπορεί να θεωρηθεί ως τυχαία μεταβλητή (τ.μ.) ***Y*** και Y_i είναι η τιμή του χαρακτηριστικού για το *i* μέλος του δείγματος. Οι δυνατές τιμές της τ.μ. Y είναι το σύνολο $\{Y_1, Y_2, \dots, Y_n\}$. .

Αν ***Y*** είναι το υπό μελέτη χαρακτηριστικό του πληθυσμού, τότε οι ποσότητες για τις οποίες πιο συχνά ενδιαφερόμαστε να εξαγάγουμε συμπεράσματα είναι συνήθως:

• *η μέση τιμή (mean value) του χαρακτηριστικού για τον πληθυσμό.*
Όταν ο υπό μελέτη πληθυσμός είναι πεπερασμένος, η μέση τιμή ορίζεται ως το άθροισμα των μετρήσεων του ***Y*** για όλα τα μέλη του πληθυσμού διαιρούμενο με το πλήθος όμως. Εάν συμβολίσουμε τη μέση τιμή του ***Y*** για τον πληθυσμό με μ ή $\bar{Y}$, τότε:

$$\mu = \bar{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i$$

• *η διασπορά (variance)* σ^2 *των τιμών του πληθυσμού γύρω από τη μέση τιμή.*

$$\sigma^2 = \frac{1}{N-1}\sum_{\iota=1}^{N}(Y_i - \bar{Y})^2$$

Η *διασπορά (variance)* σ^2 δίνει πληροφορία σχετικά με την ετερογένεια ή ομοιογένεια των μετρήσεων του χαρακτηριστικού στο σύνολο του πληθυσμού. Όσο μεγαλύτερη η διασπορά σ^2 για έναν πληθυσμό, τόσο πιο ετερογενής είναι ο πληθυσμός. Ο ρόλος της διασποράς του πληθυσμού για το χαρακτηριστικό παίζει ουσιαστικό ρόλο σχεδόν σε κάθε στάδιο της έρευνας: στον σχεδιασμό και την επιλογή της δειγματοληπτικής μεθόδου, στον καθορισμό του μεγέθους του δείγματος, και βέβαια στην ανάλυση των δεδομένων και τον υπολογισμό των ιδιοτήτων των εκτιμητών.

- *το ποσοστό (percentage) όμως χαρακτηριστικού στον πληθυσμό.*
  Το ποσοστό συμβολίζεται με **p** και ορίζεται ως:

$$p = \frac{A}{N}$$

όπου ***A*** είναι το πλήθος των μελών του πληθυσμού που ανήκουν στην υπό μελέτη κατηγορία.

Δειγματικές ποσότητες

Για τα δείγματα πιθανότητας, οι n παρατηρήσεις του δείγματος είναι τυχαίες μεταβλητές, γιατί εμπεριέχουν τον παράγοντα της τυχαιότητας. Η τυχαία μεταβλητή X_i , με $i = 1.2 \ldots . n$ είναι η μεταβλητή που καταγράφει την i −οστή μέτρηση του δείγματος. Οι δυνατές τιμές της είναι το σύνολο $\{Y_1, Y_2, \ldots, Y_n\}$. Τότε ορίζονται:

- η *δειγματική μέση τιμή (sample mean value)* η οποία συμβολίζεται με $\bar{X}$ και υπολογίζεται ως:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

- η *δειγματική διασπορά (sample variance)* των τιμών του δείγματος που συμβολίζεται με s^2.

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{x})^2$$

Η διασπορά s^2 υπολογίζεται αριθμητικά μετά την συλλογή και καταγραφή των μετρήσεων του δείγματος και η πληροφορία που προσφέρει είναι ανάλογη εκείνης της σ . Όσο μεγαλύτερη είναι η s , τόσο μεγαλύτερη ετερογένεια παρατηρείται στις μετρήσεις του δείγματος, και αντίστροφα.

- *το δειγματικό ποσοστό (percentage) όμως χαρακτηριστικού.*
  Το ποσοστό συμβολίζεται με $\hat{p}$ και ορίζεται ως:

$$\hat{p} = \frac{\alpha}{n}$$

Όπου α το πλήθος των μελών του δείγματος που ανήκουν στην υπό μελέτη κατηγορία.

2.3 Εκτίμηση παραμέτρων πεπερασμένων πληθυσμών

Ένας όρος που μας απασχολεί είναι η **στατιστική συνάρτηση (σ.σ.)** (statistical function). Στατιστική συνάρτηση ονομάζεται μια συνάρτηση των τ.μ. του δείγματος.

Εκτιμητής (estimator) $\hat{\theta}$ του θ ονομάζεται μια συνάρτηση των τυχαίων μεταβλητών του δείγματος, που χρησιμοποιείται με σκοπό την εκτίμηση του .

Σύμφωνα με τον ορισμό, θα είναι $\hat{\theta}(X_1 X_{2\ldots} X_n)$ και, κατά συνέπεια: (α) ο εκτιμητής είναι όμως μια τυχαία μεταβλητή ως συνάρτηση τυχαίων μεταβλητών και (β) είναι εφικτό να υπολογίσουμε την αριθμητική τιμή του αμέσως μετά τη διεξαγωγή της έρευνας, όταν θα είναι γνωστές οι αριθμητικές τιμές των X_i .

Σύμφωνα με τον ορισμό του εκτιμητή, ενας εκτιμητής είναι στατιστική συνάρτηση. Τα περισσότερα από τα συμπεράσματα στη θεωρία δειγματοληψίας βασίζονται στη **δειγματική κατανομή** (sampling distribution). Ως δειγματική κατανομή μιας στατιστικής συνάρτησης ορίζεται η κατανομή των τιμών της στατιστικής συνάρτησης που προκύπτουν, εάν γίνει εξάντληση όλων των δυνατών δειγμάτων σύμφωνα με τον τρόπο που ακολουθείται για την επιλογή του δείγματος.

Η **αναμενόμενη τιμή** (expected value) $t = t(X_1 X_2 \ldots X_n) = t(s)$ μιας σ.σ. υπολογίζεται λαμβάνοντας υπόψη όμως δυνατές τιμές όμως συνάρτησης $t = t(s)$, δηλαδή για όλα τα δυνατά δείγματα s και όμως αντίστοιχες πιθανότητες πραγματοποίησης του κάθε δείγματος, σύμφωνα με τον τρόπο δειγματοληψίας. Θα είναι συνεπώς:

$$E(t) = \sum_{s} p(s)t(s)$$

όπου ο δείκτης s του αθροίσματος λαμβάνει όλες τις δυνατές τιμές του δείγματος, $t(s)$ είναι η τιμή της σ.σ t υπολογισμένη για το δείγμα s , και $p(s)$ η πιθανότητα επιλογής του s.

Ο ορισμός του εκτιμητή είναι αρκετά γενικός, με αποτέλεσμα να επιτρέπει οποιαδήποτε συνάρτηση (οποιασδήποτε μορφής) να είναι θεωρητικά εκτιμητής της ποσότητας που μας ενδιαφέρει. Υπάρχει συνεπώς ανάγκη για αξιολόγηση και σύγκριση των εκτιμητών μεταξύ τους. Η αξιολόγηση γίνεται με βάση μια σειρά κριτηρίων, ορισμένα εκ των οποίων αποτελούν ταυτόχρονα και ιδιότητες των εκτιμητών. Παραθέτουμε στη συνέχεια τα πιο σημαντικά από αυτά.

Αμεροληψία (unbiasedness) Ένας εκτιμητής $\hat{\theta}$ θα λέγεται αμερόληπτος εκτιμητής μιας παραμέτρου θ εάν η αναμενόμενη τιμή του ισούται με θ, δηλ. αν $E(\hat{\theta}) = \theta$. Η αναμενόμενη τιμή του εκτιμητή υπολογίζεται από τον τύπο:

$$E(\mathrm{t}) = \sum_{s} p(s)t(s)$$

Το σύνολο των δυνατών δειγμάτων που προκύπτουν με μια μέθοδο δειγματοληψίας ονομάζεται **δειγματοληπτικός χώρος (sampling space)** και συμβολίζεται συνήθως με $\mathcal{S}$.

Το **ποσό μεροληψίας** (bias) ενος εκτιμητή $\hat{\theta}$, συμβολίζεται ως $bias(\hat{\theta})$ δίνεται από τη σχέση:

$$bias(\hat{\theta}) = E(\hat{\theta}) - \theta.$$

Εάν $bias(\hat{\theta}) > 0$ αυτό σημαίνει ότι ο εκτιμητής παρουσιάζει θετική μεροληψία, δηλ. για το δοθέν δειγματοληπτικό σχέδιο αναμένουμε η δειγματική κατανομή του $\hat{\theta}$ να έχει μέση τιμή μεγαλύτερη του και, κατά συνέπεια, ο εκτιμητής να υπερ-εκτιμά την παράμετρο.

Εάν $bias(\hat{\theta}) < 0$ έχουμε αρνητική μεροληψία ή ισοδύναμα ο εκτιμητής υπο-εκτιμά την παράμετρο.

Εάν $bias(\hat{\theta}) = 0$ ο εκτιμητής είναι αμερόληπτος.

Ένα άλλο κριτήριο σύγκρισης εκτιμητών είναι η **ακρίβεια (accuracy).** Η ακρίβεια ως ιδιότητα ενός εκτιμητή δίνει ένα μέτρο της συγκέντρωσης, ή, αντίθετα, της απόκλισης που παρουσιάζουν μεταξύ τους οι δυνατές τιμές του εκτιμητή. Όσο πιο πυκνά είναι οι δυνατές τιμές του εκτιμητή, όσο δηλαδή μεγαλύτερη είναι η συγκέντρωση, τόσο μεγαλύτερη είναι η ακρίβεια του εκτιμητή.

Ένα μέτρο της ακρίβειας του εκτιμητή είναι το **μέσο τετραγωνικό σφάλμα** (mean square error) που συμβολίζεται με MSE. Το μέσο τετραγωνικό σφάλμα ενός εκτιμητή $\widehat{\theta}$ ορίζεται ως η αναμενόμενη τιμή της τετραγωνικής απόκλισης του εκτιμητή από την προς εκτίμηση ποσότητα:

$$\mathrm{MSE}(\hat{\theta}) = \mathrm{E}(\hat{\theta} - \theta)^2$$

Η αναμενόμενη τιμή στον ορισμό του MSE υπολογίζεται από τη σχέση:

$$\mathrm{MSE}(\hat{\theta}) = \sum_{s \in \mathcal{S}} \pi(s) \left(\hat{\theta}(s) - \theta\right)^2$$

Ακριβής (accurate) είναι ένας εκτιμητής με μικρό μέσο τετραγωνικό σφάλμα.

Το MSE σχετίζεται με το γνωστό μέτρο της διακύμανσης ενός εκτιμητή. Η **διακύμανση** (variance) ενός εκτιμητή ορίζεται από τη σχέση:

$$Var(\hat{\theta}) = \mathrm{E}\big(\hat{\theta} - E(\hat{\theta})\big)^2$$

Εάν $E(\hat{\theta}) = \theta$, δηλαδή εάν ο εκτιμητής $\hat{\theta}$ είναι αμερόληπτος, τότε $MSE(\hat{\theta}) = Var(\hat{\theta})$. Γενικότερα, αποδεικνύεται ότι η σχέση που συνδέει το MSE όμως εκτιμητή και τη διακύμανσή του είναι:

$$MSE(\hat{\theta}) = \mathrm{Var}(\hat{\theta}) + \mathrm{bias}^2(\hat{\theta})$$

Μεταξύ όλων των αμερόληπτων εκτιμητών, ο εκτιμητής με την ελάχιστη διακύμανση λέγεται **αποτελεσματικός** (efficient). Αν $\hat{\theta}_1$ και $\hat{\theta}_2$ είναι δύο αμερόληπτοι εκτιμητές με $Var(\hat{\theta}_1) < Var(\hat{\theta}_2)$ τότε ο εκτιμητής $\hat{\theta}_1$ λέγεται σχετικά αποτελεσματικός.

Τυπικό σφάλμα (standard error) του εκτιμητή $\hat{\theta}$, συμβολικά $se(\hat{\theta})$, ονομάζεται η θετική τετραγωνική ρίζα όμως διακύμανσης του εκτιμητή, δηλ.

$$\mathrm{se}(\hat{\theta}) = \sqrt{\mathrm{Var}(\hat{\theta})}$$

Για τον υπολογισμό της αναμενόμενης τιμής, του μέσου τετραγωνικού σφάλματος ή της διακύμανσης ενός εκτιμητή λαμβάνεται υπόψη ο δειγματικός χώρος και οι πιθανότητες επιλογής των δειγμάτων. Συνεπώς, όλες οι παραπάνω ποσότητες εξαρτώνται από το δειγματοληπτικό σχέδιο που υιοθετήθηκε στην έρευνα.

Παρακάτω δίνονται όλες οι βασικές ιδιότητες της αναμενόμενης τιμής, όπως διακύμανσης και ποσοτήτων που ορίζονται με τη βοήθεια αυτών, όπως η **συμμεταβλητότητα** (covariance) και η **συσχέτιση** (correlation). Τα σύμβολα *X, Y*, και *Z* αντιπροσωπεύουν τυχαίες μεταβλητές, ενώ τα *a, b, c* πραγματικούς αριθμούς.

Αναμενόμενη Τιμή **1)** $E(a) = a$
2) $E(aX + bY) = aE(X) + bE(Y)$

Διακύμανση **1)** $Var(a) = 0$
2) $Var(aX) = a^2 Var(X)$
3) Αν *X, Y* ασυσχέτιστες τότε $Var(X \pm Y) = Var(X) + Var(Y)$

Συμμεταβλητότητα $Cov(X,Y) = E[(X - \bar{X})(Y - \bar{Y})] = E(XY) - E(X)E(Y)$
1) Αν $Cov(X,Y) = 0$ τότε X, Y ασυσχέτιστες
2) $Cov(aX, Y) = aCov(X,Y)$
3) $Cov(aX, bY) = abCov(X,Y)$
4) $Var(X + Y) = Var(X) + Var(Y) \pm 2Cov(X,Y)$

5) Ισχύει ότι:

$$Var\left(\sum_{i=1}^{N} X_i\right) = \sum_{i,j=1}^{N} Cov(X_i, X_j) = \sum_{i=1}^{N} VarX_i + \sum_{i \neq j} Cov(X_i, X_j)$$

Συσχέτιση **Ορισμός:** $\rho(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)\,Var(Y)}}$

$$-1 \leq \rho(X,Y) \leq 1$$

ΒΑΣΙΚΕΣ ΠΟΣΟΤΗΤΕΣ ΠΛΗΘΥΣΜΟΥ ΚΑΙ ΔΕΙΓΜΑΤΟΣ

	ΠΛΗΘΥΣΜΟΣ	ΔΕΙΓΜΑ
ΜΕΓΕΘΟΣ	N	n
ΜΕΣΗ ΤΙΜΗ	$\mu = \frac{1}{N}\sum_{i=1}^{N} Y_i$	$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} X_i$
ΔΙΑΣΠΟΡΑ	$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N} (Y_i - \mu)^2$	$s^2 = \frac{1}{n-1}\sum_{i=1}^{n} (X_i - \bar{x})^2$
ΤΥΠΙΚΗ ΑΠΟΚΛΙΣΗ	$\sigma = \sqrt{\sigma^2}$	$s = \sqrt{s^2}$
ΑΡΙΘΜΟΣ ΜΟΝΑΔΩΝ ΑΝΑ ΚΑΤΗΓΟΡΙΑ	A	α
ΠΟΣΟΣΤΟ ΜΟΝΑΔΩΝ ΣΕ ΜΙΑ ΚΑΤΗΓΟΡΙΑ	$p = \frac{A}{N}$	$\hat{p} = \frac{a}{n}$

ΚΕΦΑΛΑΙΟ 3

Μέθοδοι δειγματοληψίας

3.1 Δείγματα πιθανότητας και δείγματα μη πιθανότητας

- Ένα δείγμα λέγεται **δείγμα πιθανότητας** όταν η κάθε μονάδα του πληθυσμού έχει μια πιθανότητα, συγκεκριμένη και μη-μηδενική, να συμπεριληφθεί στο δείγμα. Η πιθανότητα αυτή είναι προκαθορισμένη πριν την επιλογή του δείγματος.
  Συνεπώς σύμφωνα με τα δείγματα πιθανότητας, η μέθοδος δειγματοληψίας δεν αποκλείει κάποιες μονάδες του πληθυσμού από το ενδεχόμενο να είναι μέρη του δείγματος. Επιπλέον η προκαθορισμένη και κατά συνέπεια γνωστή πιθανότητα επιλογής όμως κάθε μονάδας του πληθυσμού, συνεπάγεται ή εγγυάται ότι στη διαδικασία επιλογής του δείγματος υπεισέρχεται ο παράγοντας της τυχαιότητας.

  Π.χ. Αν ενδιαφερόμαστε για την ψήφο των δημοτών μιας πόλης και επιλέξουμε τυχαία 500 άτομα από το δημοτολόγιο όμως πόλης, τότε το δείγμα είναι ένα δείγμα πιθανότητας. Συγκεκριμένα, οι πιθανότητες επιλογής μεταξύ των μελών είναι ίσες για όλα τα μέλη του πληθυσμού.

- Ένα δείγμα λέγεται **δείγμα μη πιθανότητας** όταν η μέθοδος επιλογής της κάθε μονάδας του δείγματος δεν διέπεται από τους νόμους της πιθανότητας, αλλά βασίζεται σε κριτήρια όπως η ευκολία, η εύκολη πρόσβαση, η διαθεσιμότητα, ο χρόνος συλλογής των δεδομένων και άλλα. Τα κριτήρια αυτά δεν εξασφαλίζουν μια θετική και προκαθορισμένη πιθανότητα επιλογής όλων των μελών του πληθυσμού. Αντίθετα η επιλογή ή μη των μελών γίνεται με βεβαιότητα.

  Π.χ. Αν ενδιαφερόμαστε ξανά για την ψήφο των δημοτών μιας πόλης αλλά αυτή τη φορά επιλέξουμε να κάνουμε την έρευνα μέσω διαδικτύου, υπάρχει ο κίνδυνος μεροληψίας γιατί αυτομάτως αποκλείονται άτομα χωρίς πρόσβαση στο διαδίκτυο και άτομα που δεν πρόλαβαν το χρονικό πλαίσιο της έρευνας.

3.2 Τεχνικές δειγματοληψίας

Οι τεχνικές δειγματοληψίας διακρίνονται σε δύο κατηγορίες:

- τη δειγματοληψία με πιθανότητες ή αντιπροσωπευτική δειγματοληψία και
- τη δειγματοληψία χωρίς πιθανότητες ή δειγματοληψία κρίσης.

Η δειγματοληψία με πιθανότητα γίνεται σύμφωνα με τους νόμους των πιθανοτήτων, είναι ελεγχόμενη ως προς τις παραμέτρους και δίνει τη δυνατότητα να γενικευτούν τα

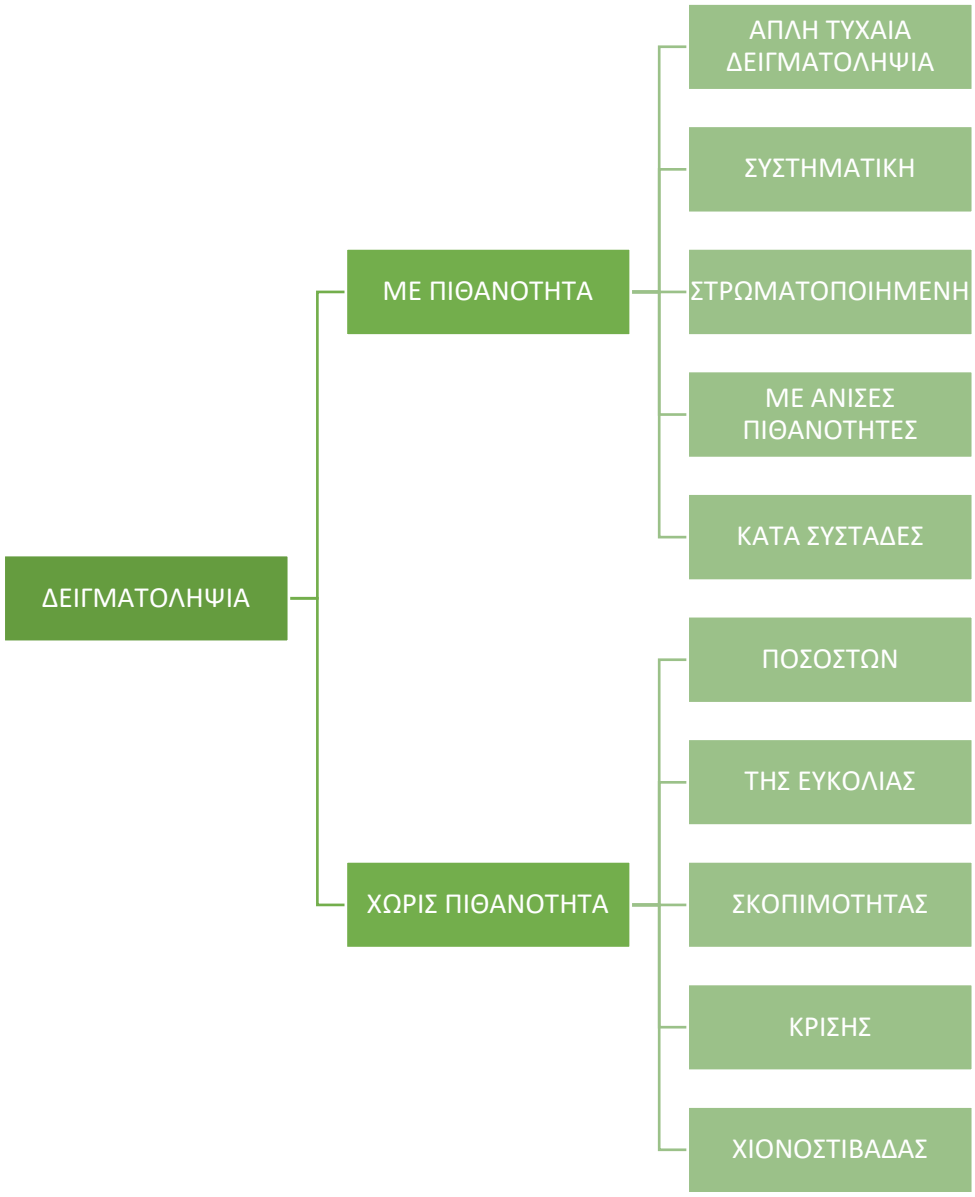
συμπεράσματα που εξάγονται από ένα δείγμα, για αυτό και δίνει επιπλέον τη δυνατότητα να υπολογίσουμε και το σφάλμα εκτίμησης.
Η δειγματοληψία χωρίς πιθανότητα γίνεται σε περιπτώσεις που δεν είναι εφικτή η δειγματοληψία με πιθανότητα ή όταν ενδιαφέρει να γίνει γρήγορα μια εφαρμογή της έρευνας, για παράδειγμα σε μια πιλοτική μελέτη. Τα αποτελέσματα μιας έρευνας που έχει γίνει με δειγματοληψία χωρίς πιθανότητα δεν είναι γενικεύσιμα, ούτε δύναται να υπολογισθεί το σφάλμα εκτίμησης, και ως εκ τούτου είναι περιορισμένης χρήσης και εφαρμογής και θα πρέπει να χρησιμοποιούνται προσεκτικά.

Στη κατηγορία των τεχνικών δειγματοληψίας με πιθανότητα περιλαμβάνονται οι εξής τεχνικές:

α) απλή τυχαία δειγματοληψία
β) συστηματική δειγματοληψία
γ) στρωματοποιημένη δειγματοληψία
δ) δειγματοληψία κατά συστάδες
ε) δειγματοληψία κατά συστάδες με άνισες πιθανότητες

ενώ στη κατηγορία των τεχνικών δειγματοληψίας χωρίς πιθανότητα εντάσσονται οι:

α) δειγματοληψία ποσοστών,
β) δειγματοληψία ευκολίας
γ) δειγματοληψία κρίσης
δ) δειγματοληψία χιονοστιβάδας
ε) δειγματοληψία σκοπιμότητας.



Παρακάτω αναλύονται οι διάφορες τεχνικές δειγματοληψίας που αναφέραμε.

ΚΕΦΑΛΑΙΟ 4

Δειγματοληψία με πιθανότητες

Η πρώτη δειγματοληπτική μέθοδος, απλής τυχαίας δειγματοληψίας, δεν είναι συνήθως πολύ πρακτική, αλλά πάνω της βασίζεται η Θεωρία της Δειγματοληψίας, και επιπλέον βοηθάει στην κατανόηση των εννοιών που αναφέρθηκαν παραπάνω όπως η εκτίμηση των παραμέτρων και η επιλογή δειγμάτων. Σπάνια η μέθοδος αυτή εφαρμόζεται αποκλειστικά σε μια δειγματοληπτική έρευνα.

4.1 Απλή Τυχαία Δειγματοληψία

Ορισμοί

Η απλή τυχαία δειγματοληψία αποτελεί τη βάση κάθε στατιστικής επεξεργασίας στοιχείων και πληροί δύο βασικές στατιστικές συνθήκες:

- Κάθε μονάδα του πληθυσμού έχει ίσες πιθανότητες να επιλεγεί για το δείγμα
- Η επιλογή μιας μονάδας του πληθυσμού για το δείγμα με κανένα τρόπο δεν επηρεάζει την επιλογή κάποιας άλλης μονάδας του πληθυσμού.

Εξασφαλίζεται δηλαδή, η ίση πιθανότητα επιλογής παρατηρήσεων και η ανεξαρτησία. Το δειγματοληπτικό υπόβαθρο μπορεί να είναι ένας κατάλογος. Η επιλογή του τυχαίου δείγματος από το υπόβαθρο αυτό γίνεται με τη χρήση πινάκων τυχαίων αριθμών.
Στην περίπτωση που ο αρχικός πληθυσμός είναι άπειρος, η λήψη ενός στοιχείου με επανάθεση, δηλαδή επανατοποθέτηση του επιλεγέντος στοιχείου κατά την προηγούμενη επιλογή, ή χωρίς επανάθεση, δεν αλλοιώνει την αρχική σύνθεση του πληθυσμού. Έτσι, στην περίπτωση αυτή η λήψη του τυχαίου δείγματος συνεπάγεται ισονομία – ίση πιθανότητα επιλογής και ανεξαρτησία.

Η χρήση της απλής, τυχαίας δειγματοληψίας δεν οδηγεί στη δημιουργία αντιπροσωπευτικών δειγμάτων. Το δείγμα μπορεί να αφήνει περιοχές του πληθυσμού ακάλυπτες και τίποτα δεν εξασφαλίζει ότι υπάρχει αντιπροσωπευτικότητα ως προς τα χαρακτηριστικά που μας ενδιαφέρουν. Η απλή, τυχαία δειγματοληψία δεν έχει το μικρότερο σφάλμα εκτίμησης, σε αντίθεση με άλλες μεθόδους (π.χ. στρωματοποιημένη δειγματοληψία), αλλά παρουσιάζει σχετική ευκολία στη χρήση της.

Έστω N το μέγεθος του πληθυσμού και n το μέγεθος του δείγματος. Συνολικά υπάρχουν $\binom{N}{n}$ δυνατά δείγματα μεγέθους n, όπου

$$\binom{N}{n} = \frac{N!}{(N-n)!\,n!}$$

$$n! = n(n-1)(n-2)\ldots 2 * 1.$$

Το σύνολο των δυνατών δειγμάτων στην περίπτωση αυτή θα είναι:

$$S = \left\{s_1, s_2, \dots, s_{\binom{N}{n}}\right\}$$

Ορισμός απλής τυχαίας δειγματοληψίας: Απλή τυχαία δειγματοληψία είναι η δειγματοληψία κατά την οποία όλα τα δυνατά δείγματα του πληθυσμού μεγέθους n $(n = 1,2, \dots, N)$ έχουν την ίδια πιθανότητα να επιλεγουν, η οποία είναι ίση με $1/\binom{N}{n}$.

Δειγματοληψία με επανατοποθέτηση: Όταν επιλέγεται μία μονάδα του πλυθησμού για το δείγμα και επανατοποθετείται στον πληθυσμό. Συνεπώς οποιαδήποτε μονάδα του πληθυσμού έχει τη δυνατότητα να εμφανιστεί περισσότερες από μια φορές στο δείγμα. Σε αυτή την περίπτωση το πλήθος των δυνατών δειγμάτων είναι N^n.

Δειγματοληψία χωρίς επανατοποθέτηση: Η μονάδα του πληθυσμού που επιλέγεται για το δείγμα, δεν ξανατοποθετείται στον πληθυσμό.

Η πιο συνηθισμένη μορφή απλής τυχαίας δειγματοληψίας γίνεται χωρίς επανατοποθέτηση, έτσι ώστε όλα τα στοιχεία του πληθυσμού να παραμένουν ισοπίθανα.

Η διαδικασία για την εκτέλεση απλής τυχαίας δειγματοληψίας είναι η εξής:

1. Αριθμούμε τις μονάδες του πλυθησμού από το 1 εως το N.
2. Με έναν υπολογιστή ή πίνακα τυχαίων αριθμών, επιλέγουμε n αριθμούς όχι ίσους μεταξύ τους.
3. Οι μονάδες του πληθυσμού που αντιστοιχούν στους αριθμούς αυτούς αποτελούν το δείγμα.

Εκτίμηση παραμέτρων στην Απλή Τυχαία Δειγματοληψία

Εάν υποθέσουμε ότι το υπό μελέτη χαρακτηριστικό του πληθυσμού είναι ποσοτικό, δηλαδή ότι η τυχαία μεταβλητή που το περιγράφει είναι συνεχής, τότε μπορούμε να εκτιμήσουμε τις διάφορες παραμέτρους που είναι χρήσιμες για τον πληθυσμό.

1. ΕΚΤΙΜΗΣΗ ΤΟΥ ΜΕΣΟΥ ΓΙΑ ΤΟ ΥΠΟ ΜΕΛΕΤΗ ΧΑΡΑΚΤΗΡΙΣΤΙΚΟ

Στην απλή τυχαία δειγματοληψία, ο δειγματικός μέσος $\bar{X}$ είναι ένας αμερόληπτος εκτιμητής του πληθυσμιακού μέσου $\bar{Y}$.

Απόδειξη: Έστω Y_i οι μονάδες του πληθυσμού και V_i οι τυχαίες μεταβλητές που ορίζονται ως εξής:

$$V_i = \begin{cases} 1, & \text{αν η μονάδα } Y_i \text{ επιλεγεί στο δείγμα} \\ 0, & \text{αν η μονάδα } Y_i \text{ δεν επιλεγεί στο δείγμα} \end{cases}$$

Τότε $V_i \sim Bernoulli(p)$ $με$ $p = \frac{n}{N}$ δηλαδή $E(V_i) = p$
Ο τύπος που δίνει την δειγματική μέση τιμή δίνεται από τον τύπο

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{N} X_i$$

η οποία θα γράφεται ισοδύναμα

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{N} V_i Y_i$$

Αν υπολογίσουμε την αναμενόμενη τιμή στην τελευταία σχέση τότε προκύπτει:

$$E(\bar{X}) = \frac{1}{n}\sum_{i=1}^{N} E(V_i)Y_i = \frac{1}{n}\sum_{i=1}^{N} \frac{n}{N} Y_i = \frac{1}{N}\sum_{i=1}^{N} Y_i = \bar{Y}$$

Συνεπώς ο δειγματικός μέσος $\bar{X}$ είναι αμερόληπτος εκτιμητής του πληθυσμιακού μέσου $\bar{Y}$. (Ιουλία Παπαγεωργίου)

2. *ΔΙΑΣΠΟΡΑ ΤΟΥ ΜΕΣΟΥ*

Η διασπορά του μέσου $\bar{X}$ ενός απλού τυχαίου δείγματος μεγέθους n από ένα πληθυσμό μεγέθους N είναι:

$$Var(\bar{X}) = \frac{\sigma^2}{n} * \frac{N-n}{N}$$

Απόδειξη: Γνωρίζουμε ότι:

$$V(\bar{X}) = E(\bar{X} - \mu)^2$$

Επιπλέον,

$$n(\bar{X} - \mu) = (X_1 - \mu) + (X_2 - \mu) + \cdots + (X_n - \mu)$$
$$n^2(\bar{X} - \mu)^2 = (X_1 - \mu)^2 + (X_2 - \mu)^2 + \cdots + (X_n - \mu)^2 + 2[2(X_1 - \mu)(X_2 - \mu) + \cdots + (X_{n-1} - \mu)(X_n - \mu)]$$

Συνεπώς,

$$n^2E(\bar{X}-\mu)^2 = E\left(\sum_{i=1}^{n}(X_i-\mu)^2\right) + 2E\left(\sum_{i<j}^{n}(X_i-\mu)(X_j-\mu)\right) *$$

Επειδή κάθε μονάδα του πληθυσμού περιέχεται στον ίδιο αριθμό δειγμάτων, η ποσότητα $E(X_1 + \ ... \ + X_n)$ πρέπει να είναι υποπολλαπλάσιο του $y_1 + y_2 + ... + y_N$. Επειδή η πρώτη ποσότητα έχει n όρους και η δεύτερη N όρους, έπεται ότι ο πολλαπλασιαστής έχει την τιμή $\frac{n}{N}$. Επομένως,

$$E\left(\sum_{i=1}^{n}(X_i-\mu)^2\right) = \frac{n}{N}\left(\sum_{i=1}^{N}(Y_i-\mu)^2\right)$$

Επιπλέον,

$$E\left(\sum_{i<j}^{n}(X_i-\mu)(X_j-\mu)\right) = \frac{n(n-1)}{N(N-1)}\cdot\sum_{i<j}^{n}(Y_i-\mu)\big(Y_j-\mu\big)$$

Διότι το αριστερό μέλος έχει $\binom{n}{2}$ όρους ενώ το δεξί έχει $\binom{N}{2}$ όρους. Τότε η σχέση * γίνεται:

$$n^2E(\bar{X}-\mu)^2 = \frac{n}{N}\left[\left(\sum_{i=1}^{N}(Y_i-\mu)^2\right) + 2\frac{n-1}{N-1}\left(\sum_{i<j}^{N}(Y_i-\mu)(Y_j-\mu)\right)\right]$$

Προσθαφαιρώντας το

$$\frac{n-1}{N-1}\sum_{i=1}^{N}(Y_i-\mu)^2$$

στο δεξί μέλος της παραπάνω ισότητας, προκύπτει ότι:

$$n^2E(\bar{X}-\mu)^2 = \frac{n}{N}\left[(1-\frac{n-1}{N-1}\left(\sum_{i=1}^{N}(Y_i-\mu)^2\right) + \frac{n-1}{N-1}\left(\sum_{i<j}^{N}(Y_i-\mu)^2\right)^2\right]$$

$$= \frac{n(N-n)}{N(N-1)}\sum_{i<j}^{N}(Y_i-\mu)^2.$$

Συνεπώς:

$$E(\bar{X}-\mu)^2 = \frac{\sigma^2}{n}\cdot\frac{N-n}{N}.$$

Παρατηρήσεις:

- *Πηλίκο Δείγματος* είναι η ποσότητα $\frac{n}{N}$ και συμβολίζεται με f. Από τον ορισμό του, ισχύει ότι $0 < f < 1$.

- *Διόρθωση πεπερασμένου πληθυσμού (fpc)* είναι η ποσότητα $1-f$. Η ποσότητα αυτή διαφοροποιεί το αποτέλεσμα της $Var(\bar{X})$ από αντίστοιχο αποτέλεσμα για άπειρους πληθυσμούς. (Ιουλία Παπαγεωργίου)

- Με βάση τα παραπάνω η διασπορά του μέσου μπορεί να πάρει τη μορφή

$$Var(\bar{X}) = \frac{1-f}{n}\cdot\sigma^2 .$$

- *Τυπικό σφάλμα εκτιμητή* $se(\bar{X}) = \sqrt{Var(\bar{X})}$

- *Συντελεστής μεταβλητότητας του εκτιμητή*

$$CV(\bar{X}) = \frac{se(\bar{X})}{\bar{Y}}$$

3. ΕΚΤΙΜΗΣΗ ΤΗΣ ΔΙΑΣΠΟΡΑΣ σ^2 ΤΟΥ ΠΛΗΘΥΣΜΟΥ

Στην απλή τυχαία δειγματοληψία η σ.σ

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2$$

είναι μια αμερόληπτη εκτιμήτρια της διασποράς σ^2 του πληθυσμού.

Δηλαδή ισχύει ότι:

$$E(S^2) = \sigma^2$$

Απόδειξη:

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{1}{n-1}\sum_{i=1}^{n}[(X_i - \mu) - (\bar{X} - \mu)]^2 =$$

$$= \frac{1}{n-1}\left[\sum_{i=1}^{n}(X_i - \mu)^2 - n(\bar{X} - \mu)^2\right]$$

Παίρνοντας τη μέση τιμή της παραπάνω σχέσης, προκύπτει ότι:

$$E(S^2) = E\left(\frac{1}{n-1}\left[\sum_{i=1}^{n}(X_i - \mu)^2 - n(\bar{X} - \mu)^2\right]\right) =$$

$$= \frac{1}{n-1}\left[E(\sum_{i=1}^{n}(X_i - \mu)^2) - E(n(\bar{X} - \mu)^2)\right]$$

$$= \frac{1}{n-1}\left[\frac{n}{N}\sum_{i=1}^{N}(Y_i - \mu)^2 - \frac{N-n}{N}\,\sigma^2\right]$$

$$= \frac{1}{n-1}\left[\frac{n(N-1)}{N}\sigma^2 - \frac{N-n}{N}\,\sigma^2\right]$$

$$= \frac{\sigma^2}{(n-1)N}[n(N-1) - (N-n)]$$

$$= \sigma^2$$

(Ιουλία Παπαγεωργίου)

4. ΕΚΤΙΜΗΣΗ ΤΗΣ ΔΙΑΣΠΟΡΑΣ ΤΟΥ ΜΕΣΟΥ $Var(\bar{X})$

Μια αμερόληπτη εκτιμήτρια της διασποράς του μέσου $Var(\bar{X})$ είναι η συνάρτηση

$$\hat{Var}(\bar{X}) = \frac{1-f}{n} \cdot s^2 .$$

5. *ΕΚΤΙΜΗΣΗ ΤΟΥ ΤΥΠΙΚΟΥ ΣΦΑΛΜΑΤΟΣ ΤΟΥ ΜΕΣΟΥ $\bar{X}$*

Μια αμερόληπτη εκτιμήτρια του τυπικού σφάλματος $se(\bar{X})$ είναι η συνάρτηση

$$\widehat{se}(\bar{X}) = \sqrt{\hat{Var}(\bar{X})} = \sqrt{\frac{1-f}{n} \cdot s^2}$$

6. *ΕΚΤΙΜΗΣΗ ΤΟΥ ΣΥΝΟΛΟΥ ΤΟΥ ΠΛΗΘΥΣΜΟΥ $X_T = N\bar{X}$*

Σύνολο του πληθυσμού ή άθροισμα των τιμών ενός χαρακτηριστικού είναι το

$$Y \text{ ή } Y_T = \sum_{i=1}^{N} Y_i.$$

Επίσης ισχύει ότι $Y_T = N\bar{Y}$.

Μια αμερόληπτη εκτιμήτρια του Y_T είναι ο $\boldsymbol{X_T = N\bar{X}}$
Επομένως ισχύει ότι $E(X_T) = Y_T$

7. *ΔΙΑΣΠΟΡΑ ΚΑΙ ΤΥΠΙΚΟ ΣΦΑΛΜΑ ΤΟΥ X_T*

Η διασπορά του εκτιμητή X_T δίνεται από τη σχέση:

$$Var(X_T) = N^2 \frac{1-f}{n} \cdot \sigma^2$$

Ενώ το τυπικό σφάλμα του X_T δίνεται από τη σχέση

$$se(X_T) = N\sqrt{\frac{1-f}{n} \cdot \sigma^2}$$

8. *ΕΚΤΙΜΗΣΗ ΔΙΑΣΠΟΡΑΣ ΚΑΙ ΤΥΠΙΚΟΥ ΣΦΑΛΜΑΤΟΣ ΤΟΥ X_T*

Οι εκτιμήτριες της διασποράς και του τυπικού σφάλματος για το σύνολο X_T είναι αμερόληπτες και δίνονται από τις σχέσεις:

$$\hat{Var}(X_T) = N^2 \frac{1-f}{n} \cdot S^2$$

Και

$$\widehat{se}(X_T) = N\sqrt{\frac{1-f}{n} \cdot S^2}$$

9. *ΕΚΤΙΜΗΣΗ ΠΟΣΟΣΤΟΥ*

Αν p το ποσοστό του πληθυσμού που ανήκει σε μία κατηγορία, A ο αριθμός των μονάδων του πληθυσμού που ανήκουν στην συγκεκριμένη κατηγορία, τότε το ποσοστό του πληθυσμού δίνεται από τον τύπο:

$$p = \frac{A}{N}$$

Ο αριθμός A μπορεί να οριστεί και ως εξής:

Έστω $Y_i = \begin{cases} 1, & \alpha\nu\ \tau o\ i\ \mu\acute{\varepsilon}\lambda o\varsigma\ \alpha\nu\acute{\eta}\kappa\varepsilon\iota\ \sigma\tau\eta\nu\ \upsilon\pi\acute{o}\ \mu\varepsilon\lambda\acute{\varepsilon}\tau\eta\ \kappa\alpha\tau\eta\gamma o\rho\acute{\iota}\alpha \\ 0, & \delta\iota\alpha\phi o\rho\varepsilon\tau\iota\kappa\acute{\alpha} \end{cases}$

Συνεπώς:

$$A = \sum_{i=1}^{N} Y_i$$

Και το ποσοστό παίρνει τη μορφή:

$$p = \frac{1}{N}\sum_{i=1}^{N} Y_i$$

Ο εκτιμητής του ποσοστού p δίνεται από τον τύπο

$$\hat{p} = \frac{a}{n}$$

όπου α είναι ο αριθμός των μονάδων του δείγματος που ανήκουν στην υπό μελέτη κατηγορία.

Ο εκτιμητής $\hat{p}$ του ποσοστού p είναι αμερόληπτος, δηλαδή ισχύει ότι:

$$E(\hat{p}) = p$$

Ο εκτιμητής του αριθμού A δίνεται από τον τύπο

$$\hat{A} = N\frac{\alpha}{n}$$

Η πληθυσμιακή διασπορά σ^2 δίνεται μέσω του ποσοστού από τον τύπο

$$\sigma^2 = \frac{Np(1-p)}{N-1}$$

Απόδειξη: Είναι

$$\sigma^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left[{Y_i}^2 - \frac{1}{N}\left(\sum_{i=1}^{N} Y_i\right)^2\right]$$

Όπου Y_i ορίζονται όπως παραπάνω ακολουθώντας κατανομή Bernoulli, συνεπώς παίρνουν τιμές 0 ή 1. Άρα:

$$\sigma^2 = \frac{1}{N-1}\left(A - \frac{A^2}{N}\right) = \frac{1}{N-1}\left(Np - \frac{Np^2}{N}\right) = \frac{Np(1-p)}{N-1}$$

Όμοια, η δειγματική διασπορά δίνεται μέσω του ποσοστού από τον τύπο:

$$s^2 = \frac{n\hat{p}(1-\hat{p})}{n-1}$$

10. ΔΙΑΣΠΟΡΑ ΚΑΙ ΤΥΠΙΚΟ ΣΦΑΛΜΑ ΠΟΣΟΣΤΟΥ

Η διασπορά του ποσοστού $\hat{p}$ δίνεται από τον τύπο:

$$Var(\hat{p}) = \frac{1-f}{n}\sigma^2 = \frac{1-f}{n} \cdot \frac{Np(1-p)}{N-1}$$

Ενώ το τυπικό σφάλμα δίνεται από τον τύπο:

$$se(\hat{p}) = \sqrt{\frac{1-f}{n}\sigma^2} = \sqrt{\frac{1-f}{n} \cdot \frac{Np(1-p)}{N-1}}$$

11. ΕΚΤΙΜΗΣΗ ΔΙΑΣΠΟΡΑΣ ΚΑΙ ΤΥΠΙΚΟΥ ΣΦΑΛΜΑΤΟΣ ΠΟΣΟΣΤΟΥ

Οι εκτιμήτριες της διασποράς και του τυπικού σφάλματος του ποσοστού δίνονται από τους τύπους:

$$\hat{Var}(\hat{p}) = \frac{1-f}{n}s^2 = (1-f) \cdot \frac{p(1-p)}{n-1}$$

Και

$$\widehat{se}(\hat{p}) = \sqrt{(1-f) \cdot \frac{p(1-p)}{n-1}}$$

Διαστήματα εμπιστοσύνης στην Απλή Τυχαία Δειγματοληψία

Αναζητούμε διαστήματα εμπιστοσύνης για τις παραμέτρους $\bar{Y}, Y_T, p, A$ του πληθυσμού.

Έχει αποδειχθεί (βλέπε Ε. Ξεκαλάκη και Ι. Πανάρετου: Πιθανότητες και Στοιχεία Στοχαστικών Ανελίξεων, Αθήνα 1993) ότι η κανονική κατανομή είναι η οριακή μορφή της κατανομής του μέσου $\bar{X}$ ενός τυχαίου δείγματος μεγέθους n, το οποίο προέρχεται από έναν άπειρο πληθυσμό με πεπερασμένη διασπορά, όταν το n τείνει στο ∞. Δηλαδή, αν μ και σ^2 συμβολίζουν την μέση τιμή και την διασπορά του πληθυσμού αντίστοιχα, τότε, όταν το δείγμα είναι αρκετά μεγάλο (n→∞) τότε η μέση τιμή ακολουθεί κανονική κατανομή, δηλαδή

$$\bar{X} \sim N(\mu, \sigma^2) \quad \text{ή ισοδύναμα} \quad \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

Το πόσο μεγάλο πρέπει να είναι το μέγεθος n του δείγματος ώστε η προσέγγιση της πραγματικής κατανομής του $\bar{X}$ από την κανονική κατανομή να είναι ικανοποιητική δεν καθορίζεται από κάποιο γενικό κανόνα. Στις περισσότερες εφαρμογές, το n δεν συνηθίζεται να είναι μικρότερο του 25 ($n \geq 25$). Το παραπάνω αποτέλεσμα είναι γνωστό ως κεντρικό οριακό θεώρημα. Η πρακτική αξία του θεωρήματος αυτού είναι μεγάλη εξ αιτίας των δυνατοτήτων που δίνει στον ερευνητή όσο αφορά την συναγωγή στατιστικών συμπερασμάτων.

Βασιζόμενοι στην υπόθεση της κανονικότητας, το $100(1-\alpha)\%$ Δ.Ε. για την πραγματική τιμή του μέσου $\bar{Y}$ με γνωστή την πληθυσμιακή διασπορά σ^2 είναι το διάστημα:

$$\left(\bar{X} - z_{\frac{a}{2}} \cdot \sigma \sqrt{\frac{1-f}{n}} \ , \qquad \bar{X} + z_{\frac{a}{2}} \cdot \sigma \sqrt{\frac{1-f}{n}} \right)$$

Όπου z_a το άνω α-εκατοστιαίο σημείο για την τυπική κανονική κατανομή, δηλαδή

$$P(Z \geq z_a) = a \quad \text{για} \quad Z \sim N(0,1)$$

Οι πιο συνηθισμένες τιμές του α είναι 5% και 1%.

Για α=5% τότε $z_{0.975} = 1.96$

Για α=1% τότε $z_{0.005} = 2.56$.

Όταν η διασπορά του πληθυσμού σ^2 είναι άγνωστη, τότε χρησιμοποιείται η εκτιμήτρια s^2

Για $n < 30$ η τ.μ. $\bar{X}$ ακολουθεί κατανομή Student-t και τότε το $100(1-\alpha)\%$ Δ.Ε. για την πραγματική τιμή του μέσου $\bar{Y}$ με άγνωστη την πληθυσμιακή διασπορά σ^2 είναι το διάστημα:

$$\left(\bar{X} - t_{n-1,\,\frac{a}{2}} \cdot \sigma\sqrt{\frac{1-f}{n}}\ ,\qquad \bar{X} + t_{n-1,\,\frac{a}{2}} \cdot \sigma\sqrt{\frac{1-f}{n}}\right)$$

(Κοκολάκης Γ., Σπηλιώτης Ι., 1999)

Για $n \geq 30$ η τ.μ. $\bar{X}$ προσεγγίζεται ικανοποιητικά από την κανονική άρα το $100(1-\alpha)\%$ Δ.Ε. για την πραγματική τιμή του μέσου $\bar{Y}$ με άγνωστη την πληθυσμιακή διασπορά σ^2 είναι το διάστημα:

$$\left(\bar{X} - z_{\frac{a}{2}} \cdot s\sqrt{\frac{1-f}{n}}\ ,\qquad \bar{X} + z_{\frac{a}{2}} \cdot s\sqrt{\frac{1-f}{n}}\right)$$

Ανάλογα με τη μέση τιμή ενός πληθυσμού μπορούμε να κατασκευάσουμε και τα διαστήματα εμπιστοσύνης για τα Y_T , p *και* A. Για την εκτιμώμενη διασπορά των εκτιμητριών τους και για μεγάλο αριθμό $n \geq 30$, τα διαστήματα αυτά είναι:

- Για το Y_T

$$\left(N\bar{X} - z_{\frac{a}{2}} \cdot Ns\sqrt{\frac{1-f}{n}}\ ,\qquad N\bar{X} + z_{\frac{a}{2}} \cdot Ns\sqrt{\frac{1-f}{n}}\right)$$

- Για το p

$$\left(\hat{p} - z_{\frac{a}{2}} \cdot \sqrt{\frac{(1-f)p(1-p)}{n-1}}\ ,\qquad \hat{p} + z_{\frac{a}{2}} \cdot \sqrt{\frac{(1-f)p(1-p)}{n-1}}\right)$$

- Για το A

$$\left(N\hat{p} - z_{\frac{a}{2}} \cdot N\sqrt{\frac{(1-f)p(1-p)}{n-1}} \ , \qquad N\hat{p} + z_{\frac{a}{2}} \cdot N\sqrt{\frac{(1-f)p(1-p)}{n-1}} \right)$$

Εκτίμηση ενός λόγου

Έστω πληθυσμός μεγέθους N και έστω X_i και Y_i οι τιμές δύο διαφορετικών χαρακτηριστικών της i μονάδας του πληθυσμού, i = 1,2,…,N. Τότε

$$X = \sum_{i=1}^{N} X_i \quad \kappa\alpha\iota \ \ Y = \sum_{i=1}^{N} Y_i$$

Αντιπροσωπεύουν τις συνολικές τιμές των δύο χαρακτηριστικών. Πολλές φορές στην πράξη μας ενδιαφέρει να εκτιμήσουμε τον λόγο των δύο αυτών ποσοστήτων, δηλαδή τον

$$R = \frac{\sum_{i=1}^{N} X_i}{\sum_{i=1}^{N} Y_i} = \frac{\frac{\sum_{i=1}^{N} X_i}{N}}{\frac{\sum_{i=1}^{N} Y_i}{N}} = \frac{\mu_X}{\mu_Y}$$

Η εκτιμήτρια του λόγου R είναι:

$$\hat{R} = \frac{\sum_{i=1}^{n} X_i}{\sum_{i=1}^{n} Y_i} = \frac{\bar{X}}{\bar{Y}}$$

Για την αξιολόγηση της ακριβείας της εκτιμήτριας αυτής και την εξαγωγή συμπερασμάτων, είναι αναγκαίος ο προσδιορισμός της κατανομής της και του τυπικού της σφάλματος. Για μικρές τιμές του n η κατανομή της $\hat{R}$ δεν είναι κανονική. Αντίθετα, είναι ασύμμετρη προς τα δεξιά (έχει μακριά δεξιά ουρά). Επί πλέον, η $\hat{R}$ δεν είναι αμερόληπτη εκτιμήτρια του R. Για μεγάλες όμως τιμές του n, η κατανομή της $\hat{R}$ τείνει στην κανονική κατανομή και ισχύει το παρακάτω θεώρημα:

ΘΕΩΡΗΜΑ: Για μεγάλο $n \to N$ η στατιστική συνάρτηση

$$\hat{R} = \frac{\sum_{i=1}^{n} X_i}{\sum_{i=1}^{n} Y_i} = \frac{\bar{X}}{\bar{Y}}$$

Είναι αμερόληπτη εκτιμήτρια του R και η διασπορά της είναι

$$V\left(\hat{R}\right) = \frac{1-f}{n\,\mu^2{}_X} \frac{\sum_{i=1}^{N}(Y_i - RX_i)^2}{N-1}$$

(Κοκολάκης Γ., Σπηλιώτης Ι., 1999)

Απόδειξη: Ισχύει ότι

$$\hat{R} - R = \frac{\bar{X}}{\bar{Y}} - R = \frac{\bar{X} - R\bar{Y}}{\bar{Y}}$$

Αλλά ο μέσος ενός απλού τυχαίου δείγματος είναι συνεπής εκτιμήτρια της μέσης τιμής του πληθυσμού. Άρα, για μεγάλες τιμές του n, η τιμή του μέσου $\bar{X}$ δεν διαφέρει πολύ από την μ_X και, επομένως, ισχύει κατά προσέγγιση ότι

$$\hat{R} - R \cong \frac{\bar{X} - R\bar{Y}}{\mu_Y}$$

Παίρνοντας μέση τιμή προκύπτει ότι

$$E\left(\hat{R} - R\right) = \frac{1}{\mu_Y} E(\bar{X} - R\bar{Y}) = \frac{1}{\mu_Y} (\mu_X - R\mu_Y) = R - R = 0$$

Άρα, με την προσέγγιση που θεωρήθηκε, αποδείχθηκε ότι η $\hat{R}$ είναι αμερόληπτη εκτιμήτρια του R. Επομένως,

$$V(\hat{R}) = E(\hat{R} - R)^2 = \frac{1}{\mu^2{}_X}\, E(\bar{Y} - R\bar{X})^2$$

Όμως η τ.μ. $\bar{Y} - R\bar{X}$ έχει μέση τιμή 0. Άρα

$$\begin{aligned}\mu^2{}_X V(\hat{R}) &= \mathrm{V}(\bar{Y} - R\bar{X}) \\ &= E(\bar{Y} - R\bar{X})^2\end{aligned}$$

$$= \frac{\sigma^2}{n}\ \left(1 - \frac{n}{N}\right)$$

$$= \frac{\sum_{i=1}^{N}(Y_i - RX_i)^2}{N-1}\ \ \frac{1 - \frac{n}{N}}{n}$$

$$= \frac{1 - \frac{n}{N}}{n} \sum_{i=1}^{N} \frac{(Y_i - RX_i)^2}{N-1}$$

Τέλος απόδειξης.

Μια εκτιμήτρια της $V(\hat{R})$ είναι η στατιστική συνάρτηση

$$\sigma_R{}^2 = \frac{1 - \frac{n}{N}}{n\bar{X}^2} \cdot \frac{\sum_{i=1}^{n}(Y_i - \hat{R}X_i)^2}{n-1}$$

Προσδιορισμός μεγέθους δείγματος

Ένα πρόβλημα που απασχολεί τους στατιστικούς μελετητές είναι ο προσδιορισμός του μεγέθους του δείγματος, καθώς αυτό παίζει καθοριστικό ρόλο στην αξιοπιστία των εκτιμητριών που θα παραχθούν. Η αξιοπιστία των εκτιμιτριών είναι ο σημαντικότερος παράγοντας των δειγματοληπτικών ερευνών, συνεπώς ένας τρόπος για την επιλογή του μεγέθους του δείγματος είναι να υπάρχουν κάποιες απαιτήσεις ως προς την αξιοπιστία των εκτιμητριών.

Τα βήματα για την επίλυση του προβλήματος εκτίμησης παραμέτρων είναι τα εξής:

- Το πρώτο βήμα είναι να τεθούν οι απαιτήσεις που πρέπει να εκπληρωθούν και να ικανοποιούνται από τα αποτελέσματα της έρευνας, όταν ολοκληρωθεί. Οι απαιτήσεις αυτές συνήθως σχετίζονται με την αξιοπιστία των εκτιμητών και εκφράζονται σε απόλυτη μορφή ή με περιθώριο λάθους.

- Οι απαιτήσεις αυτές πρέπει να εκφράζονται με μαθηματικό τρόπο ώστε με κατάλληλους μετασχηματισμούς το πρόβλημα να μεταφραστεί σε μαθηματική σχέση από την οποία θα προκύπτει ο άγνωστος n.

- Η εξίσωση που θα διατυπωθεί για τον προσδιορισμό του μεγέθους δείγματος n ενδέχεται να περιέχει ποσότητες του πληθυσμού οι οποίες είναι άγνωστες κατά το στάδιο αυτό, και θα πρέπει να προσδιοριστούν με καποιο τρόπο από από την διεξαγωγή της έρευνας.

- Από τη λύση της μαθηματικής σχέσης αυτής, προκύπτει το προτεινόμενο μέγεθος τους δείγματος για την έρευνα. Το μέγεθος αυτό που προκύπτει πρέπει να μην υπερβαίνει τα όρια που είναι διαθέσιμα ως προς το κόστος και το χρόνο διεξαγωγής της έρευνας

- Μια παράμετρος που πρέπει να υπολογιστεί επίσης είναι το ποσοστό μη-απόκρισης (Non-response) κάποων μονάδων του πληθυσμού, το οποίο ανάλογα και με το είδος της έρευνας, μπορεί να είναι και αρκετά μεγάλο.

Εύρεση του ελάχιστα απαιτούμενου μεγέθους δείγματος

Για την εκτίμηση της μέσης τιμής μ και της διασποράς σ^2, το ερώτημα σχετικά με το μέγεθος του δείγματος μπορεί, λαμβάνοντας υπ' όψη τα παραπάνω, να διατυπωθεί ως εξής: ″Πόσο μεγάλο πρέπει να είναι το μέγεθος n του δείγματος, ώστε, με πιθανότητα 1-α, το σφάλμα που κάνει ο ερευνητής εκτιμώντας την άγνωστη μέση τιμή μ με τον μέσο $\bar{X}$ του δείγματος να μην υπερβαίνει την τιμή e;″

Το ερώτημα αυτό ισοδυναμεί με το ερώτημα: ″Ποια είναι η τιμή του n για την οποία ισχύει:

$$P(|\bar{X} - \mu| \leq e) = 1 - a \text{ ;"}$$

Γνωρίζουμε ότι ισχύει:

$$P(|\bar{X} - \mu| \leq e) = 1 - a$$

$$P\left(\left|\frac{\bar{X}-\mu}{\sigma_{\bar{X}}}\right| \leq \frac{e}{\sigma_{\bar{X}}}\right) = 1 - a$$

Η παραπάνω σχέση μπορεί να οδηγήσει στον προσδιορισμό της τιμής του n, αν, για το συγκεκριμένο πρόβλημα, μπορεί να υποτεθεί ότι ο πληθυσμός είναι κανονικός ή κατά προσέγγιση κανονικός. Στην περίπτωση αυτή,

$$\frac{\bar{X}-\mu}{\sigma_{\bar{X}}} \sim N(0,1)$$

Άρα πρέπει να ισχύει

$$\frac{e}{\sigma_{\bar{X}}} = z_{1-\alpha/2}$$

Δηλαδή,

$$\frac{\sigma}{\sqrt{n}}\sqrt{1-\frac{n}{N}} = \frac{e}{z_{1-\alpha/2}}$$

Και λύνοντας ως προς n προκύπτει ότι:

$$n = \frac{N\sigma^2 z^2{}_{1-\alpha/2}}{Ne^2 + \sigma^2 z^2{}_{1-\alpha/2}}$$

Όταν το N είναι πολύ μεγάλο, θεωρητικά άπειρο, τότε

$$n = \left(\frac{\sigma z_{1-\alpha/2}}{e}\right)^2$$

Αν η διασπορά σ^2 του πληθυσμού είναι άγνωστη, τότε χρησιμοποιείται μια εκτίμησή της βασισμένη σε κάποιο προκαταρκτικό δείγμα μεγέθους ≥ 30.

Ένας άλλος τρόπος για τον υπολογισμό του μεγέθους του δείγματος είναι με βάση κάποιο ποσοστό p

$$P(|\hat{p} - p| \leq e) = 1 - a$$

$$P\left(\left|\frac{\hat{p} - p}{\sigma_{\hat{p}}}\right| \leq \frac{e}{\sigma_{\hat{p}}}\right) = 1 - a$$

$$\frac{e}{\sigma_{\hat{p}}} = z_{1-\alpha/2}$$

Όμως

$$\sigma_{\hat{p}} = \frac{p(1-p)}{n}\frac{N-n}{N-1}$$

Άρα

$$n = \frac{N\, z^2{}_{1-\alpha/2}\, p(1-p)}{Ne^2 + z^2{}_{1-\frac{\alpha}{2}}\, p(1-p) - e^2}$$

Η παραπάνω σχέση δεν μπορεί να δώσει την λύση, αφού η τιμή του n εξαρτάται από την τιμή της άγνωστης παραμέτρου p. Στην πράξη, συνήθως χρησιμοποιείται μια εκτίμηση της παραμέτρου p που βασίζεται σε παλαιότερη εμπειρία. Αν δεν υπάρχουν ενδείξεις όσον αφορά το ποια περίπου είναι η τιμή του p, τότε χρησιμοποιείται η συντηρητική τιμή $p = 1/2$, η οποία μεγιστοποιεί το γινόμενο $p(1-p)$.

(Πράγματι, το γινόμενο $p(1-p) = p - p^2$ γίνεται μέγιστο όταν

$$(p - p^2)' = 0 \Leftrightarrow 1 - 2p = 0 \Leftrightarrow p = \frac{1}{2}) .$$

4.2 Στρωματοποιημένη Δειγματοληψία

Η στρωματοποιημένη δειγματοληψία είναι από τις πιο διαδεδομένες μεθόδους δειγματοληψίας. Το μεγαλύτερό της πλεονέκτημα είναι ότι δίνει τα μικρότερα τυπικά σφάλματα σε σχέση με τις άλλες μεθόδους δειγματοληψίας.

Στην στρωματοποιημένη δειγματοληψία, ο πληθυσμός χωρίζεται σε στρώματα, δηλαδή σε υποσύνολα του πληθυσμού τα οποία έχουν κάποια συγκεκριμένη ιδιότητα. Στη συνέχεια σε κάθε στρώμα λαμβάνεται κάποιο δείγμα στο οποίο πραγματοποιείται ξεχωριστή δειγματοληψία, και το τελικό δείγμα του συνολικού πληθυσμού αποτελείται από το σύνολο των επιμέρους δειγμάτων των στρωμάτων. Συνεπώς το τελικό δείγμα περιέχει μονάδες από κάθε στρώμα του πληθυσμού. Όσο περισσότερο ομοιογενή είναι τα στρώματα, τόσο πιο αντιπροσωπευτικό είναι το τελικό δείγμα. Δηλαδή οι μονάδες του πληθυσμού που είναι πιο παρόμοιες ως προς το χαρακτηριστικό που μελετάμε, τοποθετούνται στο ίδιο στρώμα.
Η εκτίμηση των παραμέτρων γίνεται σε κάθε στρώμα ξεχωριστά, και η εκτίμηση των παραμέτρων του πληθυσμού γίνεται με τον κατάλληλο συνδυασμό των επιμέρους εκτιμήσεων κάθε στρώματος.

Η στρωματοποιημένη δειγματοληψία είναι κοινή στρατηγική. Πρακτικοί αλλά και θεωρητικοί λόγοι οδηγούν σ'αυτή όπως για παράδειγμα: για κάποιο μέρος του πληθυσμού απαιτείται ιδιαίτερη ακρίβεια στις εκτιμήσεις. Άλλος λόγος είναι αν ο πληθυσμός είναι ήδη στρωματοποιημένος (π.χ. ένα σύνολο λιμνών, το σύνολο των δήμων μιας πόλης), ή αν ο πληθυσμός είναι ετερογενής αλλά στο εσωτερικό του περιέχει μέρη (ομάδες, συνιστώσες) που παρουσιάζουν σχετική ομοιογένεια. Αυτά τα μέρη θα αποτελέσουν τις στρώσεις.

Γενικά, η στρωματοποιημένη δειγματοληψία είναι πιο ακριβής από την απλή τυχαία δειγματοληψία.

Πλεονεκτήματα:

1. Είναι μια ευκολοπροσάρμοστη στρατηγική που συνδυάζεται και με άλλες οδηγώντας σε περίπλοκους σχεδιασμούς που όμως επιτρέπουν τον υπολογισμό της ακρίβειας των εκτιμητών και τη δημιουργία διαστημάτων εμπιστοσύνης.

2. Επιτρέπει την κατ'επιλογή μεγαλύτερη συμμετοχή στο δείγμα ατόμων του πληθυσμού που προέρχονται από συγκεκριμένες στρώσεις (αυτό μπορεί να εξυπηρετήσει παράλληλες μελέτες).

3. Επιτρέπει την ανάλυση της επίδρασης πάνω στα άτομα του πληθυσμού της παραμέτρου που χρησιμοποιήθηκε για τη στρωματοποίηση.

4. Επιτρέπει τη διεξαγωγή της δειγματοληψίας ακόμα κι αν διακυμάνσεις στην κατανομή της προσπάθειας στο χώρο ή το χρόνο είναι αναπόφευκτες (σε κάποιες περιοχές η

πρόσβαση είναι δύσκολη ή υπάρχουν δυσχέρειες για κάποιες περιόδους π.χ. νύχτα ή αργίες).

5. Ακόμα κι αν γίνουν λάθη στη στην κατανομή των ατόμων στις στρώσεις οι εκτιμήσεις παραμένουν αμερόληπτες.

6. Ακόμα κι αν δεν υπάρχουν πληροφορίες για μια έστω και στοιχειώδη στρωματοποίηση η στρατηγική αυτή εφαρμόζεται κατόπιν διπλής δειγματοληψίας (μια πρώτη χαλαρή δειγματοληψία για τη μελέτη των χαρακτηριστικών του πληθυσμού και του περιβάλλοντος του και στη συνέχεια με βάση αυτή την πληροφορία μια στρωματοποιημένη δειγματοληψία για τις τελικές εκτιμήσεις). Σ'αυτή την περίπτωση το κέρδος στην ακρίβεια της εκτίμησης φυσικά μειώνεται.

Συμβολισμοί

Διάφοροι συμβολισμοί:

- **Πλήθος στρωμάτων k**
- **Πληθυσμιακά μεγέθη κάθε στρώματος** $N_1, N_2, \dots, N_k$
- **Μέγεθος δείγματος στρώματος** N_i **:** $n_i \;\; i = 1, \dots, k$
- Y_{ij} **η τιμή του** j **μέλους του στρώματος** i
- $W_h = \frac{N_h}{N}$ **Βάρος του στρώματος h**

  Ισχύει ότι:

$$\sum_{h=1}^{k} W_h = 1$$

- **Μέση τιμή για το στρώμα** h

$$\overline{Y}_h = \frac{1}{N_h} \sum_{j=1}^{N_h} Y_{hj}$$

- **Πληθυσμιακή Διασπορά στρώματος** h

$${\sigma_h}^2 = \frac{1}{N_h - 1} \sum_{j=1}^{N_h} (Y_{hj} - \overline{Y}_h)^2$$

- X_{ij} **η** j **μονάδα του δείγματος που ανήκει στο στρώμα** i

- $n_1, n_2, \dots, n_k$
- $f_h = \frac{n_h}{N_h}$

- **Ο δειγματικός μέσος για το στρώμα h**

$$\bar{X}_h = \frac{1}{N_h} \sum_{j=1}^{N_h} X_{hj}$$

- **Δειγματική διασπορά για το στρώμα h**

$$s_h = \frac{1}{n_h - 1} \sum_{j=1}^{n_h} (X_{hj} - \bar{X}_h)^2$$

Ένα πρόβλημα που συνδέεται με την στρωματοποιημένη δειγματοληψία είναι ο επιμερισμός του συνολικού αριθμού των μονάδων του δείγματος, n, στα επί μέρους στρώματα δηλαδή ο προσδιορισμός των μεγεθών των k απλών δειγμάτων.

Ο επιμερισμός του n μπορεί να γίνει με τρεις τρόπους:.

α. Αναλογικός Επιμερισμός του n *(Proportional Allocation)*

β. Βέλτιστος Επιμερισμός του n με σταθερό κόστος ανά δειγματοληπτική μονάδα *(Optimum Allocation with constant cost per unit)*

γ. Βέλτιστος Επιμερισμός *(Optimum Allocation)*

Το πρόβλημα που συνδέεται άμεσα με την τεχνική της στρωματοποιημένης δειγματοληψίας είναι ο καταμερισμός του συνολικού δειγματικού μεγέθους n στα k διαθέσιμα στρώματα, δηλαδή ο καθορισμός των τιμών των μεγεθών $n_1, n_2, \dots, n_k$ των k απλών τυχαίων δειγμάτων. Αν το δειγματοληπτικό κόστος ανά μονάδα είναι το ίδιο σε όλα τα στρώματα και οι διασπορές των στρωμάτων δεν διαφέρουν σημαντικά, τα μεγέθη $n_1, n_2, \dots, n_k$ συνηθίζεται να επιλέγονται έτσι ώστε

$$n_i = n \cdot \frac{N_i}{N} \quad i = 1, \dots, k$$

Ο σχεδιασμός αυτός είναι γνωστός ως αναλογικός καταμερισμός του n (proportional allocation) και η δειγματοληπτική τεχνική ονομάζεται αναλογική στρωματοποιημένη τυχαία δειγματοληψία (proportional stratified random sampling).

Το μέγεθος του δείγματος από ένα στρώμα είναι ανάλογο του ποσοστού των μονάδων του πληθυσμού που το στρώμα εκπροσωπεί. Υπάρχουν, όμως, περιπτώσεις όπου οι τιμές του πληθυσμού έχουν μεγαλύτερη διακύμανση σε μερικά στρώματα από ό,τι σε άλλα. Διαφέρουν δηλαδή σημαντικά οι διασπορές των στρωμάτων. Επομένως, για να αντιπροσωπευθούν επαρκώς τα στρώματα αυτά στο δείγμα, θα πρέπει ο λόγος $\frac{N_i}{N}$ να είναι ανάλογος της τυπικής απόκλισης σ_i του στρώματος. Αυτό σημαίνει ότι στρώματα με μεγαλύτερη διακύμανση από άλλα πρέπει να εκπροσωπούνται από μεγαλύτερο τμήμα του δείγματος, για να αυξηθεί η ακρίβεια των εκτιμήσεων. Υποθέτοντας ότι το δειγματοληπτικό κόστος ανά μονάδα είναι το ίδιο για όλα τα στρώματα, αποδεικνύεται ότι η $V(\hat{\mu}_n)$ γίνεται ελάχιστη αν τα $n_1, n_2, \dots, n_k$ επιλεγούν έτσι ώστε

$$\boldsymbol{n_i = n\frac{N_i\sigma_i}{\sum_{j=1}^{k} N_j\sigma_j} \quad i = 1, \dots, k}$$

Ο σχεδιασμός αυτός είναι γνωστός ως βέλτιστος καταμερισμός του n με σταθερό κόστος ανά δειγματοληπτική μονάδα (optimum allocation with constant cost per unit) ή καταμερισμός κατά Neyman (Neyman allocation).

Αν το δειγματοληπτικό κόστος ανά μονάδα διαφέρει από στρώμα σε στρώμα, τότε είναι φυσικό να προσπαθήσει ο ερευνητής να αυξήσει την ακρίβεια των εκτιμήσεών του επιλέγοντας τα n_i , αντιστρόφως ανάλογα των c_i , $i = 1, 2, \dots, k$.

Έστω ότι το συνολικό κόστος c μιας δειγματοληψίας είναι συνάρτηση των

c_i , $i = 1, 2, \dots, k$ δηλαδή, έστω ότι

$$c = c_0 + \sum_{i=1}^{k} n_i c_i \ , \qquad c_0 > 0$$

Τότε, αποδεικνύεται ότι αν το κόστος c έχει μια δοθείσα τιμή, οι τιμές των n_i που ελαχιστοποιούν την διασπορά n $V(\hat{\mu})$ δίνονται από τον τύπο

$$\boldsymbol{n_i = n\frac{N_i\sigma_i/\sqrt{c_i}}{\sum_{j=1}^{k} N_j\sigma_j/\sqrt{c_i}} \quad i = 1, \dots, k}$$

Εκτίμηση παραμέτρων στην Στρωματοποιημένη Δειγματοληψία

Το δείγμα κάθε στρώματος έχει δικό του μέσο και διασπορά, όπως φαίνονται στον πίνακα.

	ΠΛΗΘΥΣΜΟΣ			**ΔΕΙΓΜΑ**		
Στρώμα	**ΜΕΓΕΘΟΣ ΣΤΡΩΜΑΤΟΣ**	**ΜΕΣΗ ΤΙΜΗ**	**ΔΙΑΣΠΟΡΑ**	**ΜΕΓΕΘΟΣ ΔΕΙΓΜΑΤΟΣ**	**ΜΕΣΟΣ**	**ΔΙΑΣΠΟΡΑ**
1	N_1	μ_1	${\sigma_1}^2$	n_1	$\bar{X}_{n_1}$	$\frac{{\sigma_1}^2}{n_1}\left(1-\frac{n_1}{N_1}\right)$
. . .						
$\boldsymbol{i}$	N_i	μ_i	${\sigma_i}^2$	n_i	$\bar{X}_{n_i}$	$\frac{{\sigma_i}^2}{n_i}\left(1-\frac{n_i}{N_i}\right)$
. . .						
k	N_k	μ_k	${\sigma_k}^2$	n_k	$\bar{X}_{n_k}$	$\frac{{\sigma_k}^2}{n_k}\left(1-\frac{n_k}{N_k}\right)$

1. Είναι προφανές ότι η μέση τιμή μ του πληθυσμού και οι μέσες τιμές $\mu_1,\ \mu_2, \dots,\ \mu_k$ των υποπληθυσμών συνδέονται με την σχέση

$$\mu = \frac{1}{N}\sum_{i=1}^{k} N_i \mu_i$$

Αν $\bar{X}_n$ ο μέσος του συνολικού δείγματος, τότε τα αντίστοιχα δειγματικά μεγέθη συνδέονται με μια παρόμοια σχέση:

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{k} n_i \bar{X}_{n_i}$$

Στην πράξη όμως, η στατιστική συνάρτηση που χρησιμοποιείται ως εκτιμήτρια του μ δεν είναι η $\bar{X}_n$, αλλά η

$$\hat{\mu}_n = \frac{1}{N}\sum_{i=1}^{k} N_i \bar{X}_{n_i}$$

στην οποία οι μέσοι των δειγμάτων από τα διάφορα στρώματα σταθμίζονται με τους συντελεστές βαρύτητας $\frac{Ni}{N}, i = 1, 2, \dots, k$ των στρωμάτων.
Προφανώς,

$$\hat{\mu}_n = \bar{X}_n \ \alpha\nu \ \frac{n_i}{n} = \frac{N_i}{N} \quad i = 1, 2, \dots, k$$

2. *Εκτίμηση του μέσου*

Η στατιστική συνάρτηση $\hat{\mu}_n$ είναι αμερόληπτη εκτιμήτρια της μέσης τιμής μ του πληθυσμού με διασπορά

$$V(\hat{\mu}_n) = \sum_{i=1}^{k} \left(\frac{N_i}{N}\right)^2 \frac{{\sigma_i}^2}{n_i}\left(1 - \frac{n_i}{N_i}\right)$$

Απόδειξη:

Για την αμεροληψία, ισχύει ότι:

$$E(\hat{\mu}_n) = \frac{1}{N}\sum_{i=1}^{k} N_i E(\bar{X}_{n_i})$$

Οι αριθμοί $\bar{X}_{n_i}$ είναι μέσοι των απλών τυχαίων δειγμάτων κάθε στρώματος και είναι αμερόληπτες εκτιμήτριες των μ_i . Άρα ισχύει ότι:

$$E\left(\bar{X}_{n_i}\right) = \mu_i$$

Συνεπώς

$$E(\hat{\mu}_n) = \frac{1}{N}\sum_{i=1}^{k} N_i\, \mu_i = \mu$$

Άρα η $\hat{\mu}_n$ είναι αμερόληπτη εκτιμήτρια του μ.
Για την διασπορά, έχουμε:

$$V(\hat{\mu}_n) = V\left(\frac{1}{N}\sum_{i=1}^{k} N_i \bar{X}_{n_i}\right)$$

$$V(\hat{\mu}_n) = \frac{1}{N^2}\left\{\sum_{i=1}^{k} {N_i}^2 V(\bar{X}_{n_i}) + 2\sum_{1\le i<j\le k} N_i N_j Cov(\bar{X}_{n_i}, \bar{X}_{n_j})\right\}$$

Όμως $Cov\left(\bar{X}_{n_i}, \bar{X}_{n_j}\right) = 0$ διότι τα τυχαία δείγματα είναι ανεξάρτητα μεταξύ τους. Συνεπώς η παραπάνω σχέση γράφεται:

$$V(\hat{\mu}_n) = \frac{1}{N^2}\sum_{i=1}^{k} {N_i}^2 V\left(\bar{X}_{n_i}\right) = \frac{1}{N^2}\sum_{i=1}^{k} {N_i}^2 \frac{{\sigma_i}^2}{n_i}\left(1 - \frac{n_i}{N_i}\right).$$

3. *Εκτίμηση και διασπορά συνολικού μεγέθους*

Η στατιστική συνάρτηση

$$\hat{Y}_{st} = N\hat{\mu}_n$$

Είναι αμερόληπτη εκτιμήτρια του συνολικού μεγέθους.

$$y = \sum_{i=1}^{k}\sum_{j=1}^{k} y_{ij}$$

Με διασπορά

$$V(\hat{Y}_{st}) = \sum_{i=1}^{k} {N_i}^2 \frac{{\sigma_i}^2}{n_i}\left(1 - \frac{n_i}{N_i}\right).$$

4. *Εκτίμηση διασποράς της εκτιμήτριας του μέσου.*

Η στατιστική συνάρτηση

$${s_{\hat{\mu}_n}}^2 = \sum_{i=1}^{k} \left(\frac{N_i}{N}\right)^2 \frac{{s_i}^2}{n_i}\left(1 - \frac{n_i}{N_i}\right)$$

Είναι αμερόληπτη εκτιμήτρια της $V(\hat{\mu}_n)$

5. *Εκτίμηση ποσοστού*

Εκτιμητής ποσοστού στο στρώμα i

$$\hat{p}_i = \frac{a_i}{n_i}$$

Εκτιμητής πληθυσμιακού ποσοστού

$$\hat{p}_{st} = \sum_{i=1}^{k} W_i \hat{p}_i \quad \text{όπου} \quad W_i = \frac{N_i}{N}$$

Διασπορά ποσοστού

$$V(\hat{p}_{st}) = \sum_{i=1}^{k} \left(\frac{N_i}{N}\right)^2 \frac{1-f_i}{n_i} \frac{N_i p_i (1-p_i)}{N_i - 1}$$

Εκτίμηση της διασποράς του ποσοστού

$$\hat{Var}(\hat{p}_{st}) = \sum_{i=1}^{k} \left(\frac{N_i}{N}\right)^2 \frac{1-f_i}{n_i - 1} \hat{p}_i (1-\hat{p}_i)$$

Τυπικό σφάλμα ποσοστού

$$se(\hat{p}_{st}) = \sqrt{\sum_{i=1}^{k} \left(\frac{N_i}{N}\right)^2 \frac{1-f_i}{n_i} \frac{N_i p_i (1-p_i)}{N_i - 1}}$$

6. *Εκτίμηση συνόλου μελών*

Εκτιμητής συνόλου μελών

$$\hat{A}_{st} = \sum_{i=1}^{k} N_i \hat{p}_i$$

Διασπορά εκτιμητή του A

$$Var(\hat{A}_{st}) = \sum_{i=1}^{k} {N_i}^2 \frac{1-f_i}{n_i} \frac{N_i p_i (1-p_i)}{N_i - 1}$$

Εκτίμηση διασποράς

$$\hat{Var}(\hat{A}_{st}) = \sum_{i=1}^{k} {N_i}^2 \frac{1-f_i}{n_i - 1} \hat{p}_i (1-\hat{p}_i)$$

Τυπικό σφάλμα εκτιμητή του A

$$se(\widehat{A}_{st}) = \sqrt{\sum_{i=1}^{k} {N_i}^2 \frac{1-f_i}{n_i} \frac{N_i p_i (1-p_i)}{N_i - 1}}$$

Προσδιορισμός μεγέθους του δείγματος

Αν το μέγεθος του δείγματος δεν είναι γνωστό, τότε η διαδικασία που ακολουθούμε είναι αντίστοιχη με αυτήν της απλής τυχαίας δειγματοληψίας. Προσπαθόντας να τηρήσουμε κάποιες απαραίτητες προδιαγραφές, δημιουργούμε τις κατάλληλες εκφράσεις οι οποίες μας οδηγούν στον προσδιορισμό του n.

Έστω ότι η απαραίτητη προϋπόθεση για τον προσδιορισμό του μεγέθους είναι η εκτίμηση του μέσου $\bar{Y}$ από τον δειγματικό $\bar{X}_{st}$ να έχουν επιτρεπτή διαδορά d με πιθανότητα σφάλματος α. Αν υποθέσουμε κανονική κατανομή, τότε ισχύει:

$$\frac{d}{\sqrt{Var(\bar{X}_{st})}} = z_{a/2}$$

(Thompson, S. K. 2012)

Όπως έχουμε αναφέρει στον καταμερισμό Neyman τα μεγέθη του δείγματος ανά στρώμα, δίνονται από την σχέση:

$$n_i = n \frac{N_i \sigma_i}{\sum_{j=1}^{k} N_j \sigma_j} \quad i = 1, \dots, k$$

Σε αυτή την περίπτωση η τιμή της ελάχιστης διακύμανσης του εκτιμητή $\bar{X}_{st}$ δίνεται από τη σχέση:

$$Var(\bar{X}_{st}) = \frac{1}{n} \left(\sum_{i=1}^{k} W_i \sigma_i \right)^2 - \frac{1}{N} \sum_{i=1}^{k} W_i {\sigma_i}^2$$

Αντικαθιστώντας την διακύμανση στον τύπο της απαραίτητης προϋπόθεσης, προκύπτει ότι:

$$n = \frac{\left(\sum_{i=1}^{k} W_i \sigma_i\right)^2}{\frac{1}{N}\sum_{i=1}^{k} W_i {\sigma_i}^2 + \left(\frac{d}{z_{a/2}}\right)^2}$$

Σύγκριση διακυμάνσεων απλής τυχαίας, αναλογικής στρωματοποιημένης και βέλτιστης στρωματοποιημένης.

Πρόταση: Αν $Var(\bar{X})$, $Var_{opt}(\bar{X}_{st})$, $Var_{prop}(\bar{X}_{st})$ οι διακυμάνσεις που προκύπτου από απλό τυχαίο δείγμα, από βέλτιστο στρωματοποιημένο και αναλογικό στρωματοποιημένο, τότε ισχύει ότι:

$$Var_{opt}(\bar{X}_{st}) \leq Var_{prop}(\bar{X}_{st}) \leq Var(\bar{X})$$

Απόδειξη: Από τον τύπο της διασποράς έχουμε:

$$\boldsymbol{(N-1)\sigma^2 = \sum_{h=1}^{k}\sum_{j=1}^{N_h}(Y_{hj} - \bar{Y})^2}$$

Προσθαφαιρώ $\bar{Y}_h$

$$\boldsymbol{(N-1)\sigma^2 = \sum_{h=1}^{k}\sum_{j=1}^{N_h}(Y_{hj} - \bar{Y}_h + \bar{Y}_h - \bar{Y})^2}$$

$$\boldsymbol{= \sum_{h=1}^{k}\sum_{j=1}^{N_h}(Y_{hj} - \bar{Y}_h)^2 + \sum_{h=1}^{k}\sum_{j=1}^{N_h}(Y_h - \bar{Y})^2 + 2\sum_{h=1}^{k}\sum_{j=1}^{N_h}(Y_{hj} - \bar{Y}_h)(Y_h - \bar{Y})}$$

$$\boldsymbol{= \sum_{h=1}^{k}(N_h - 1)\,{\sigma_h}^2 + \sum_{h=1}^{k} N_h\,(Y_h - \bar{Y})^2 + 2\sum_{h=1}^{k}(Y_h - \bar{Y})\sum_{h=1}^{N_h}(Y_{hj} - \bar{Y}_h)}$$

Από το οποίο προκύπτει ότι:

$$\boldsymbol{(N-1)\sigma^2 = \sum_{h=1}^{k}(N_h - 1)\,{\sigma_h}^2 + \sum_{h=1}^{k} N_h\,(Y_h - \bar{Y})^2}$$

Αν θεωρήσουμε ότι οι όροι $^1/_{N_h}$ είναι αμελητέοι προκύπτει ο τύπος:

$$\boldsymbol{\sigma^2 = \sum_{h=1}^{k} W_h \sigma_h{}^2 + \sum_{h=1}^{k} W_h (Y_h - \bar{Y})^2}$$

Πολλαπλασιάζω και τα 2 μέλη με $\frac{1-f}{n}$

$$\frac{1-f}{n}\boldsymbol{\sigma^2} = \frac{1-f}{n}\sum_{h=1}^{k} \boldsymbol{W_h \sigma_h{}^2} + \frac{1-f}{n}\sum_{h=1}^{k} \boldsymbol{W_h (Y_h - \bar{Y})^2}$$

Δηλαδή

$$Var(\bar{X}) = Var_{prop}(\bar{X}_{st}) + \frac{1-f}{n}\sum_{h=1}^{k} \boldsymbol{W_h (Y_h - \bar{Y})^2}$$

Συνεπώς

$$Var(\bar{X}) \geq Var_{prop}(\bar{X}_{st}).$$

(Thompson, S. K. 2012)

Παρατήρηση: Η στρωματοποιημένη δειγματοληψία σε κάθε περίπτωση βελτιώνει την απλη τυχαία δειγματοληψία.

4.3 Συστηματική Δειγματοληψία

Η μέθοδος της συστηματικής δειγματοληψίας χρησιμοποιείται κυρίως επειδή είναι εύκολη και απλή κατά την υλοποίησή της από τον ερευνητή. Ένα πλεονέκτημα της μεθόδου αυτής είναι ότι το δείγμα που επιλέγει ο ερευνητής είναι ομοιόμορφα κατανεμημένο ως προς τον πληθυσμό του, καθώς οι μονάδες που επιλέγονται απέχουν την ίδια απόσταση μεταξύ τους.

Η συστηματική δειγματοληψία είναι μια στρατηγική που μοιάζει απλή και λογική και επιλέγεται συχνά αυθόρμητα από μελετητές.. Η αυθόρμητη αυτή προτίμηση οφείλεται στο ότι διευκολύνει την επιλογή των δειγματοληπτικών μονάδων λόγω του απλού της πρωτοκόλου.
Το μέγεθος του δείγματος που επιθυμούμε να μελετήσουμε καθορίζει και τον τρόπο επιλογής των ατόμων. Έτσι άν θέλουμε να σχηματίσουμε ένα δείγμα που να αντιπροσωπεύει το 1/p του πληθυσμού, αρκεί από την λίστα των ατόμων του πληθυσμού ή από την σειρά προσέλευσης, ή εμφάνισης τους να επιλέγουμε ένα κάθε p άτομα (p=N/n είναι το βήμα της δειγματοληψίας, N τα άτομα του πληθυσμού και ν τα του δείγματος).

Για την επιλογή δειγματων με συστηματικό τρόπο από πληθυσμούς που κατανέμονται στο χώρο και για τους οποίους δεν υπάρχει κατάλογος των ατόμων (σχεδόν όλοι οι φυσικοί πληθυσμοί) ακολουθούμε την εξής διαδικασία. Χωρίζουμε τον πληθυσμό μας (ή την έκταση που καλύπτει αυτός) σε ισομεγέθεις και ιδίου σχήματος πρωτογενείς μονάδες. Στη συνέχεια κάθε πρωτογενής μονάδα χωρίζεται με το ίδιο τρόπο σε p δευτερογενείς. Από τίς p μονάδες μιάς πρωτογενούς μονάδος επιλέγεται τυχαία μια (χρήση υπολογιστή, πίνακας τυχαίων αριθμών κ.λ.π.) και στη συνέχεια σε κάθε πρωτογενή μονάδα επιλέγεται η δευρογενής με τις ίδιες συντεταγμένες. Το σύνολο των δευτερογενών αυτών μονάδων αποτελεί ένα συστηματικό δείγμα ίσο με το 1/p του πληθυσμού στόχου. Είναι προφανές ότι το μέγεθος και κατά συνέπεια ο αριθμός των πρωτογενών μονάδων καθορίζει και τα χαρακτηριστικά του δείγματος. Μικρές και πολυπληθείς πρωτογενείς μονάδες οδηγούν σε μια καλύτερη "κάλυψη" του πληθυσμού. Καλύτερη κάλυψη σημαίνει ότι διατρέχουμε όλον τον πληθυσμό με λεπτομερή τρόπο και δεν αφήνουμε ακάλυπτες μεγάλες περιοχές.

Η συστηματική δειγματοληψία είναι εξαιρετικά μεροληπτική στην περίπτωση που ο πληθυσμός έχει περιοδικά χαρακτηριστικά. Φυσικά η μεγάλη μεροληψία είναι ανεπιθύμητη. Το δεύτερο πρόβλημα προέρχεται από το γεγονός ότι το συστηματικό δείγμα δεν είναι ένα τυχαίο δείγμα του πληθυσμού. Αυτό δυσχεραίνει την εκτίμηση βασικών παραμέτρων και κυρίως της διασποράς. Κατ'αρχήν από τον τρόπο επιλογής των ατόμων του δείγματος φαίνεται ότι πρόκειται για μια ειδική περίπτωση δισταδιακής δειγματοληψίας (πρωτογενείς μονάδες που περιέχουν δευτερογενείς από τις οποίες επιλέγεται μία). Από θεωρητικής πλευράς η συστηματική δειγματοληψία παραβιάζει βασικές συνθήκες της τυχαίας επιλογής.
Δεν επιτρέπει την επιλογή ατόμων με πιθανότητες που να σχετίζονται με την σχετική τους σπουδαιότητα και το ρόλο τους μέσα στον πληθυσμού.

Στη φύση τα φαινόμενα με περιοδικό χαρακτήρα είναι αρκετά συχνά (παλίρροια, ημερονύκτιες και εποχιακές αλλαγές κ.λ.π.) γι'αυτό και η επιλογή της συστηματικής δειγματοληψίας καθώς και του βήματος της πρέπει να γίνονται με μεγάλη προσοχή. Εάν τα αποτελέσματα μιας συστηματικής δειγματοληψίας δείχνουν ασυνήθιστα υψηλή ομοιογένεια τότε μια προσεκτική ανάλυση των χαρακτηριστικών του πληθυσμού σε σχέση με το βήμα της δειγματοληψίας επιβάλεται.

Ορισμοί

Η διαδικασία υλοποίησης της συστηματικής δειγματοληψίας έχει ως εξής:

1. Κατασκευή του δειγματοληπτικού πλαισίου του πληθυσμού. Έστω ότι $\{Y_1, Y_2, \dots, Y_N\}$ οι μονάδες του πληθυσμού που ανήκουν στο δειγματοληπτικό πλαίσιο.
2. Υπολογισμός του βήματος k της δειγματοληψίας. Το βήμα είναι συνήθως $k = \frac{N}{n}$.
3. Επιλέγεται τυχαία ένας αριθμός έστω i όπου $0 \leq i \leq k$. Η μονάδα του πληθυσμού Y_i που αντιστοιχεί στον αριθμό i λέγεται αφετηρία της συστηματικής δειγματοληψίας.
4. Το συστηματικό δείγμα αποτελείται από τις μονάδες:

$$\{Y_i, Y_{i+k}, Y_{i+2k}, \dots, Y_{i+(n-1)k}\}$$

Παρατηρήσεις:

- Στην περίπτωση που το κλάσμα $\frac{N}{n}$ είναι δεκαδικός αριθμός, τότε ως βήμα k λαμβάνεται ο πλησιέστερος ακέραιος αριθμός, με όλες τις επιπτώσεις στο μέγεθος του δείγματος. Σε αυτή την περίπτωση το μέγεθος του δείγματος ενδέχεται να βγει μεγαλύτερο ή ικρότερο του n.
- Ανάλογα με την επιλογή του αριθμού i, υπάρχουν k πιθανά δείγματα, τα οποία ονομάζονται συστηματικά δείγματα.
- Τα δυνατά δείγματα στη συστηματική δειγματοληψία δεν έχουν κανένα κοινό στοιχείο μεταξύ τους.

Τυχαία συστηματική δειγματοληψία: Ο αριθμός i όπου $0 \leq i \leq k$ για την αφετηρία του συστηματικού δείγματος επιλέγεται τυχαία.

Κεντρικά τοποθετημένη συστηματική δειγματοληψία: Εάν αντί της τυχαίας επιλογής του αριθμού i όπου $0 \leq i \leq k$ επιλέξουμε το κεντρικό σημείο του πρώτου διαστήματος μήκους k, τότε η δειγματοληψία ονομάζεται κεντρικά τοποθετημένη συστηματική.

Εκτίμηση παραμέτρων στη συστηματική δειγματοληψία

Έστω ένα τυχαίο δείγμα που έχει προκύψει με τυχαία συστηματική δειγματοληψία.

$$\{X_1, X_2, X_3, \dots, X_n\} = \{Y_i, Y_{i+k}, Y_{i+2k}, \dots, Y_{i+(n-1)k}\}$$

1. ΕΚΤΙΜΗΣΗ ΤΟΥ ΜΕΣΟΥ

Ο εκτιμητής του πληθυσμιακού μέσου $\bar{Y}$ είναι ο δειγματικός μέσος:

$$\bar{X}_{sy} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Ο οποίος είναι αμερόληπτος όταν το μέγεθος του πληθυσμού N διαιρείται ακριβώς με το μέγεθος του δείγματος n.

Απόδειξη:

Ο εκτιμητής

$$\bar{X}_{sy} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

είναι αμερόληπτος εάν ισχύει ότι:

$$E(\bar{X}_{sy}) = \bar{Y}\,.$$

Από τον ορισμό, γνωρίζουμε ότι:

$$E(\bar{X}_{sy}) = \sum_{s \in S} \bar{X}_{sy}^{(s)}\, \pi(s)$$

Όπου s ένα από τα δυνατά δείγματα του S , $\bar{X}_{sy}^{(s)}$ η τιμή του εκτιμητή για το δείγμα s , και $\pi(s)$ η πιθανότητα επιλογής του δείγματος s , η οποία είναι ίση με $1/k$

Οι δειγματικοί μέσοι των k συστηματικών δειγμάτων είναι:

$s_1 = (Y_1, Y_{1+k}, \dots, Y_{1+(n-1)k})$ για τον τυχαίο αριθμό $j = 1$ με μέσο $\bar{X}_{sy}^{(s_1)} \dfrac{1}{n} \displaystyle\sum_{i=1}^{n} Y_{1+(i-1)k}$

$s_2 = (Y_2, Y_{2+k}, \dots, Y_{2+(n-1)k})$ για τον τυχαίο αριθμό $j = 2$ με μέσο $\bar{X}_{sy}^{(s_2)} \dfrac{1}{n} \displaystyle\sum_{i=1}^{n} Y_{2+(i-1)k}$

…

$s_k = (Y_k, Y_{2k}, \dots, Y_{nk})$ για τον τυχαίο αριθμό $j = k$ με μέσο $\bar{X}_{sy}^{(s_k)} \dfrac{1}{n} \displaystyle\sum_{i=1}^{n} Y_{ik}$

Συνεπώς έχουμε:

$$E\left(\bar{X}_{sy}\right) = \sum_{s \in S} \bar{X}_{sy}^{(s)}\, \pi(s) = \sum_{j=1}^{k} \frac{1}{k} \bar{X}_{sy}^{(s_j)} = \sum_{j=1}^{k} \frac{n}{N} \frac{1}{n} \sum_{i=1}^{n} Y_{j+(i-1)k}$$

$$= \frac{1}{N} \sum_{j=1}^{k} \sum_{i=1}^{n} Y_{j+(i-1)k} = \bar{Y}$$

Τέλος απόδειξης.

2. *ΔΙΑΣΠΟΡΑ ΤΟΥ ΕΚΤΙΜΗΤΗ* $\bar{X}_{sy}$

Η διασπορά του εκτιμητή του πληθυσμιακού μέσου, δίνεται από τη σχέση:

$$\boldsymbol{Var(\bar{X}_{sy}) = \frac{N-1}{N}\sigma^2 - \frac{k(n-1)}{N}{\sigma_{wsy}}^2}$$

Όπου

$$\boldsymbol{{\sigma_{wsy}}^2 = \frac{1}{k(n-1)}\sum_{j=1}^{k}\sum_{i=1}^{n}(Y_{ji} - \bar{Y}_j)^2}$$

Και $\boldsymbol{\bar{Y}_j}$ ο μέσος του j συστηματικού δείγματος.

Απόδειξη:

$$\boldsymbol{Var(\bar{X}_{sy}) = E(\bar{X}_{sy} - \bar{Y})^2}$$

η οποία υπολογίζεται σύμφωνα με το σύνολο των δυνατών δειγμάτων και των πιθανοτήτων $\pi(s)$. Σύμφωνα με τη συστηματική δειγματοληψία θα είναι:

$$\boldsymbol{Var(\bar{X}_{sy}) = \frac{1}{k}\sum_{i=1}^{k}\left(\bar{X}_{sy}^{(s_i)} - \bar{Y}\right)^2}$$

Όπως έχουμε δει και στην στρωματοποιημένη δειγματοληψία, ισχύει η σχέση:

$$\boldsymbol{(N-1)\sigma^2 = \sum_{i=1}^{k}\sum_{j=1}^{n}\left(Y_{ij} - \bar{Y}_j\right)^2 + n\sum_{i=1}^{k}(\bar{Y}_i - \bar{Y})^2}$$

$$\boldsymbol{\frac{(N-1)}{N}\sigma^2 = \frac{k(n-1)}{N}{\sigma_{wsy}}^2 + \frac{n}{N}\sum_{i=1}^{k}(\bar{Y}_i - \bar{Y})^2}$$

$$\boldsymbol{\frac{(N-1)}{N}\sigma^2 = \frac{k(n-1)}{N}{\sigma_{wsy}}^2 + Var\left(\bar{X}_{sy}\right)}$$

Τέλος απόδειξης.

3. ΤΥΠΙΚΟ ΣΦΑΛΜΑ ΤΟΥ ΕΚΤΙΜΗΤΗ ΤΟΥ ΜΕΣΟΥ

Το τυπικό σφάλμα του εκτιμητή $\boldsymbol{\bar{X}_{sy}}$ δίνεται από τη σχέση:

$$\boldsymbol{se(\bar{X}_{sy}) = \sqrt{\frac{N-1}{N}\sigma^2 - \frac{k(n-1)}{N}{\sigma_{wsy}}^2}}$$

4. ΕΚΤΙΜΗΣΗ ΤΗΣ ΔΙΑΚΥΜΑΝΣΗΣ

Για να εκτιμηθεί η διακύμανση του μέσου στη συστηματική δειγματοληψία χρειάζεται να εκτιμηθούν δύο διασπορές που εμφανίζονται στον τύπο της, δηλαδή η συνολική πληθυσμιακή διασπορά $\boldsymbol{\sigma^2}$ και η μέση διασπορά των συστηματικών δειγμάτων στο εσωτερικό τους $\boldsymbol{\sigma_{wsy}{}^2}$. Η $\boldsymbol{\sigma^2}$ μπορεί να εκτιμηθεί από την αντίστοιχη δειγματική διακύμανση $\boldsymbol{s^2}$ αλλά η $\boldsymbol{\sigma_{wsy}{}^2}$ δεν μπορεί να εκτιμηθεί διότι είναι ένας μέσος όρος k τιμών $(\sigma_1^2, \sigma_2^2, \dots, \sigma_k^2)$ και μέσω της συστηματικής δειγματοληψίας θα είναι ίση με μόνο μία εξ'αυτών.

Για την εκτίμηση της $\boldsymbol{\sigma_{wsy}{}^2}$ χρησιμοποιείται η μέθοδος της επαναλαμβανόμενης συστηματικής δειγματοληψίας.
Επαναλαμβανόμενη συστηματική δειγματοληψία είναι η παραγωγή πολλαπλών συστηματικών δειγμάτων, με σκοπό να είναι δυνατός ο υπολογισμός περισσότερων από μιας διακύμανσης $\boldsymbol{\sigma_i{}^2}$.

Τα βήματα της μεθόδου είναι τα εξής:

- Αν k το βήμα της συστηματικής δειγματοληψίας, τότε επιλέγουμε έναν αριθμό m σχετικό με τον k, ο οποίος δηλώνει τον αριθμό των επαναλήψεων της συστηματικής δειγματοληψίας. (και ο οποίος να διαιρεί τον n)
- Ορίζουμε ως βήμα $k' = mk$ και εφαρμόζουμε m ανεξάρτητες συστηματικές δειγματοληψίες.
- Το τελικό δείγμα της επαναλαμβανόμενης συστηματικής δειγματοληψίας είναι το δείγμα που αποτελείται από τη συλλογή των μετρήσεων που θα προκύψουν από τις m διαδοχικές συστηματικές δειγματοληψίες.

Στην περίπτωση της επαναλαμβανόμενης συστηματικής δειγματοληψίας, ο εκτιμητής του μέσου υπολογίζεται από τον μέσο των μέσων των επιμέρους συστηματικών, δηλαδή

$$\bar{X}_{sy,rep} = \bar{X}$$

Πλεονεκτήματα της επαναλαμβανόμενης συστηματικής δειγματοληψίας:

- Καλύτερη κάλυψη του πληθυσμού
- Αποφυγή τυχόν περιοδικότητας του πληθυσμού ως προς το υπό μελέτη χαρακτηριστικό.
- Εκτίμηση της διακύμανσης $\boldsymbol{\sigma_{wsy}{}^2}$ και κατά συνέπεια της διακύμανσης του εκτιμητή $\boldsymbol{\bar{X}_{sy}}$ βάση ενός δείγματος μεγέθους n.
  (Κοκολάκης Γ., Σπηλιώτης Ι., 1999)

Η εκτίμηση της $\boldsymbol{\sigma_{wsy}{}^2}$ με τη βοήθεια της επαναλαμβανόμενης συστηματικής δειγματοληψίας, δίνεται από τον τύπο:

$$\widehat{\boldsymbol{\sigma}}_{\boldsymbol{wsy}}{}^{\boldsymbol{2}} = \frac{\boldsymbol{1}}{\boldsymbol{m}}\sum_{i=1}^{m}\boldsymbol{\sigma}_{\boldsymbol{i}}^{\boldsymbol{2}}$$

Για την εκτίμηση της $\boldsymbol{Var}(\overline{\boldsymbol{X}}_{\boldsymbol{sy}})$ επειδή η επιλογή των m δειγμαάτων έγινε με απλή τυχαία δειγματοληψία, δίνεται ο τύπος:

$$\boldsymbol{Var}(\overline{\boldsymbol{X}}_{\boldsymbol{sy}}) = \left(\boldsymbol{1} - \frac{\boldsymbol{m}}{\boldsymbol{k'}}\right)\frac{\boldsymbol{1}}{\boldsymbol{m}}\left[\frac{\boldsymbol{1}}{\boldsymbol{m}-\boldsymbol{1}}\sum_{\boldsymbol{i}=\boldsymbol{1}}^{\boldsymbol{m}}\left(\overline{\boldsymbol{X}}^{(\boldsymbol{s_i})} - \overline{\boldsymbol{X}}\right)^{\boldsymbol{2}}\right]$$

Όπου $\overline{\boldsymbol{X}}^{(\boldsymbol{s_i})}$ για $i = 1,2,\dots,m$ είναι η μέση δειγματική τιμή του i συστηματικού δείγματος από τα m που επιλέγονται.

Συνεπώς οι εκτιμήτριες της διακύμανσης και του τυπικού σφάλματος είναι:

$$\widehat{\boldsymbol{Var}}(\overline{\boldsymbol{X}}_{\boldsymbol{sy}}) = \left(\boldsymbol{1} - \frac{\boldsymbol{m}}{\boldsymbol{k'}}\right)\frac{\boldsymbol{1}}{\boldsymbol{m}}\left[\frac{\boldsymbol{1}}{\boldsymbol{m}-\boldsymbol{1}}\sum_{\boldsymbol{i}=\boldsymbol{1}}^{\boldsymbol{m}}\left(\overline{\boldsymbol{X}}^{(\boldsymbol{s_i})} - \overline{\boldsymbol{X}}\right)^{\boldsymbol{2}}\right]$$

Και

$$\widehat{\boldsymbol{se}}(\overline{\boldsymbol{X}}_{\boldsymbol{sy}}) = \sqrt{\left(\boldsymbol{1} - \frac{\boldsymbol{m}}{\boldsymbol{k'}}\right)\frac{\boldsymbol{1}}{\boldsymbol{m}}\left[\frac{\boldsymbol{1}}{\boldsymbol{m}-\boldsymbol{1}}\sum_{\boldsymbol{i}=\boldsymbol{1}}^{\boldsymbol{m}}(\overline{\boldsymbol{X}}^{(\boldsymbol{s_i})} - \overline{\boldsymbol{X}})^{\boldsymbol{2}}\right]}$$

5. ΕΚΤΙΜΗΣΗ ΣΥΝΟΛΟΥ

Επειδή ισχύει ότι $Y_T = N\bar{Y}$ για την εκτίμηση του συνόλου θα έχουμε:

$$\widehat{\boldsymbol{Y}}_{\boldsymbol{T,sy}} = \boldsymbol{N}\,\overline{\boldsymbol{X}}_{\boldsymbol{sy}}$$

Συνεπώς οι ιδιότητες του $\widehat{\boldsymbol{Y}}_{\boldsymbol{T,sy}}$ είναι άμεσες συνέπειες των ιδιοτήτων του $\overline{\boldsymbol{X}}_{\boldsymbol{sy}}$.

6. ΕΚΤΙΜΗΣΗ ΠΟΣΟΣΤΟΥ

Η εκτιμήτρια του ποσοστού δίνεται από τον τύπο:

$$\hat{p}_{sy} = p_{sy}$$

Όπου p_{sy} το ποσοστό στο συστηματικό δείγμα που επιλέχθηκε.

Η διακύμανση του εκτιμητή του ποσοστού, είναι:

$$\boldsymbol{Var\left(\hat{p}_{sy}\right) = p(1-p) - \frac{k(n-1)}{N}\sigma_{wsy}{}^{2}}$$

όπου

$$\boldsymbol{\sigma_{wsy}{}^{2} = \frac{1}{k}\sum_{i=1}^{k}\frac{np_i(1-p_i)}{n-1}}$$

Όπου $\boldsymbol{p_i}$ το ποσοστό στο i συστηματικό δείγμα.

Η εκτιμώμενη διακύμανση του ποσοστού δίνεται από τον τύπο:

$$\boldsymbol{\hat{Var}\left(\hat{p}_{sy}\right) = \left(1-\frac{1}{k}\right)\frac{1}{m}\left[\frac{1}{m-1}\sum_{i=1}^{m}\left(p^{(s_i)} - \overline{p}\right)^{2}\right]}$$

Όπου $\boldsymbol{p^{(s_i)}}$ το ποσοστό στο i επαναληπτικό συστηματικό δείγμα και $\boldsymbol{\overline{p}}$ ο μέσος των $\boldsymbol{p^{(s_i)}}$ για τα διάφορα i.

Τυπικό σφάλμα εκτίμησης διασποράς ποσοστού:

$$\boldsymbol{se\left(\hat{p}_{sy}\right) = \sqrt{Var\left(\hat{p}_{sy}\right)}}$$

Εκτίμηση τυπικού σφάλματος:

$$\boldsymbol{\widehat{se}\left(\hat{p}_{sy}\right) = \sqrt{\hat{Var}\left(\hat{p}_{sy}\right)}}$$

Συντελεστής συσχέτισης

Ένας διαφορετικός τρόπος για την εκτίμηση της διασποράς του εκτιμητή του μέσου είναι ο συντελεστής συσχέτισης. Ορίζεται ως ο συντελεστής συσχέτισης ανά δύο των στοιχείων που ανήκουν στο ίδιο συστηματικό δείγμα. Δηλαδή:

$$\rho_w = \frac{2}{(n-1)(N-1)\sigma^2}\sum_{i=1}^{k}\left[\sum_{j=1}^{n}\sum_{\substack{u=1 \\ u>j}}^{n}(X_{ij}-\bar{Y})(X_{iu}-\bar{Y})\right]$$

Όπου:

X_{ij} η μονάδα του δείγματος που βρίσκεται στο i συστηματικό δείγμα και ανήκει στην j θέση,

$\bar{Y}$ η πληθυσμιακή μέση τιμή και

σ^2 η πληθυσμιακή διασπορά.

Ο ρ_w δίνει αντίστοιχη πληροφορία με το $\boldsymbol{\sigma_{wsy}}^{\boldsymbol{2}}$. Και τα δύο δηλώνουν τη θέση μεταξύ των μονάδων που ανήκουν στο ίδιο συστηματικό δείγμα. Όσο μεγαλύτερος ο συντελεστής συσχέτισης ρ_w τόσο μικρότερη αναμένεται να είναι η διασπορά $\boldsymbol{\sigma_{wsy}}^{\boldsymbol{2}}$**.**

Ένας διαφορετικός τύπος της διακύμανσης του εκτιμητή του μέσου, με βάση τον συντελεστή συσχέτισης, είναι ο εξής:

$$\boldsymbol{Var(\bar{X}_{sy}) = \frac{\sigma^2(N-1)}{nN}[1+(n-1)\rho_w]}$$

Παρατήρηση:

- Αν $\boldsymbol{\rho_w = 0}$ δηλαδή οι μονάδες που ανήκουν στο ίδιο δείγμα είναι ασυσχέτιστες, τότε

$$\boldsymbol{Var(\bar{X}_{sy}) = \frac{\sigma^2(N-1)}{nN} \cong \frac{\sigma^2}{n}}$$

που είναι η διακύμανση του εκτιμητή του μέσου στην απλή τυχαία δειγματοληψία.
Συνεπώς στην περίπτωση που $\boldsymbol{\rho_w = 0}$ ο υπολογισμός της διακύμανσης του εκτιμητή του μέσου μπορεί να γίνει με βάση τα αποτελέσματα της απλής τυχαίας.
Σε αυτή την περίπτωση, η συστηματική δειγματοληψία δίνει αποτελέσματα ίσης αποτελεσματικότητας με αυτά της απλής τυχαίας δειγματοληψίας.

- Αν $\boldsymbol{\rho_w > 0}$ τότε η διακύμανση του εκτιμητή $\boldsymbol{\bar{X}_{sy}}$ είναι μεγαλύτερη από την αντίστοιχη διακύμανση της απλής τυχαίας ίσου μεγέθους, συνεπώς και λιγότερο αποτελεσματική.

Σύγκριση συστηματικής με απλή τυχαία δειγματοληψία

Η συστηματική δειγματοληψία είναι πιο αποτελεσματική από την απλή τυχαία δειγματοληψία ως προς την εκτίμηση του πληθυσμιακού μέσου, αν και μόνο αν

$$\boldsymbol{{\sigma_{wsy}}^2 > \sigma^2}$$

Απόδειξη:
Υπολογίζουμε την διαφορά των διακυμάνσεων των εκτιμητών των πληθυσμιακών μέσων:

$$\boldsymbol{Var(\bar{X}) - Var(\bar{X}_{sy}) = \frac{1-\frac{n}{N}}{n}\sigma^2 - \frac{N-1}{N}\sigma^2 + \frac{k(n-1)}{N}{\sigma_{wsy}}^2}$$

$$\boldsymbol{= \left(\frac{1-\frac{n}{N}}{n} - \frac{N-1}{N}\right)\sigma^2 + \frac{k(n-1)}{N}{\sigma_{wsy}}^2}$$

Αν $N = kn$ τότε:

$$\boldsymbol{Var(\bar{X}) - Var(\bar{X}_{sy}) = \frac{n-1}{n}\left({\sigma_{wsy}}^2 - \sigma^2\right) = \frac{k(n-1)}{N}\left({\sigma_{wsy}}^2 - \sigma^2\right)}$$

Συνεπώς $\boldsymbol{Var(\bar{X}) - Var(\bar{X}_{sy}) > 0}$ άρα η συστηματική δειγματοληψία είναι πιο αποτελεσματική.

Τέλος απόδειξης.

Πλεονεκτήματα και μειονεκτήματα της συστηματικής δειγματοληψίας

Πλεονεκτήματα

- Η συστηματική δειγματοληψία είναι πολύ πιο εύκολη στην προτιμασία και την πραγματοποίησή της από τις άλλες στρατηγικές, ειδικά όταν πρόκειται για τη συλλογή πληροφορίας από πληθυσμούς που κατανέμονται σε μεγάλες εκτάσεις.
- Το πρωτόκολο συλλογής μειώνει στο ελάχιστο την πιθανότητα λάθους (π.χ. λάθος στίς συντεταγμένες σταθμού ή παράληψη παρατήρησης.
- Αποφεύγεται η μεροληπτική επιλογή ατόμων. Για παράδειγμα όταν ο ερευνητής προσπαθεί να εκτιμήσει το μέσο μέγεθος είδους ψαριού στην ιχθυόσκαλα έχει συχνά αυθόρμητα την τάση να επιλέγει για μέτρηση τα μεγαλύτερα άτομα.

- Όταν τα άτομα του πληθυσμού παρουσιάζονται με τυχαία σειρά ή κατανέμονται τυχαία στο χώρο τότε η συστηματική δειγματοληψία διατηρεί όλα τα πλεονεκτήματα της τυχαίας δειγματοληψίας αλλά επιπλέον είναι πιο ακριβής διότι οι παρατηρήσεις κατανέμονται πιο ομοιόμορφα και διατρέχουν όλον τον πληθυσμό.
- Εάν ο πληθυσμός παρουσιάζει φαινόμενα αυτοσυσχέτησης, τότε η συστηματική δειγματοληψία είναι πιο ακριβής διότι αποφεύγεται η πιθανότητα συλλογής δειγμάτων από μια συγκεκριμένη περιοχή και που φυσικά θα είναι επηρεασμένα από τα τοπικά χαρακτηριστικά. Τα φαινόμενα αυτοσυσχέτισης είναι πολύ κοινά στη φύση όπου οι κλίμακες των φαινομένων είναι μεγάλες. Για παράδειγμα δυο μετεωρολογικοί σταθμοί που βρίσκονται κοντά δίνουν πληροφορίες που μοιάζουν πολύ μεταξύ τους σε σχέση με δυο σταθμούς που απέχουν πολύ μεταξύ τους.
- Επίσης και όταν υπάρχει γραμμική διακύμανση στον πληθυσμό η συστηματική δειγματοληψία είναι πιο ακριβής σε σχέση με την απλή τυχαία για τον προαναφερθέντα λόγο.

Μειονεκτήματα

- Δέν επιτρέπει με σίγουρο τρόπο την εκτίμηση της διασποράς των εκτιμητών από τη συλλογή ενός μόνο συστηματικού δείγματος
- Εάν ο πληθυσμός παρουσιάζει περιοδικές διακυμάνσεις τότε η στρατηγική αυτή εγκυμονεί μεγάλους κινδύνους και απαιτείται μεγάλη προσοχή στην επιλογή του βήματος δειγματοληψίας.
- Δεν επιτρέπει την επιλογή ατόμων με πιθανότητες που να σχετίζονται με την σχετική τους σπουδαιότητα και το ρόλο τους μέσα στον πληθυσμού.
- Το πλάνο αυτό δεν εκμεταλλεύεται εύκολα προϋπάρχουσες πληρφορίες για τη δομή και την κατανομή του πληθυσμού.
- Εκτός από την κατανομή παρατηρήσεων στο χώρο και τον χρόνο (σταθμούς) το πρωτόκολο αυτό απαιτεί γνώση της διάταξης των ατόμων στον πληθυσμό, πράγμα δύσκολο στις περισσότερες περιπτώσεις.

4.4 Δειγματοληψία κατά συστάδες

Μία άλλη τεχνική δειγματοληψίας, πέραν αυτών που παρουσιάστηκαν, είναι η δειγματοληψία κατά συστάδες (ΔκΣ), κατά την οποία χωρίζονται τα μέλη-άτομα του πληθυσμού Π σε διάφορες ομάδες, συστάδες (clusters), οι οποίες είναι όσο το δυνατό πιο ομοιόμορφες μεταξύ τους, έστω και με ανομοιογενή στοιχεία η καθεμία.
Πολλοί είναι οι λόγοι που υπαγορεύουν να χρησιμοποιηθεί δειγματοληψία κατά συστάδες. Η δομή του πληθυσμού και η δυνατότητα πρόσβασης ή μη στα άτομα του πληθυσμού Π είναι ένας λόγος για τη χρήση των συστάδων και της ΔκΣ. Ένας άλλος λόγος είναι το κόστος που προκύπτει, αν χρησιμοποιηθεί η μέθοδος αυτή.
Κατά τη ΔκΣ ως μονάδα του πληθυσμού νοείται μία ομάδα ατόμων του που συνδέονται μεταξύ τους κατά κάποιο τρόπο (π.χ. διοικητικές μονάδες, σχολεία, τμήματα φυτώριου κ.λπ.)) και ονομάζεται «συστάδα» (cluster).

Ο πληθυσμός Π αποτελείται από Ν συστάδες και από αυτές θα συμπεριληφθούν στο δείγμα n συστάδες (units). Γενικά το πλήθος των στοιχείων στις συστάδες δεν είναι σταθερό. Η κάθε συστάδα μπορεί να περιέχει διαφορετικό αριθμό ατόμων. Περισσότερο ενδιαφέρει, τουλάχιστον στο παρόν σύγγραμμα, η περίπτωση όπου οι συστάδες είναι ισομεγέθεις και ακροθιγώς θα εξεταστεί και η γενικότερη περίπτωση. Προφανές είναι ότι οι συστάδες ενός πληθυσμού είναι δυνατόν να αποτελούνται και αυτές από άλλες μικρότερες συστάδες και να έχουμε κατά αυτό τον τρόπο τουλάχιστον δύο στάδια ΔκΣ.

Γενικότερα μπορούμε να ισχυριστούμε ότι υπάρχουν δύο είδη ΔκΣ:

(Α) Μονοσταδιακή και η
(Β) Πολυσταδιακή (με απλούστερη τη δισταδιακή).

Στη μονοσταδιακή ΔκΣ επιλέγουμε από τον κατάλογο των Ν το πλήθος συστάδων με ΑΤΔ ένα δείγμα από n το πλήθος συστάδες. Μετά την επιλογή αυτή η μελέτη αφορά πια όλα τα άτομα όλων των συστάδων που επιλέχτηκαν να συμπεριληφθούν στο δείγμα. Πάνω σε όλα τα στοιχεία όλων των συστάδων του δείγματος θα στηριχτούμε για την εξαγωγή των συμπερασμάτων μας μέσα από τη στατιστική ανάλυση.

Διαφορά στρωματοποιημένης με δειγματοληψία κατά συστάδες

- Στη στρωματοποιημένη δειγματοληψία, ο πληθυσμός χωρίζεται σε ομοιογενείς ομάδες που ονομάζονται στρώματα, χρησιμοποιώντας ένα χαρακτηριστικό των δειγμάτων. Στη συνέχεια επιλέγονται μέλη από κάθε στρώμα και ο αριθμός των δειγμάτων που λαμβάνονται από αυτά τα στρώματα είναι ανάλογος με την παρουσία των στρωμάτων στον πληθυσμό.

- Στη δειγματοληψία συστάδων, ο πληθυσμός ομαδοποιείται σε συστάδες, κυρίως με βάση την τοποθεσία και στη συνέχεια επιλέγεται τυχαία ένα σύμπλεγμα.

- Στη δειγματοληψία κατά συστάδες, ένα σύμπλεγμα επιλέγεται τυχαία, ενώ σε στρωματοποιημένα μέλη δειγματοληψίας δεν επιλέγονται τυχαία.

- Στη στρωματοποιημένη δειγματοληψία, κάθε ομάδα που χρησιμοποιείται (στρώματα) περιλαμβάνει ομοιογενή μέλη ενώ, στη δειγματοληψία συστάδων, ένα σύμπλεγμα είναι ετερογενές.

- Η στρωματοποιημένη δειγματοληψία είναι πιο αργή ενώ η δειγματοληψία συστάδων είναι σχετικά ταχύτερη.

- Σκοπός της στρωματοποίησης είναι η βελτίωση των τυπικών σφαλμάτων και των εκτιμητών, ενώ σκοπός της δημιουργίας συστάδων είναι η ευκολία στην πρόσβαση και η μείωση κόστους και χρόνου της έρευνας.

- Τα στρωματοποιημένα δείγματα έχουν μικρότερο σφάλμα λόγω του factoring με την παρουσία κάθε ομάδας εντός του πληθυσμού και την προσαρμογή των μεθόδων ώστε να επιτευχθεί καλύτερη εκτίμηση.

- Η δειγματοληψία κατά συστάδες έχει εγγενές υψηλότερο ποσοστό σφάλματος.

Συμβολισμοί

1. Αριθμός συστάδων M
2. Μέγεθος i συστάδας $\boldsymbol{K_i}$
   Ισχύει ότι:
$$K_1 + K_2 + \cdots + K_M = N$$
3. $\boldsymbol{Y_{ij}}$ η τιμή του χαρακτηριστικού Y για το j μέλος της i συστάδας.
4. $\boldsymbol{U_i} = \{\boldsymbol{Y_{i1}}, \boldsymbol{Y_{i2}}, \dots, \boldsymbol{Y_{iK_i}}\}$ τα μέλη της i συστάδας.
5. Ο μέσος της i συστάδας για το χαρακτηριστικό Y
$$\bar{\boldsymbol{U}}_{\boldsymbol{i}} = \frac{\mathbf{1}}{\boldsymbol{K_i}} \sum_{\boldsymbol{j=1}}^{\boldsymbol{K_i}} \boldsymbol{Y_{ij}}$$
6. Το άθροισμα της συστάδας i για το χαρακτηριστικό
$$\boldsymbol{t_i} = \sum_{\boldsymbol{j=1}}^{\boldsymbol{K_i}} \boldsymbol{Y_{ij}}$$
7. Το άθροισμα του πληθυσμού για το χαρακτηριστικό
$$Y_T = \sum_{i=1}^{M} t_i$$
8. Ο μέσος του πληθυσμού για το χαρακτηριστικό
$$\bar{Y} = \frac{1}{N} Y_T$$
9. Η διακύμανση στο εσωτερικό της i συστάδας

$$\sigma_i^2 = \frac{1}{K_i - 1}\sum_{j=1}^{K_i}\left(Y_{ij} - \bar{U}_i\right)^2$$

10. Η διακύμανση του πληθυσμου

$$\sigma^2 = \frac{1}{N - 1}\sum_{i=1}^{M}\sum_{j=1}^{K_i}\left(Y_{ij} - \bar{Y}\right)^2$$

11. m το πλήθος συστάδων που επιλέγονται από τις Μ συνολικά
12. k_i το μέγεθος του δείγματος που επιλέγεται από την συστάδα i .
13. X_{ij} το j στοιχείο του δείγματος της i συστάδας.
14. Ο δειγματικός μέσος της συστάδας i

$$\bar{\boldsymbol{X}}_{\boldsymbol{i}} = \frac{\mathbf{1}}{\boldsymbol{k_i}}\sum_{\boldsymbol{j=1}}^{\boldsymbol{k_i}}\boldsymbol{X_{ij}}$$

15. Η δειγματική διακύμανση της συστάδας i

$$s_i^2 = \frac{1}{k_i - 1}\sum_{j=1}^{k_i}\left(X_{ij} - X_i\right)^2$$

Δειγματοληψία κατά συστάδες σε ένα στάδιο

Εκτίμηση παραμέτρων σε συστάδες ίσου μεγέθους με ίσες πιθανότητες

Όταν οι συστάδες είναι ίδιου μεγέθους θα ισχύει

$$K = K_i \quad i = 1,2,\dots,M$$

Απ'όπου προκύπτει ότι $N = MK$ για τον πληθυσμό και $n = mK$ για το δείγμα.

1. ΕΚΤΙΜΗΣΗ ΠΛΗΘΥΣΜΙΑΚΟΎ ΜΈΣΟΥ

Ένας αμερόληπτος εκτιμητής του πληθυσμιακού μέσου είναι ο απλός μέσος του δείγματος

$$\bar{X}_{cl} = \frac{1}{n}\sum_{i=1}^{m} t_i$$

Και λόγω της ισότητας των συστάδων θα ισχύει ότι:

$$\bar{X}_{cl} = \frac{1}{m}\sum_{i=1}^{m} \bar{U}_i$$

2. ΔΙΑΚΥΜΑΝΣΗ ΤΟΥ ΕΚΤΙΜΗΤΗ ΤΟΥ ΜΕΣΟΥ

Η διακύμανση του εκτιμητή $\bar{X}_{cl}$ δίνεται από τον τύπο:

$$Var(\bar{X}_{cl}) = \frac{1-f}{m}\sum_{i=1}^{M}\frac{(\bar{U}_i - \bar{Y})^2}{M-1}$$

3. ΕΚΤΙΜΗΣΗ ΤΗΣ ΔΙΑΚΥΜΑΝΣΗΣ ΤΟΥ ΕΚΤΙΜΗΤΗ ΤΟΥ ΜΕΣΟΥ

$$\hat{Var}(\bar{X}_{cl}) = \frac{1-f}{m}\sum_{i=1}^{m}\frac{(\bar{U}_i - \bar{X}_{cl})^2}{m-1}$$

4. ΔΙΑΣΤΗΜΑΤΑ ΕΜΠΙΣΤΟΣΥΝΗΣ

Για μεγάλο αριθμό m , το διάστημα εμπιστοσύνης είναι:

$$\bar{X}_{cl} \pm z_{a/2}\sqrt{\hat{Var}(\bar{X}_{cl})}$$

Ενώ για μικρό m

$$\bar{X}_{cl} \pm t_{m-1,\ \frac{\alpha}{2}} \sqrt{\hat{Var}(\bar{X}_{cl})}$$

Σύγκριση απλής τυχαίας δειγματοληψίας με δειγματοληψία κατά συστάδες ίσου μεγέθους.

Για να είναι η δείγματοληψία κατά συστάδες ίσου μεγέθους πιο αποτελεσματική από την απλή τυχαία δειγματοληψία πρέπει να ισχύει ότι:

$$\bar{\sigma}^2 \geq \sigma^2$$

Όπου $\bar{\sigma}^2$ η διακύμανση των συστάδων του πληθυσμού.

Απόδειξη: Όμοια με προηγούμενη απόδειξη παίρνουμε την διαφορά των δύο διακυμάνσεων

$$Var(\bar{X}_{srs}) - Var(\bar{X}_{cl}) = \frac{1-\frac{n}{N}}{n}\ \sigma^2 - \frac{1-f}{m}\ \sum_{i=1}^{M} \frac{(\bar{U}_i - \bar{Y})^2}{M-1}$$

Σύμφωνα με απόδειξη στην στρωματοποιημένη δειγματοληψία για τα στοιχεία του πληθυσμού ισχύει η σχέση:

$$(N-1)\sigma^2 = \sum_{i=1}^{M}(K-1)\sigma_i^2 + \sum_{i=1}^{M} K\,(\bar{U}_i - \bar{Y})^2$$

Συνεπώς:

$$Var(\bar{X}_{srs}) - Var(\bar{X}_{cl}) = \frac{(1-f)M(K-1)}{mK(M-1)}(\bar{\sigma}^2 - \sigma^2) > 0$$

Όπου

$$\bar{\sigma}^2 = \frac{1}{M}\sum_{i=1}^{M} \sigma_i^2$$

Τέλος απόδειξης

Ορίζεται και εδώ ο συντελεστής συσχέτισης αντίστοιχα με την περίπτωση της συστηματικής δειγματοληψίας. Όσο μεγαλύτερος ο συντελεστής συσχέτισης ρ_w τόσο μικρότερη η μέση διασπορά $\bar{\sigma}^2$ το οποίο δηλώνει ομάδες με ομοιογένεια στο εσωτερικό τους.

Ο συντελεστής συσχέτισης δίνεται από τον τύπο:

$$\rho_w = 1 - \left(\frac{MK}{MK-1}\right)\frac{\bar{\sigma}^2}{\sigma^2}$$

Ερμηνεία:

- Αν $\rho_w > 0$ τότε υπάρχει ομοιογένεια στο εσωτερικό της συστάδας και ο εκτιμητής $\bar{X}_{cl}$ έχει μεγαλύτερο στατιστικό σφάλμα από αυτόν της απλής τυχαίας.
- Αν $\rho_w = 0$ τότε οι δύο εκτιμητές ταυτίζονται.
- Αν $\rho_w < 0$ τότε ο εκτιμητής $\bar{X}_{cl}$ έχει μικρότερο σφάλμα από τον $\bar{X}_{srs}$

Εκτίμηση παραμέτρων σε συστάδες άνισου μεγέθους

1. ΕΚΤΙΜΗΣΗ ΤΟΥ ΜΕΣΟΥ

Ένας αμερόληπτος εκτιμητής του πληθυσμιακού μέσου $\bar{Y}$ είναι:

$$\bar{X}_{cl,u} = \frac{1}{N}\frac{M}{m}\sum_{i=1}^{m} t_i$$

Ο οποίος γράφεται ισοδύναμα:

$$\bar{X}_{cl,u} = \frac{1}{N}\frac{M}{m}\sum_{i=1}^{m} K_i\bar{U}_i$$

2. ΔΙΑΚΥΜΑΝΣΗ ΤΟΥ ΕΚΤΙΜΗΤΗ ΤΟΥ ΜΕΣΟΥ

$$Var\left(\bar{X}_{cl,u}\right) = \frac{M^2}{N^2}\frac{1-f}{m}\sum_{i=1}^{M}\frac{(K_i\bar{U}_i - \bar{t})^2}{M-1} = \frac{M^2}{N^2}\frac{1-f}{m}\sum_{i=1}^{M}\frac{(t_i - \bar{t})^2}{M-1}$$

Όπου $\bar{t}$ το μέσο t_i για τις ομάδες.

3. ΕΚΤΙΜΗΣΗ ΤΗΣ ΔΙΑΣΠΟΡΑΣ ΤΟΥ ΕΚΤΙΜΗΤΗ ΤΟΥ ΜΕΣΟΥ

$$\hat{Var}(\bar{X}_{cl,u}) = \frac{M^2}{N^2}\frac{1-f}{m}\sum_{i=1}^{m}\frac{(k_i\bar{U}_i - \hat{t})^2}{m-1} = \frac{M^2}{N^2}\frac{1-f}{m}\sum_{i=1}^{m}\frac{(t_i - \hat{t})^2}{m-1}$$

Όπου $\hat{t}$ το μέσο εκτιμώμενο άθροισμα με βάση το δείγμα. Δηλαδή

$$\hat{t} = \frac{1}{m}\sum_{i=1}^{m} t_i$$

Εκτίμηση παραμέτρων σε συστάδες ίσου μεγέθους με άνισες πιθανότητες

1. Ο ΕΚΤΙΜΗΤΗΣ HORVITZ-THOMPSON (HT)

Ο εκτιμητής HT είναι αμερόληπτος και δίνεται από τη σχέση:

$$\hat{\bar{Y}}_{HT} = \frac{1}{N}\sum_{i=1}^{u}\frac{t_i}{\pi_i}$$

Όπου u το πλήθος των συστάδων που είναι διακριτές από τις m που επιλέγονται, π_i η πιθανότητα να επιλεγεί η i συστάδα στο δείγμα και t_i το άθροισμα του περιεχομένου της i συστάδας.
Το τυπικό σφάλμα του εκτιμητή HT δίνεται από τον τύπο:

$$se(\hat{Y}_{HT}) = \sqrt{\sum_{i=1}^{M}\frac{1-\pi_i}{\pi_i}\,t_i^2 + \sum_{i=1}^{M}\sum_{j\neq i}\left(\frac{\pi_{ij}-\pi_i\pi_j}{\pi_i\pi_j}\right)t_i t_j}$$

Για την εκτίμηση του τυπικού σφάλματος υπάρχουν δύο προτάσεις:

$$\widehat{se}_1(\hat{Y}_{HT}) = \sqrt{\sum_{i=1}^{u} \frac{1-\pi_i}{\pi_i^2}\, t_i^2 + \sum_{i=1}^{u}\sum_{j\neq i}\left(\frac{\pi_{ij}-\pi_i\pi_j}{\pi_i\pi_j\pi_{ij}}\right) t_i t_j}$$

Από τους Horvitz-Thompson, και

$$\widehat{se}_2(\hat{Y}_{HT}) = \sqrt{\sum_{i=1}^{u}\sum_{j>i}\left(\frac{\pi_i\pi_j-\pi_{ij}}{\pi_{ij}}\right)\left(\frac{t_i}{\pi_i}-\frac{t_j}{\pi_j}\right)^2}$$

Από τους Yates & Grundy.

2. Ο ΕΚΤΙΜΗΤΗΣ HANSEN-HURWITZ

Ο εκτιμητής ΗΗ είναι αμερόληπτος και δίνεται από τη σχέση:

$$\hat{Y}_{HH} = \frac{1}{m}\sum_{i=1}^{m}\frac{t_i}{\pi'_i} = \frac{1}{m}\sum_{i=1}^{m} u_i = \bar{u}$$

Όπου π'_i η πιθανότητα της i συστάδας να επιλεγεί σε καθεμια από τις m ανεξάρτητες επιλογές μιας μονάδας του πληθυσμού.

Τυπικό σφάλμα του εκτιμητή $\hat{Y}_{HH}$.

$$se(\hat{Y}_{HH}) = \sqrt{\frac{1}{m}\sum_{i=1}^{M}(u_i - Y_T)^2\pi'_i}$$

Αμερόληπτος εκτιμητής του τυπικού σφάλματος

$$\widehat{se}(\hat{Y}_{HH}) = \sqrt{\frac{1}{m}\sum_{i=1}^{m}\frac{(u_i-\bar{u})^2}{m-1}}$$

ΚΕΦΑΛΑΙΟ 5
Δειγματοληψία χωρίς πιθανότητες

5.1 Δειγματοληψία ποσοστών

Σύμφωνα με αυτή τη μέθοδο δειγματοληψίας, ο στατιστικός αναλυτής συμπεριλαμβάνει στο δείγμα του μονάδες του πληθυσμού ώστε το τελικό δείγμα να έχει εκπροσώπους από κάθε κατηγορία του πληθυσμού, και μάλιστα με αναλογία ίση με εκείνη που ισχύει για τον πληθυσμό. Οι κατηγορίες ορίζονται συνήθως με βάση ένα δημογραφικό κριτήριο, π.χ φύλο, ηλικιακές ομάδες κτλ. Τα κριτήρια που καθορίζονται τα στρώματα συνδέονται με το θέμα που εξετάζουμε.

Για παράδειγμα, αν η κοινωνικοοικονομική κατάσταση είναι σημαντική σε μια έρευνα, τότε μπορούν να χρησιμοποιηθούν κατηγορίες όπως εργοδότες, επιχειρηματίες, εργάτες, εργαζόμενοι κπλ. Δίνεται ένα συγκεκριμένο ποσοστό από κάθε κατηγορία, και μέσα σε κάθε κατηγορία πραγματοποιείται δειγματοληψία ευκολίας.

Η δειγματοληψία με προκαθορισμένα ποσοστά έχει ομοιότητες με τη στρωματοποιημένη, και ειδικότερα την αναλογική στρωματοποιημένη, αλλά δεν αποτελεί δειγματοληψία πιθανότητας όπως η στρωματοποιημένη.

5.2 Δειγματοληψία ευκολίας

Ένα δείγμα ευκολίας είναι ένα δείγμα μη πιθανότητας στο οποίο ο ερευνητής χρησιμοποιεί τα θέματα που είναι πλησιέστερα και διαθέσιμα για να συμμετάσχει στην ερευνητική μελέτη. Αυτή η τεχνική αναφέρεται επίσης ως «τυχαία δειγματοληψία», και χρησιμοποιείται συνήθως σε πιλοτικές μελέτες πριν από την έναρξη ενός μεγαλύτερου ερευνητικού έργου.

Βασικές επιλογές: Δείγματα ευκολίας

- Ένα δείγμα ευκολίας αποτελείται από ερευνητικά θέματα που επιλέχθηκαν για μια μελέτη επειδή θα μπορούσαν να προσληφθούν εύκολα.
- Ένα μειονέκτημα της δειγματοληψίας ευκολίας είναι ότι τα υποκείμενα σε ένα δείγμα ευκολίας μπορεί να μην είναι αντιπροσωπευτικά του πληθυσμού που ο ερευνητής ενδιαφέρεται να μελετήσει.
- Ένα πλεονέκτημα της εύκολης δειγματοληψίας είναι ότι τα δεδομένα μπορούν να συλλεχθούν γρήγορα και με χαμηλό κόστος.
- Τα δείγματα ευκολίας χρησιμοποιούνται συχνά σε πιλοτικές μελέτες, μέσω των οποίων οι ερευνητές μπορούν να βελτιώσουν μια ερευνητική μελέτη πριν δοκιμάσουν ένα μεγαλύτερο και πιο αντιπροσωπευτικό δείγμα.

Όταν μια ερευνητής είναι πρόθυμη να ξεκινήσει τη διεξαγωγή έρευνας με άτομα ως αντικείμενα, αλλά μπορεί να μην έχει μεγάλο προϋπολογισμό ή τον χρόνο και τους πόρους που θα επέτρεπαν τη δημιουργία ενός μεγάλου, τυχαιοποιημένου δείγματος, μπορεί να επιλέξει να χρησιμοποιήσει την τεχνική της δειγματοληψίας ευκολίας. Αυτό θα μπορούσε να σημαίνει τη διακοπή των ανθρώπων καθώς περπατούν κατά μήκος ενός πεζοδρομίου, ή την παρακολούθηση των περαστικών σε ένα εμπορικό κέντρο, για παράδειγμα. Θα μπορούσε επίσης να σημαίνει έρευνα φίλων, μαθητών ή συναδέλφων στους οποίους ο ερευνητής έχει τακτική πρόσβαση.

Με ένα δείγμα ευκολίας, ο ερευνητής δεν μπορεί να ελέγξει την αντιπροσωπευτικότητα του δείγματος. Αυτή η έλλειψη ελέγχου μπορεί να προκαλέσει προκατειλημμένο δείγμα και αποτελέσματα έρευνας, και έτσι περιορίζει την ευρύτερη εφαρμογή της μελέτης.

Ενώ τα αποτελέσματα των μελετών που χρησιμοποιούν δείγματα ευκολίας μπορεί να μην είναι απαραίτητα εφαρμόσιμα στον μεγαλύτερο πληθυσμό, τα αποτελέσματα θα μπορούσαν να είναι ακόμη χρήσιμα. Για παράδειγμα, ο ερευνητής θα μπορούσε να θεωρήσει την έρευνα πιλοτική μελέτη και να χρησιμοποιήσει τα αποτελέσματα για να βελτιώσει ορισμένες ερωτήσεις στην έρευνα ή να βρει περισσότερες ερωτήσεις για να συμπεριληφθεί σε μεταγενέστερη έρευνα. Τα δείγματα ευκολίας χρησιμοποιούνται συχνά για το σκοπό αυτό: για να δοκιμάσετε συγκεκριμένες ερωτήσεις και να δείτε τι είδους απαντήσεις προκύπτουν και να χρησιμοποιήσετε αυτά τα αποτελέσματα ως εφαλτήριο για να δημιουργήσετε ένα πιο λεπτομερές και χρήσιμο ερωτηματολόγιο .

Ένα δείγμα ευκολίας έχει επίσης το πλεονέκτημα ότι επιτρέπει τη διεξαγωγή ερευνητικής μελέτης χαμηλού έως χωρίς κόστος, επειδή χρησιμοποιεί τον πληθυσμό που είναι ήδη διαθέσιμος. Είναι επίσης αποδοτικό ως προς το χρόνο, διότι επιτρέπει την έρευνα να διεξάγεται κατά τη διάρκεια της καθημερινής ζωής του ερευνητή. Ως εκ τούτου, ένα δείγμα ευκολίας επιλέγεται συχνά όταν δεν είναι δυνατόν να επιτευχθούν άλλες τυχαιοποιημένες τεχνικές δειγματοληψίας . (*Nicki Lisa Cole, Ph.D.*)

5.3 Δειγματοληψία κρίσης

Στα δείγματα κρίσης ο ερευνητής επιλέγει τις μονάδες του πληθυσμού με βάση την προσωπική του κρίση, ή την εμπειρία του από προηγούμενες έρευνες με παρόμοιο θέμα στο ίδιο σύνολο πληθυσμού. Για παράδειγμα, σε δημοσκοπήσεις με στόχο πολιτικές έρευνες, αποτελέσματα εκλογών κτλ., ο ερευνητής μπορεί να επιλέξει το δείγμα του συμπεριλαμβάνοντας με βεβαιότητα περιοχές του πληθυσμού που έχουν ιδιαίτερα χαρακτηριστικά, π.χ. στις πιο πρόσφατες εκλογές, τα εκλογικά αποτελέσματα των περιοχών αυτών ήταν πολύ κοντά στα τελικά αποτελέσματα όλης της επικράτειας. Οι περιοχές αυτές αποκαλούνται «περιοχές βαρόμετρο» και επιλέγονται με βεβαιότητα στο δείγμα, γιατί βάσει της εμπειρίας από προηγούμενες εκλογές θεωρούνται αντιπροσωπευτικές. (Ρούσσος Π)

5.4 Δειγματοληψία χιονόμπαλας

Στα δείγματα χιονοστιβάδας, το δείγμα γίνεται προσβάσιμο στον ερευνητή μέσω ενός μικρού αρχικού συνόλου δείγματος που είναι διαθέσιμο σε εκείνον. Η κάθε μιά δειγματοληπτική μονάδα του αρχικού δείγματος προσφέρει τα στοιχεία, και άρα την πρόσβαση, σε ένα σύνολο από άλλα μέλη του πληθυσμού, τα οποία συμπεριλαμβάνονται στο δείγμα, και τα οποία με τη σειρά τους προσφέρουν πρόσβαση σε ένα άλλο σύνολο κοκ. Ο στατιστικός αναλυτής, συνεπώς, αποκτά το δείγμα μέσω των αρχικών εκπροσώπων, χωρίς προσπάθεια εντοπισμού και χωρίς ανάγκη να διαθέτει στοιχεία για τον πληθυσμό (λίστα μελών, πλήθος κτλ).

Η δειγματοληψία χιονόμπαλας είναι μια δημοφιλής τεχνική μεταξύ των κοινωνικών επιστημόνων που επιθυμούν να συνεργαστούν με έναν πληθυσμό που είναι δύσκολο να εντοπιστεί ή να εντοπιστεί. Αυτό συμβαίνει συχνά όταν ο πληθυσμός είναι κάπως περιθωριοποιημένος, όπως άστεγοι ή πρώην φυλακισμένοι ή εκείνοι που εμπλέκονται σε παράνομες δραστηριότητες. Είναι επίσης συνηθισμένο να χρησιμοποιείται αυτή η τεχνική δειγματοληψίας με άτομα των οποίων η συμμετοχή σε μια συγκεκριμένη ομάδα δεν είναι ευρέως γνωστή, όπως γκέι άνθρωποι ή αμφιφυλόφιλοι ή τρανσέξουαλ. (Ρούσσος Π)

5.5 Δειγματοληψία σκοπιμότητας

Αναφέρεται στην επιλογή στο δείγμα ορισμένων ομάδων (ή περιπτώσεων) του πληθυσμού που ικανοποιούν ορισμένες υποθέσεις. Ο ερευνητής επιλέγει ως δείγμα όσους θεωρεί ότι ανταποκρίνονται σε συγκεκριμένα χαρακτηριστικά. Κατά συνέπεια, το δείγμα βασίζεται στην κρίση του ερευνητή.

ΒΙΒΛΙΟΓΡΑΦΙΑ

- Diamond I. and J., (2006), Αρχίζοντας στατιστική. Μια εισαγωγή για τους κοινωνικούς επιστήμονες, ελληνική μετάφραση Μαρία Συμεωνάκη, εκδόσεις Παπαζήση
- Thompson, S. K. (2012). Sampling (3rd Edition). Hoboken, NJ: John Wiley and Sons.
- Ε. Ξεκαλάκη και Ι. Πανάρετου: Πιθανότητες και Στοιχεία Στοχαστικών Ανελίξεων, Αθήνα 1993
- Ιουλία Παπαγεωργίου: Θεωρία Δειγματοληψίας, Αθήνα 2005
- Κατερίνα Δημάκη: Επιλογή δείγματος, Τμήμα Στατιστικής, Οικονομικό Πανεπιστήμιο Αθηνών.
- M. Saundres, P. Lewis και A. Thornhill: Μέθοδοι Έρευνας στις Επιχειρήσεις και την Οικονομία
- Κ. Κουτσικόπουλος: Δειγματοληψία-Οικολογία, Πάτρα 2002
- Κοκολάκης Γ., Σπηλιώτης Ι., (1999), Εισαγωγή στη θεωρία πιθανοτήτων και στατιστική, Συμεών, Αθήνα
- Μπένος Β., (1991), Μέθοδοι και τεχνικές δειγματοληψίας, Σταμούλης, Πειραιάς
- Ρούσσος Π.: Δειγματοληπτική έρευνα και ερωτηματολόγια, Πανεπιστήμιο Αθηνών.