ΣΥΝΟΠΤΙΚΕΣ ΣΗΜΕΙΩΣΕΙΣ

ΣΤΑΤΙΣΤΙΚΗΣ

Εισαγωγή – Βασικές Έννοιες

Ορισμός

Στατιστική (Statistics) είναι ένα σύνολο αρχών και μεθοδολογιών για

Α. το σχεδιασμό της διαδικασίας συλλογής δεδομένων

Β. τη συνοπτική και αποτελεσματική παρουσίασή τους

Γ. την εξαγωγή αντίστοιχων συμπερασμάτων

Η Στατιστική επιτυγχάνει τη συλλογή, επεξεργασία, παρουσίαση και ανάλυση των στατιστικών στοιχείων (αριθμητικών δεδομένων) με τη εφαρμογή κατάλληλων για κάθε περίπτωση στατιστικών μεθόδων, οι οποίες και συνιστούν το περιεχόμενό της.

Ορισμός

Κάθε σύνολο αντικειμένων ή ατόμων που έχουν κάποιο κοινό μετρήσιμο χαρακτηριστικό αποτελεί έναν **πληθυσμό (population)**.

Ορισμός

Κάθε υποσύνολο του πληθυσμού αποτελεί ένα **δείγμα (sample)** από τον πληθυσμό.

Ορισμός

Ένα **τυχαίο δείγμα (random sample)** είναι το δείγμα του πληθυσμού, όπου τα άτομα διαλέγονται το ένα μετά το άλλο, με κύριο χαρακτηριστικό, ότι τα υπόλοιπα άτομα του πληθυσμού κάθε φορά, έχουν τις ίδιες πιθανότητες να περιληφθούν στο τυχαίο δείγμα.

Ορισμός

Εάν μετά από κάθε διαλογή τα άτομα επιστρέφουν στον πληθυσμό, τότε έχουμε **δειγματοληψία με επανατοποθέτηση.**

Ορισμός

Εάν η επιλογή του επόμενου ατόμου του δείγματος γίνει μόνο από τα υπόλοιπα άτομα, τότε έχουμε **δειγματοληψία χωρίς επανατοποθέτηση.**

Παρατήρηση

Οι όροι πληθυσμός και δείγμα μπορεί να αναφέρονται είτε στα άτομα, είτε στις μετρήσεις του κοινού χαρακτηριστικού τους. Τότε υπάρχει μια **κατανομή** των μετρήσεων του δείγματος, η οποία συνήθως μελετάται και μια κατανομή των μετρήσεων όλου του πληθυσμού που συνήθως υπάρχει αλλά είναι δύσκολο να προσδιοριστεί.

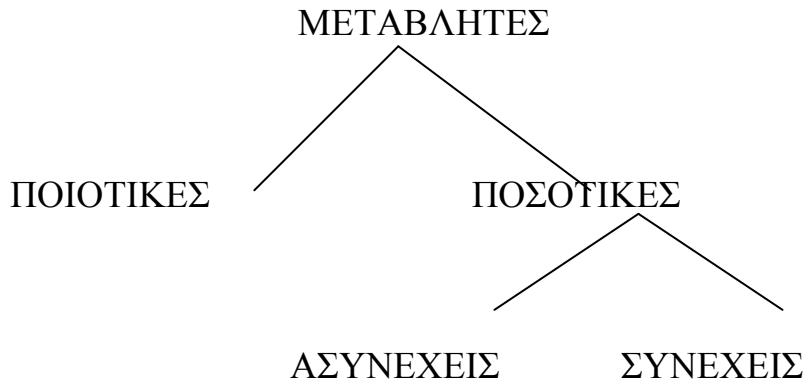
Ένα ενδιαφέρον πρόβλημα είναι η απόφαση για το τι είδος πληροφορίες δύναται να εξαχθούν για την κατανομή του πληθυσμού, από την παρατήρηση και τη μελέτη της κατανομής του τυχαίου δείγματος.

Ένα άλλο είδος **κατανομής** μπορεί να ληφθεί από την κατανομή των μετρήσεων που έγιναν σε ένα από όλα τα δυνατά δείγματα, σταθερού μεγέθους, που θα μπορούσαν να ληφθούν από έναν πληθυσμό.

Ορισμός

Τα χαρακτηριστικά ή ιδιότητες των στατιστικών μονάδων ως προς τα οποία εξετάζουμε έναν πληθυσμό ονομάζονται **μεταβλητές (variables).** Συμβολίζονται με κεφαλαία γράμματα και οι τιμές τους με τα αντίστοιχα μικρά.

Διάκριση μεταβλητών



1. **Ποιοτικές (qualitative)** χαρακτηρίζονται οι μεταβλητές που δεν επιδέχονται αριθμητική μέτρηση.

2. **Ποσοτικές (quantitative)** είναι οι μεταβλητές που δύναται να επιδέχονται αριθμητική μέτρηση.

Οι ποσοτικές μεταβλητές διακρίνονται σε δυο ακόμα κατηγορίες:

Α. **Ασυνεχείς ή Διακριτές (Discrete Variables)** είναι εκείνες που παίρνουν ακέραιες τιμές (αριθμός λευκών ή ερυθρών αιμοσφαιρίων, αριθμός υπαλλήλων ενός λογιστηρίου, αριθμός παιδιών μιας οικογένειας, αριθμός ραδιενεργών κρούσεων, αριθμός ελαττωματικών προϊόντων).

Β. **Συνεχείς (Continuous Variables)** είναι εκείνες που μπορούν να πάρουν όλες τις τιμές ενός διαστήματος πραγματικών αριθμών (βάρος, ύψος).

Κλίμακες Μέτρησης

Ευρέως χρησιμοποιούνται οι εξής τέσσερις κλίμακες: κατηγορίας, διάταξης, διαστήματος και αναλογίας. Οι δυο πρώτες κλίμακες μέτρησης αφορούν τις ποιοτικές μεταβλητές ενώ οι δυο τελευταίες τις ποσοτικές.

Κατηγορίας (nominal) είναι οι μεταβλητές των οποίων το σύνολο των τιμών δεν έχει καμία ιδιότητα. Για τη μεταβλητή αυτή, μοναδική σημασία έχουν οι διαφορετικές τιμές (το πλήθος των κατηγοριών της)

που μπορεί να πάρει. Η μοναδική σχέση που μπορεί να προσδιοριστεί μεταξύ των κατηγοριών αυτών είναι απλά η ύπαρξη διαφοράς.

Διάταξης (ordinal) είναι οι μεταβλητές που για το σύνολο τιμών τους μπορούμε να ορίσουμε μια σχέση διάταξης, δηλαδή να τοποθετηθούν στη σειρά. Η διάταξη μπορεί να είναι από τη μικρότερη τιμή προς τη μεγαλύτερη ή και αντίστροφα. Οι ίσες διαφορές μεταξύ των τιμών μιας τέτοιας μεταβλητής δεν συνεπάγονται και ίσες διαφορές για το χαρακτηριστικό που μετράει η μεταβλητή. Δηλαδή, δεν υπάρχει αντιστοίχιση σε υποδιαιρέσεις ή πολλαπλάσια κάποιας μονάδας. Η διάταξη δηλαδή, το μόνο που εξασφαλίζει είναι τον προσδιορισμό της μεγαλύτερης, καλύτερης, προτιμότερης κατηγορίας αλλά όχι πόσο μεγαλύτερη, καλύτερη, προτιμότερη είναι σε σχέση με κάποια από τις υπόλοιπες.

Διαστήματος (interval) είναι οι μεταβλητές των οποίων οι ίσες διαφορές μεταξύ των τιμών τους συνεπάγονται και ίσες διαφορές για το χαρακτηριστικό που μετράει η μεταβλητή (π.χ. ηλικία, θερμοκρασία). Η κλίμακα αυτή δεν επιτρέπει μόνο την ιεράρχηση των υποκειμένων αλλά προσδιορίζει επίσης και την ακριβή διαφορά τους. Η απόσταση μεταξύ δυο οποιονδήποτε διαδοχικών τιμών της μεταβλητής αυτής είναι ίση με την απόσταση δυο άλλων τυχαίων διαδοχικών τιμών της. Επίσης, δεν

έχει νόημα ο υπολογισμός αναλογιών. Βασικό χαρακτηριστικό των μεταβλητών αυτής της διάταξης είναι ο αυθαίρετος ορισμός του μηδενός, που δεν υποδηλώνει παντελή έλλειψη του μετρήσιμου χαρακτηριστικού.

Αναλογίας (rate) είναι οι μεταβλητές των οποίων οι τιμές αντιστοιχούν αναλογικά στην ποσότητα του χαρακτηριστικού που μετρούν. Εδώ το μηδέν ανήκει στο διάστημα τιμών της μεταβλητής και δηλώνει την πλήρη απουσία. Και επίσης για τις τιμές των μεταβλητών αυτών έχει έννοια ο υπολογισμός των αναλογιών.

Περιγραφή Δεδομένων – Τρόποι Παρουσίασης

Τα στατιστικά δεδομένα πρέπει να παρουσιάζονται με τρόπο απλό και σαφή, έτσι ώστε να είναι εύκολη η κατανόησή τους από τον κάθε ενδιαφερόμενο. Η παρουσίαση μπορεί να γίνει με μορφή

Α. Πινάκων

Β. Γραφικών Παραστάσεων

Πίνακες

Σε κάθε πίνακα, που έχει συνταχθεί σωστά, εκτός από το κύριο σώμα, που περιέχει διαχωρισμένα μέσα στις γραμμές και στήλες τα στατιστικά δεδομένα, παρατηρούνται και τα εξής ειδικότερα στοιχεία:

Α. τον **τίτλο**, που γράφεται στο πάνω μέρος και πρέπει να δηλώνει με σαφήνεια και με περιληπτικό τρόπο το περιεχόμενο του πίνακα

Β. τις **επικεφαλίδες των στηλών (και γραμμών)**, που δείχνουν συνοπτικά τη φύση και τη μονάδα μετρήσεως των δεδομένων

Γ. την **πηγή** που γράφεται στο κάτω μέρος του πίνακα και δείχνει την προέλευση των δεδομένων

Δ. τις **υποσημειώσεις** που γράφονται στο κάτω μέρος του πίνακα και πριν από την πηγή, αν θεωρηθεί απαραίτητο να δοθούν κάποιες επεξηγήσεις.

Τύποι πινάκων

Οι πίνακες μπορεί να είναι **απλής εισόδου** ή **διπλής εισόδου.**

Οι πίνακες απλής εισόδου χρησιμοποιούνται όταν οι μονάδες του εξεταζόμενου πληθυσμού ερευνώνται ως προς ένα ποιοτικό ή ποσοτικό χαρακτηριστικό.

Ενώ οι πίνακες διπλής εισόδου όταν οι μονάδες του εξεταζόμενου πληθυσμού μελετώνται ταυτοχρόνως ως προς δυο ποιοτικά ή ποσοτικά χαρακτηριστικά.

Πίνακες κατανομής συχνοτήτων

Οι πίνακες αυτοί συντάσσονται με κατάλληλη κατάταξη και συστηματική ομαδοποίηση των τιμών της μεταβλητής που εξετάζεται. Ο τρόπος κατασκευής τους εξαρτάται από το είδος των χαρακτηριστικών.

Α. **Ασυνεχή ή Διακριτά**

Αν τα χαρακτηριστικά είναι διακριτά και τα δυνατά αποτελέσματα της μέτρησης σχετικά λίγα τότε ο πίνακας παίρνει την ακόλουθη μορφή:

Δυνατές τιμές της μεταβλητής	**Αριθμός φορών που παρατηρήθηκε η κάθε τιμή (Συχνότητα)**
x_1	f_1
x_2	f_2
.	.
.	.
.	.
x_k	f_k
Σύνολο	$\sum_{i=1}^{k} f_i = n$

Τα $x_{1,}\ldots,\ x_k$ είναι οι τιμές της διακριτής μεταβλητής Χ, οι οποίες τοποθετούνται κατά τη φυσική τους σειρά, από τη μικρότερη στη μεγαλύτερη. Και τα $f_1, \ldots, f_k$ εκφράζουν πόσες φορές εμφανίζεται στο συνολικό πληθυσμό κάθε τιμή της μεταβλητής.

Όταν η τιμή x_i εμφανίζεται f_i φορές τότε λέμε ότι f_i είναι η **απόλυτη συχνότητα** ή απλά **συχνότητα** της x_i. Ενώ η **σχετική συχνότητά της**

είναι η ποσότητα $\frac{f_i}{\sum f_i}$ (πολλές φορές χρησιμοποιείται η παραπάνω ποσότητα εκφρασμένη επί %).

Β. **Συνεχή**

Αν τα χαρακτηριστικά είναι συνεχή ή διακριτά με μεγάλο πλήθος δυνατών τιμών, τότε δυσχεραίνεται η μορφή του πίνακα, οπότε κρίνεται απαραίτητη η ομαδοποίηση των παρατηρήσεων. Η ομαδοποίηση αυτή πραγματοποιείται με το χωρισμό του διαστήματος μεταβολής (α_0 , α_1) της μεταβλητής Χ σε υποδιαστήματα της μορφής [α_{i-1} , α_i), που ονομάζονται **τάξεις** ή **ομάδες** ή **κλάσεις.**

Τα άκρα των τάξεων καλούνται αντίστοιχα, το μεν α_{i-1} **κατώτερο όριο**, το δε α_i **ανώτερο όριο.** Η διαφορά των δυο ορίων καλείται **πλάτος** της τάξεως και συμβολίζεται με **δ**. Το ημιάθροισμα των ορίων της κάθε τάξης καλείται **κεντρική τιμή της τάξεως,** δηλαδή

$$x_i = \frac{a_{i-1} + a_i}{2}$$

Οι συχνότητες εδώ δίνουν τον αριθμό των παρατηρήσεων που περιέχονται στις αντίστοιχες τάξεις. Ακόμα με M και m συμβολίζονται η μέγιστη και η ελάχιστη τιμή αντίστοιχα της μεταβλητής.

Μια προφανής δυσκολία που υπάρχει στην ομαδοποίηση των παρατηρήσεων είναι ο προσδιορισμός του αριθμού των τάξεων κ, που θα χρησιμοποιηθούν.

Στην πράξη, συνήθως ο αριθμός των τάξεων κυμαίνεται κατά μέσο όρο 8 με 10. Επίσης, συχνά έχουμε τάξεις ίσου πλάτους. Φυσικά, υπάρχουν και οι περιπτώσεις άνισου πλάτους, όπως για παράδειγμα στις κατανομές δαπανών, ημερών ανεργίας κ.ο.κ.

Μεθοδολογία

1. Επιλογή των τάξεων.

Αν η επιλογή δεν είναι αυθαίρετη γίνεται η χρήση του κανόνα του Sturges:

$$\kappa = 1 + 3,322 \log_{10} n$$

όπου κ ο αριθμός των τάξεων και n ο αριθμός των παρατηρήσεων.

2. Επιλογή του πλάτους των τάξεων.

Με χρήση του τύπου:

$$\delta = \frac{M - m}{\kappa}$$

3. Καθορισμός Διαστημάτων.

Το πρώτο διάστημα διαλέγεται συνήθως έτσι ώστε να περιέχει τη μικρότερη παρατήρηση και το τελευταίο τη μεγαλύτερη. Καλό θα ήταν η επιλογή του σημείου αρχής του πρώτου διαστήματος να γίνεται έτσι ώστε καμία παρατήρηση να μην συμπίπτει με άκρο του διαστήματος για να αποφεύγονται πιθανές αμφισβητήσεις σχετικά με το διάστημα στο οποίο βρίσκεται κάθε παρατήρηση.

Παρατήρηση

Οι στρογγυλοποιήσεις που πιθανόν να χρειαστούν κατά τον υπολογισμό του κ και δ πρέπει να γίνουν προς τα πάνω ώστε τα κ διαστήματα πλάτους δ να καλύψουν όλες τις διαθέσιμες παρατηρήσεις.

Τάξεις	**Κεντρικές τιμές**	**Συχνότητες**
$\alpha_0 - \alpha_1$	x_1	f_1
$\alpha_1 - \alpha_2$	x_2	f_2
.	.	.
$\alpha_{i-1} - \alpha_i$	x_i	f_i
.	.	.
$\alpha_{κ-1} - \alpha_κ$	$x_κ$	$f_κ$
		$\sum_{i=1}^{k} f_i = n$

Αθροιστικές Κατανομές Συχνοτήτων

Πολλές φορές χρειάζεται να γνωρίζουμε πόσες ή τι ποσοστό από τις παρατηρήσεις μιας μεταβλητής περιλαμβάνεται **μέχρι** ενός ορισμένου διαστήματος ή το πλήθος των παρατηρήσεων που είναι μικρότερες ή ίσες από μια ορισμένη τιμή της μεταβλητής.

Οι αθροιστικές συχνότητες F_i δίνουν την κατάλληλη απάντηση στο παραπάνω ερώτημα. Δηλαδή, αν f_1, f_2, …, f_k οι συχνότητες της μεταβλητής X, τότε η αθροιστική συχνότητα της τιμής x_i είναι $F_i = f_1 + f_2 + \ldots + f_i$.

Οι σχετικές αθροιστικές συχνότητες εκφράζονται με τον ίδιο τρόπο που εκφράζονται οι σχετικές συχνότητες.

Τα ίδια ισχύουν και στα ομαδοποιημένα σε τάξεις χαρακτηριστικά. Η αθροιστική συχνότητα μιας τάξης μας δείχνει πόσες (ή ποιο ποσοστό αν αναφερόμαστε σε σχετική) παρατηρήσεις είναι μικρότερες από το άνω όριο της τάξης αυτής.

Παρατήρηση

Ένας πίνακας καλό είναι να περιέχει τις ακόλουθες στήλες:

Τάξεις-Κεντρικές τιμές – Συχνότητες - Σχ.Συχνότητες – Αθροιστικές - Σχ. Αθροιστικές.

Γραφικές Παραστάσεις

Όπως και στους πίνακες έτσι και εδώ θα πρέπει μια γραφική παράσταση να περιέχει τα ακόλουθα στοιχεία: τίτλο, κλίμακα μεγεθών, υπόμνημα, πηγή.

Παρακάτω, αναφέρονται μερικά από τις βασικότερες γραφικές παραστάσεις:

1. Ακιδωτά διαγράμματα ή Ραβδογράμματα (Bar Charts)

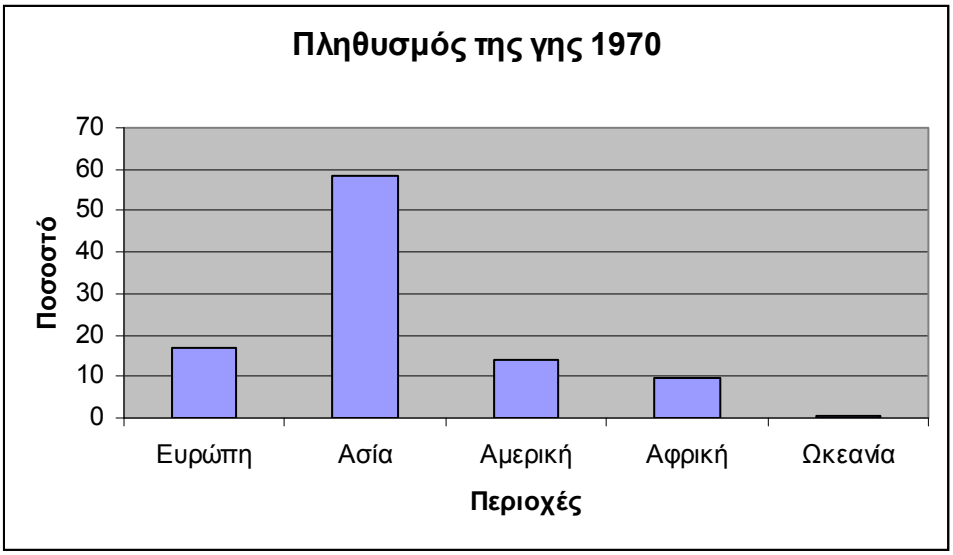
Χρησιμοποιούνται κυρίως για τη γραφική απεικόνιση ποιοτικών δεδομένων .

Ένα τέτοιο διάγραμμα αντιστοιχεί στα δεδομένα του παρακάτω πίνακα:

Πληθυσμός της γης, ανά περιοχή, το 1970

Περιοχές	**Ποσοστό**
Ευρώπη	16,7
Ασία	58,5
Αμερική	14
Αφρική	9,5
Ωκεανία	0,6
Σύνολο	100

Το ακιδωτό διάγραμμα των παραπάνω δεδομένων είναι το ακόλουθο:

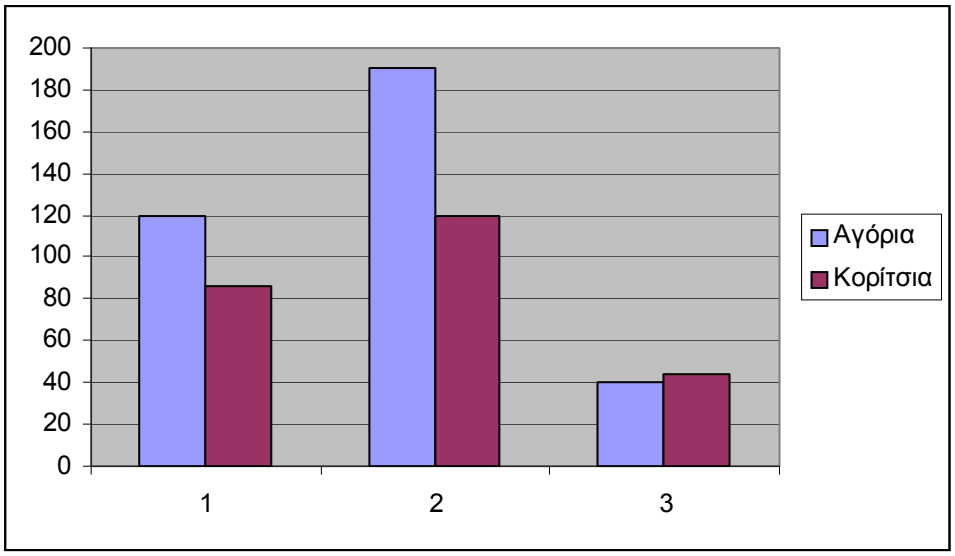


2. Σύνθετα Ακιδωτά Διαγράμματα

Παράδειγμα

Κατανομή 600 μαθητών ενός σχολείου ως προς το φύλο και το χρώμα των ματιών τους.

	Χρώμα Ματιών			
Φύλο	Μαύρα	Καστανά	Γαλανά	Σύνολο
Αγόρια	120	190	40	350
Κορίτσια	86	120	44	250
Σύνολο	206	310	84	600



3. Κυκλικά Διαγράμματα (Pie Charts)

Αυτά είναι ένας κύκλος χωρισμένους σε κυκλικούς τομείς και κάθε κυκλικός τομέας αντιστοιχεί σ' ένα τμήμα του απεικονιζόμενου συνόλου (σε μια τιμή της μεταβλητής). Επειδή τα απόλυτα μεγέθη των τμημάτων μετατρέπονται σε ποσοστά επί τοις 100 του συνόλου (η γωνία του κυκλικού τομέα είναι ανάλογη με την αντίστοιχη σχετική συχνότητα της τιμής της μεταβλητής), γι' αυτό τα κυκλικά διαγράμματα ενδείκνυνται όταν ζητείται η παρουσίαση της ποσοστιαίας σύνθεσης του εξεταζόμενου συνόλου ή η σύγκριση των ποσοστιαίων συνθέσεων δυο ή περισσοτέρων συνόλων. Ενδεικτικό είναι το ακόλουθο παράδειγμα.

Παράδειγμα

Δίνεται η ποσοστιαία σύνθεση (%) του προσωπικού μιας επιχειρήσεως ως προς το μορφωτικό τους επίπεδο.

Επίπεδο Μορφώσεως	**Ποσοστό**
Πτυχιούχοι Τρι/θμιας Εκπαιδεύσεως	25
Απόφοιτοι Λυκείου	50
Απόφοιτοι Γυμνασίου	25

Να απεικονιστούν τα παραπάνω δεδομένα με κυκλικό διάγραμμα.

Για να παρουσιαστούν τα παραπάνω δεδομένα σε ένα κύκλο, πρέπει πρώτα να βρεθεί πόση γωνία αντιστοιχεί σε κάθε ποσοστό από τα παραπάνω, με χρήση μιας απλής μεθόδου:

Στα 100 αντιστοιχούν 25

Στις 360 μοίρες ;

Άρα x=25 (360/100)=90 μοίρες

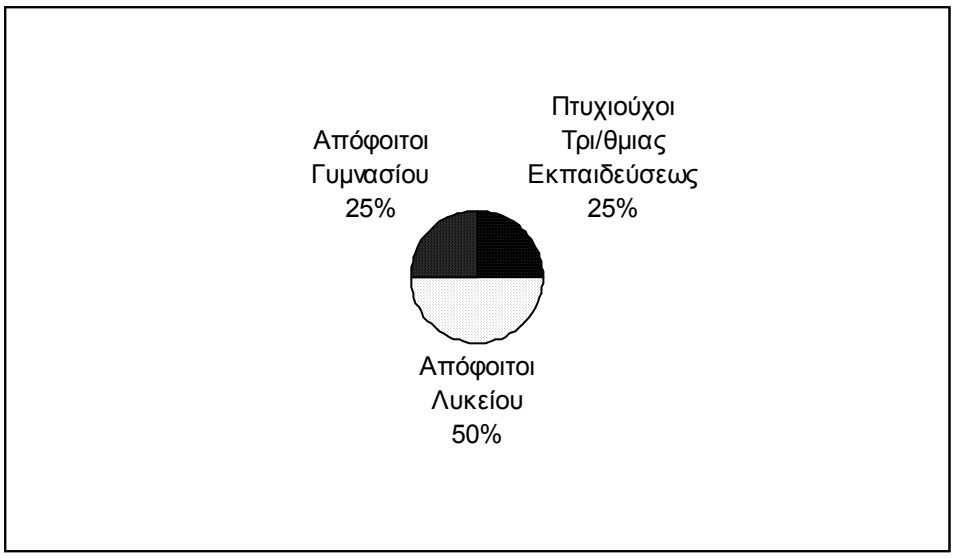
Όμοια

Στα 100 αντιστοιχούν 50

Στις 360 μοίρες ;

Άρα x=50 (360/100)=180 μοίρες

Αρχίζοντας από ένα τυχαίο σημείο του κύκλου και φέρνοντας γωνίες 90, 180 και 90 μοιρών, σχηματίζεται το κυκλικό διάγραμμα:



4. Ιστογράμματα (Histograms)

Χρησιμοποιούνται για τη γραφική απεικόνιση ποσοτικών κατανομών, αποτελούνται από διαδοχικά ορθογώνια, που έχουν βάσεις ίσες με τα διαστήματα των τάξεων τοποθετημένες πάνω στον οριζόντιο άξονα. Το εμβαδόν κάθε ορθογωνίου ισούται με τη συχνότητα της αντίστοιχης τάξεως.

Επίσης, αν ενώσουμε τα μέσα των επάνω βάσεων των ορθογωνίων ενός ιστογράμματος, σχηματίζεται τεθλασμένη γραμμή, που λέγεται **πολύγωνο συχνοτήτων (frequency polygon)**. Επίσης, θα μπορούσαμε, αφού χωριστεί ο οριζόντιος άξονας σε διαστήματα, που αντιστοιχούν στις τάξεις ίσου πλάτους, να ορίσουμε πάνω από το μέσο κάθε διαστήματος και σε ύψος ίσο με την αντίστοιχη συχνότητα κάθε τάξεως. Στη συνέχεια, ενώνοντας αυτά τα σημεία με ευθύγραμμα τμήματα σχηματίζεται το πολύγωνο συχνοτήτων.

Αν στον κάθετο άξονα αντί για τις απλές συχνότητες έχουμε τις σχετικές, τότε έχουμε το ιστόγραμμα και το πολύγωνο αντίστοιχα των σχετικών συχνοτήτων.

Και στην περίπτωση, που έχουμε τις αθροιστικές συχνότητες στον κάθετο άξονα, τότε έχουμε το **ιστόγραμμα των αθροιστικών συχνοτήτων**. Το **πολύγωνο των αθροιστικών συχνοτήτων** κατασκευάζεται όπως και το πολύγωνο συχνοτήτων, μόνο που συνδέουμε τις πάνω δεξιά κορυφές των ορθογωνίων μεταξύ τους, αντί για τα μέσα των πάνω βάσεων.

Επίσης, εάν οι τάξεις είναι άνισου πλάτους, τότε στον κάθετο άξονα τοποθετούνται οι τιμές f_i / δ_i.

5. Χρονολογικά διαγράμματα ή Χρονοδιαγράμματα (time charts)

Αυτά χρησιμοποιούνται για τη γραφικά απεικόνιση της διαχρονικής εξελίξεως ενός μεγέθους οικονομικού, δημογραφικού ή άλλου είδους. Ο οριζόντιος άξονας χρησιμοποιείται ως άξονας μετρήσεως του χρόνου και ο κάθετος ως άξονας μετρήσεως της εξεταζόμενης μεταβλητής.

6. Διαγράμματα Σημείων (Dot Diagrams)

Όταν έχουμε ένα μικρό μόνο αριθμό παρατηρήσεων, η κατανομή τους μπορεί να περιγραφεί εύκολα με ένα διάγραμμα σημείων. Το διάγραμμα αυτό είναι απλά η τοποθέτηση των διαθέσιμων τιμών πάνω σε ένα ευθύγραμμο τμήμα. Εάν υπάρχουν δυο ή περισσότερες τιμές οι οποίες συμπίπτουν, τοποθετούνται η μια πάνω στην άλλη.

Χαρακτηριστικά μέτρα θέσης και μεταβλητότητας

Εκτός από τις κατάλληλες γραφικές παραστάσεις, είναι απαραίτητα κάποια αριθμητικά μεγέθη, που είναι γνωστά ως αριθμητικά περιγραφικά μέτρα (numerical descriptive measures). Αυτά τα μέτρα χρησιμοποιούνται επίσης για τη θεωρία της στατιστικής συμπερασματολογίας.

Διακρίνονται σε δυο κατηγορίες:

- τα **μέτρα θέσεως (location measures)(central tendancy measures)** που καθορίζουν τη θέση των τιμών στο χώρο
- τα **μέτρα μεταβλητότητας (measures of variability)** που καθορίζουν πως μεταβάλλονται οι τιμές της μεταβλητής.

Μέτρα θέσης για μη ομαδοποιημένες παρατηρήσεις

Μέσος αριθμητικός ή **Μέση τιμή (Mean)** ορίζεται ως το πηλίκο των τιμών της μεταβλητής δια το πλήθος των τιμών της.

Συμβολισμός: μ αν αναφερόμαστε σε πληθυσμό και $\overline{x}$ όταν αναφερόμαστε σε δείγμα.

Τύπος

$$\mu = \frac{x_1 + x_2 + \ldots + x_N}{N} = \frac{\sum_{i=1}^{N} x_i}{N}$$

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Διάμεσος (M) (Median) είναι η τιμή εκείνη της μεταβλητής που χωρίζει το σύνολο των τιμών σε δυο ίσα μέρη, ώστε ο αριθμός των παρατηρήσεων που είναι μικρότερες από το M, να είναι ίσος με τον αριθμό αυτών που είναι μεγαλύτερες από το M. Είναι το σημείο της κατανομής που αφήνει 50% των παρατηρήσεων προς τα πάνω και 50% προς τα κάτω.

Για να βρούμε τη διάμεσο, οι παρατηρήσεις κατατάσσονται κατά τη φυσική τους διάταξη. Στην περίπτωση που οι τιμές της μεταβλητής δεν περιέχονται σε πίνακα συχνοτήτων, η διάμεσος δίνεται από τον όρο (N+1)/2, όπου N το πλήθος των παρατηρήσεων.

Εάν το N είναι **περιττός** αριθμός η διάμεσος είναι η παρατήρηση που βρίσκεται στη (N+1)/2 θέση, γιατί αυτή η παρατήρηση αφήνει (N-1)/2 παρατηρήσεις προς τα κάτω και (N-1)/2 παρατηρήσεις προς τα πάνω.

Ενώ εάν το Ν είναι **άρτιος**, τότε στη μέση των τιμών υπάρχουν δυο τιμές, οπότε η διάμεσος είναι ο μέσος όρος των δυο αυτών μεσαίων τιμών.

Κορυφή (Mode) ή **Επικρατούσα Τιμή (M_o)** είναι η τιμή της μεταβλητής με τη μεγαλύτερη συχνότητα. Η κορυφή δεν καθορίζεται πάντοτε μονοσήμαντα.

Το **κ-ποσοστημόριο ($P_κ$) (Percentiles)** ενός συνόλου τιμών είναι η τιμή εκείνη για την οποία το κ% των παρατηρήσεων είναι μικρότερες από αυτή την τιμή. Π.χ., αν κ=90, τότε το P_{90} είναι η τιμή που αφήνει προς τα κάτω το 90 % των παρατηρήσεων.

Μια πολύ ενδιαφέρουσα κατηγορία των ποσοστημορίων είναι τα **τεταρτημόρια (quartiles)**, που είναι οι τιμές της μεταβλητής που χωρίζουν το σύνολο των τιμών της σε 4 ισοπληθείς ομάδες. Το πρώτο τεταρτημόριο Q_1 είναι η τιμή της μεταβλητής κάτω της οποίας βρίσκεται το 25 % των παρατηρήσεων και το υπόλοιπο 75% βρίσκεται πάνω από αυτήν την τιμή. Όπως είναι προφανές ότι το Q_2 είναι η διάμεσος Μ και τέλος το τρίτο τεταρτημόριο Q_3 είναι η τιμή της μεταβλητής κάτω από την οποία βρίσκεται το 75% των παρατηρήσεων και το υπόλοιπο 25 % πάνω. Στην περίπτωση που τα δεδομένα δεν είναι ταξινομημένα, το Q_1 εντοπίζεται στη θέση (N+1)/4, ενώ το Q_3 στη θέση 3(N+1)/4.

Παρατήρηση

Ακολουθώντας διαδικασία ανάλογη της ευρέσεως των τεταρτημορίων υπολογίζονται απλά και τα **δεκατημόρια (deciles) ($D_κ$)** που είναι εκείνες οι τιμές που χωρίζουν το σύνολο των δεδομένων σε 10 ίσα μέρη.

Υπολογισμός μέτρων θέσεως από πίνακα συχνοτήτων

Για να υπολογίσουμε το μέσο από πίνακα συχνοτήτων, πολλαπλασιάζουμε την κάθε τιμή της μεταβλητής επί την αντίστοιχη συχνότητα (που δείχνει πόσες φορές εμφανίστηκε η τιμή της μεταβλητής στο σύνολο των τιμών) και τα γινόμενα που θα προκύψουν τα αθροίζουμε. Κατόπιν, διαιρούμε το άθροισμα αυτό με το άθροισμα των συχνοτήτων, δηλαδή με το σύνολο των παρατηρήσεων. Οπότε ισχύει:

$$\mu = \frac{x_1 f_1 + \ldots + x_k f_k}{\sum_{i=1}^{k} f_i} = \frac{\sum_{i=1}^{k} x_i f_i}{\sum_{i=1}^{k} f_i}$$

Για τον υπολογισμό της διαμέσου, υπολογίζεται η στήλη των αθροιστικών συχνοτήτων. Στη συνέχεια βρίσκουμε μεταξύ ποιων αθροιστικών συχνοτήτων F_{i-1} και F_i βρίσκεται ο αριθμός N/2. Η τιμή της μεταβλητής που αντιστοιχεί στη μεγαλύτερη αθροιστική

συχνότητα, δηλαδή στην F_i είναι η διάμεσος. Όμοια υπολογίζονται και τα άλλα τεταρτημόρια και ποσοστημόρια.

Μέτρα θέσης για ομαδοποιημένες παρατηρήσεις

Ο **μέσος** υπολογίζεται από τον ακόλουθο τύπο:

$$\mu = \frac{x_1 f_1 + \ldots + x_k f_k}{\sum_{i=1}^{k} f_i} = \frac{\sum_{i=1}^{k} x_i f_i}{\sum_{i=1}^{k} f_i}$$

όπου k ο αριθμός των τάξεων, x_i η κεντρική τιμή της i-τάξεως και f_i η αντίστοιχη συχνότητα.

Για τον υπολογισμό της **διαμέσου**, υπολογίζουμε τη στήλη των αθροιστικών συχνοτήτων. Κατόπιν, εντοπίζουμε μεταξύ ποιών αθροιστικών συχνοτήτων βρίσκεται το N/2 και προσδιορίζουμε την τάξη της μεγαλύτερης αθροιστικής συχνότητας. Η διάμεσος βρίσκεται εντός των ορίων αυτής της τάξης, δίνεται από τον τύπο:

$$M = a_{i-1} + \frac{\delta}{f_i}\left(\frac{N}{2} - F_{i-1}\right)$$

όπου δ το πλάτος της τάξης.

Παρόμοια, υπολογίζονται και τα υπόλοιπα τεταρτημόρια.

$$Q_1 = a_{i-1} + \frac{\delta}{f_i}\left(\frac{N}{4} - F_{i-1}\right)$$

$$Q_3 = a_{i-1} + \frac{\delta}{f_i}\left(\frac{3N}{4} - F_{i-1}\right)$$

Και γενικότερα για τα δεκατημόρια:

$$D_k = a_{i-1} + \frac{\delta}{f_i}\left(\frac{kN}{10} - F_{i-1}\right)$$

Η δε **κορυφή** υπολογίζεται από τον τύπο:

$$M_o = a_{i-1} + \left(\frac{\Delta_1}{\Delta_1 + \Delta_2}\right)\delta$$

όπου a_{i-1} το κατώτερο όριο της τάξης με τη μεγαλύτερη συχνότητα, Δ_1 η διαφορά της συχνότητας αυτής με τη συχνότητα της προηγούμενης τάξης, ενώ Δ_2 η διαφορά της συχνότητας της τάξης αυτής με τη συχνότητα της επόμενης τάξης.

Μέτρα μεταβλητότητας για μη ομαδοποιημένες παρατηρήσεις

Το **εύρος (range) ή έκταση** είναι το πιο απλό και δείχνει το πλάτος των τιμών της μεταβλητής. Υπολογίζεται εύκολα, αφού τοποθετηθούν

οι παρατηρήσεις κατά τη φυσική τους κατάταξη, $R = M - m$, δηλαδή αφαιρούμε από τη μέγιστη τιμή την ελάχιστη.

Το μειονέκτημα του εύρους είναι ότι εξαρτάται από τις ακραίες τιμές της μεταβλητής. Παρόλα αυτά, χρησιμοποιείται και από τους οικονομολόγους σε αρκετές περιπτώσεις (στο χρηματιστήριο για τις τιμές των μετοχών, στη μελέτη του εύρους μεταξύ χαμηλών και υψηλών τιμών κ.ο.κ).

Ένα άλλο μέτρο διασποράς είναι η **μέση απόκλιση (Μ.Α.) (mean deviation)** που ορίζεται ως ο μέσος αριθμητικός των απολύτων διαφορών των τιμών της μεταβλητής από το μ. Δίνεται από τον τύπο:

$$M.A. = \frac{\sum |x_i - \mu|}{N}$$

όπου Ν ο αριθμός των παρατηρήσεων.

Όσο πιο μικρό είναι το αποτέλεσμα, τόσο πιο κοντά στο μ βρίσκονται οι παρατηρήσεις, που σημαίνει ότι τόσο αντιπροσωπευτικός και αξιόπιστος είναι ο μ. Λόγω των απολύτων τιμών, δεν είναι εύκολος ο υπολογισμός του Μ.Α., γι' αυτό χρησιμοποιούνται άλλα μέτρα διασποράς.

Τα πλέον συχνά χρησιμοποιούμενα μέτρα διασποράς είναι η **διακύμανση ή διασπορά (variance)** και η **τυπική απόκλιση (standard deviation)**.

Η **διακύμανση** είναι ο μέσος αριθμητικός των τετραγώνων των διαφορών των τιμών μιας μεταβλητής από το μέσο αριθμητικό της. Συμβολίζεται με σ^2 όταν αναφερόμαστε σε πληθυσμό και με s^2 για δείγμα. Δίνεται από τον τύπο:

$$\sigma^2 = \frac{\sum (x_i - \mu)^2}{N} = \frac{\sum x_i^{\,2}}{N} - \mu^2$$

$$s^2 = \frac{1}{n-1}\sum (x - \bar{x})^2 = \frac{1}{n-1}[\sum x_i^2 - \frac{(\sum x_i)^2}{n}]$$

Παρατηρούμε ότι η δειγματική διασπορά διαιρείται με n-1αντί με n, όπως ο αντίστοιχος τύπος της διασποράς του πληθυσμού. Αυτό γιατί η δειγματική διασπορά χρησιμοποιείται για την εκτίμηση της διασποράς του πληθυσμού και με τον τρόπο που ορίζεται έχει την ιδιότητα της **αμεροληψίας.**

Επειδή η διακύμανση εκφράζεται μέσω του τετραγώνου της μεταβλητής, γι' αυτό παίρνουμε τη θετική τετραγωνική ρίζα της διακύμανσης που ονομάζεται **τυπική απόκλιση** και η οποία

εκφράζεται με τις ίδιες μονάδες μέτρησης με τη μονάδα μέτρησης της μεταβλητής. Η τυπική απόκλιση ορίζεται: $\sigma = \sqrt{\sigma^2}$ και $s = \sqrt{s^2}$.

Όσο μικρότερες είναι οι τιμές της διασποράς και της τυπικής απόκλισης, τόσο πιο συγκεντρωμένες γύρω από το μ βρίσκονται οι τιμές της μεταβλητής. Επίσης, είναι φανερό ότι οι τιμές της διασποράς κυμαίνονται μεταξύ 0 και ∞.

Ένα άλλο μέτρο μεταβλητότητας είναι ο **συντελεστής μεταβλητότητας (coefficient of variation).** Ορίζεται ως εξής:

$$CV = \frac{\sigma}{\mu} 100$$

$$CV = \frac{s}{\bar{x}} 100$$

Είναι καθαρός αριθμός, απαλλαγμένος από μονάδες μέτρησης της μεταβλητής. Και εκφράζει το ‘άπλωμα’ των τιμών σε σχέση με το μέσο. Επίσης, χρησιμοποιείται για συγκρίσεις ομάδων μεταξύ τους (είτε οι ομάδες εκφράζονται με ίδιες μονάδες μέτρησης είτε όχι). Επίσης, χρησιμοποιείται για την εξέταση της ομοιογένειας μέσα στην ίδια ομάδα. Επίσης, όταν ο CV δεν ξεπερνά το 10%, θα λέμε ότι το δείγμα είναι ομοιογενές.

Το **ενδοτεταρτημοριακό εύρος (interquantile range)** είναι η διαφορά του πρώτου από το τρίτο τεταρτημόριο. Στο μεταξύ τους διάστημα το 50% των τιμών της κατανομής. Επομένως, όσο μικρότερο είναι αυτό το διάστημα, τόσο μεγαλύτερη θα είναι η συγκέντρωση των τιμών και άρα μικρότερη η διασπορά των τιμών της μεταβλητής. Το μισό του ενδοτεταρτημοριακού εύρους είναι γνωστό ως **ημιενδοτεταρτημοριακό εύρος (semi-interquantile range)** και συμβολίζεται με Q. Μετριέται με τις ίδιες μονάδες της μεταβλητής και δεν εξαρτάται από όλες τις τιμές αλλά μόνο από εκείνες που περιλαμβάνονται στον υπολογισμό του πρώτου και τρίτου τεταρτημορίου.

Μέτρα μεταβλητότητας για ομαδοποιημένες παρατηρήσεις

Για τη **διασπορά** και την **τυπική απόκλιση** που κυρίως μας ενδιαφέρει, ισχύουν:

$$\sigma^2 = \frac{\sum f_i (x_i - \mu)^2}{N} = \frac{\sum f_i x_i^{\,2}}{N} - \mu^2$$

$$s^2 = \frac{1}{n-1}\sum f_i (x - \overline{x})^2 = \frac{1}{n-1}[\sum f_i x_i^2 - \frac{(\sum f_i x_i)^2}{n}]$$

όπου x_i η κεντρική τιμή της i τάξης και f_i η αντίστοιχη συχνότητα.

Προφανώς, από τα υπόλοιπα μέτρα μεταβλητότητας, υπολογίζονται με βάση της γενικής μεθοδολογίας των ομαδοποιημένων παρατηρήσεων.

Ιδιότητες μέσου και διασποράς

Ιδιότητες του μέσου:

1. Αν σε όλες τις τιμές μιας μεταβλητής προσθέσουμε (ή αφαιρέσουμε) μια σταθερή ποσότητα, τότε ο μ αυξάνεται (ή μειώνεται) κατά την ποσότητα αυτή.
2. Αν όλες οι τιμές της μεταβλητής είναι ίσες με μια σταθερά, τότε ο μ ισούται με την σταθερά αυτή.
3. Αν όλες οι τιμές της μεταβλητής πολλαπλασιασθούν επί μια σταθερά, τότε ο μ πολλαπλασιάζεται με αυτή την σταθερά.
4. Αν από όλες τις τιμές της μεταβλητής αφαιρέσουμε το μ, τότε το άθροισμα των διαφορών x-μ, ισούται με μηδέν.
5. Αν ένας πληθυσμός χωρισθεί σε κ υποπληθυσμούς, που ο καθένας έχει N_1, …, $N_κ$ μονάδες και οι αντίστοιχοι μέσοι είναι $μ_1$, …, $μ_κ$, τότε ο μέσος όλου του πληθυσμού είναι:

$$\mu = \frac{N_1\mu_1 + ... + N_\kappa\mu_\kappa}{N_1 + ... + N_\kappa}$$

Ιδιότητες της διασποράς:

1. Αν έχουμε ότι όλες της μεταβλητής ίσες με μια σταθερά, τότε η διασπορά ισούται με μηδέν.
2. Αν σε όλες τις τιμές μιας μεταβλητής προσθέσουμε (ή αφαιρέσουμε) μια σταθερή ποσότητα, τότε η διασπορά και η διακύμανση παραμένουν αμετάβλητες.
3. Αν όλες οι τιμές μιας μεταβλητής πολλαπλασιασθούν (ή διαιρεθούν) με μια σταθερή ποσότητα, τότε η διασπορά πολλαπλασιάζεται (ή διαιρείται) με το τετράγωνο της ποσότητας αυτής, ενώ η τυπική απόκλιση πολλαπλασιάζεται (ή διαιρείται) με την ποσότητα αυτή.

Μέτρα Ασυμμετρίας (Measures of skewness)

Η κατανομή ενός συνόλου δεδομένων μπορεί να είναι είτε **συμμετρική** είτε **μη συμμετρική.**

Σαν αριθμητικά μέτρα καθορισμού της ασυμμετρίας έχουν προταθεί διάφοροι παράμετροι εκ' των οποίων σπουδαιότεροι είναι οι εξής:

1. Συντελεστές ασυμμετρίας κατά Pearson

Ορίζονται από τις σχέσεις:

$$\gamma_1 = \frac{\overline{x} - M_o}{s}, \gamma_2 = \frac{3(\overline{x} - M)}{s}$$

και λέγονται **πρώτος** και **δεύτερος** συντελεστής ασυμμετρίας του Pearson αντίστοιχα.

Σε περίπτωση μέτριας ασυμμετρίας ισχύει:

$$\gamma_1 \approx \gamma_2 = \gamma$$

Είναι φανερό ότι:

- γ=0, συμμετρία και ισχύει $\overline{x} = M = M_o$
- γ<0, αρνητική ασυμμετρία και ισχύει $\overline{x} < M < M_o$
- γ>0, θετική ασυμμετρία και ισχύει $\overline{x} > M > M_o$

2. Συντελεστής ασυμμετρίας του Bowley

Ένα άλλο μέτρο ασυμμετρίας είναι και η ποσότητα:

$$s_A = \frac{Q_1 + Q_3 - 2M}{Q_3 - Q_1} = \frac{(Q_3 - M) - (M - Q_1)}{Q_3 - Q_1}$$

που λέγεται συντελεστής ασυμμετρίας του Bowley ή τεταρτημοριακός συντελεστής. Παίρνει τιμές μεταξύ –1 και 1. Και ισχύουν τα ακόλουθα:

- $s_A = 0$, συμμετρία κατά Bowley
- $0 < s_A < 1$, θετική ασυμμετρία κατά Bowley
- $-1 < s_A < 0$, αρνητική ασυμμετρία κατά Bowley

Επίσης ισχύουν:

- Αν $s_A = 0$, συμμετρία και ισχύει

$$Q_3 - M = M - Q_1 \Leftrightarrow M = \frac{Q_3 + Q_1}{2}$$

- Αν $s_A = 1$, μεγαλύτερη θετική ασυμμετρία με το πρώτο τεταρτημόριο να προσεγγίζει τη διάμεσο
- Αν $s_A = -1$, μεγαλύτερη αρνητική ασυμμετρία με το τρίτο τεταρτημόριο να τείνει στη διάμεσο

3. Συντελεστής ασυμμετρίας με βάση τις ροπές

Γενικεύοντας την έννοια της διασποράς μπορεί κανείς να ορίσει τις λεγόμενες **κεντρικές ροπές (central moments)** t-τάξης από τη σχέση:

$$\mu_t = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^t}{n}, t = 2,3,...$$

$$\mu_t = \frac{\sum_{i=1}^{n} f_i(x_i - \bar{x})^t}{\sum_{i=1}^{n} f_i}, t = 2,3,...$$

αν t=2, τότε η κεντρική ροπή δεύτερης τάξης συμπίπτει με τη διασπορά.

Ο συντελεστής ασυμμετρίας με βάση τις ροπές (moment coefficient of skewness) ορίζεται σαν το πηλίκο

$$\beta_1 = \frac{\mu_3^2}{\mu_2^2} \text{ η } \beta_1 = \frac{\mu_3}{\mu_2^{3/2}}$$

και εκφράζει τη συμμετρία αν είναι ίσο με μηδέν και αν είναι θετικό δηλώνει ασυμμετρία. Το είδος της ασυμμετρίας καθορίζεται από το πρόσημο της κεντρικής ροπής τρίτης τάξης. Αν $\mu_3 > 0$, θετική ασυμμετρία, ενώ $\mu_3 < 0$, αρνητική ασυμμετρία.

Μέτρα Κύρτωσης (Measures of kurtosis)

Τα μέτρα αυτά αφορούν το βαθμό συγκέντρωσης των δεδομένων γύρω από το μέσο και τα άκρα της κατανομής.

Μια κατανομή η οποία έχει σχετικά μεγάλη συχνότητα (κορυφή) και επομένως μεγάλη συγκέντρωση τιμών γύρω από το μέσο λέγεται **λεπτόκυρτη (leptokurtic)**, ενώ αν η μέγιστη συχνότητά της είναι σχετικά μικρή λέγεται **πλατύκυρτη (platykurtic)**. Κατανομές που προσεγγίζονται από την κανονική κατανομή λέγονται **μεσόκυρτες (mesokurtic)**.

Ένα μέτρο που εκφράζει το βαθμό κυρτότητας μιας κατανομής είναι ο συντελεστής κύρτωσης του Pearson ο οποίος ορίζεται από τον τύπο:

$$\beta_2 = \frac{\mu_4}{\mu_2^2}$$

Επειδή για κανονικές κατανομές έχουμε $\beta_2 = 3$ συνηθίζεται να μετράμε την κυρτότητα με τη διαφορά $\beta_2 - 3$, η οποία για λεπτόκυρτες κατανομές παίρνει θετικές τιμές (θετική κύρτωση), ενώ για πλατύκυρτες κατανομές γίνεται αρνητική (αρνητική κύρτωση).

Τυποποιημένες Τιμές (Standardized Values)

Είναι ένα μέτρο σχετικής θέσης των τιμών σε ένα σύνολο μετρήσεων. Συμβολίζεται ως z-τιμές (z-values).

$$Z = \frac{\mathrm{X} - \mu}{\sigma}$$ (πληθυσμό)

όπου Χ, η συγκεκριμένη παρατήρηση που μας ενδιαφέρει, μ ο μέσος και σ η τυπική απόκλιση. Επίσης, για δείγμα έχουμε αντίστοιχα

$$Z = \frac{X - \overline{x}}{s}$$

Παρατήρηση

Οι τυποποιημένες τιμές είναι καθαροί αριθμοί, γι' αυτό και χρησιμοποιούνται για σύγκριση αποδόσεων που έχουν μετρηθεί σε διαφορετικές κλίμακες.

Παρατήρηση – Εμπειρικός κανόνας

Πριν δοθεί η ερμηνεία των τυποποιημένων τιμών αξίζει να σημειωθεί, ότι αν το ιστόγραμμα των δεδομένων που εξετάζονται μοιάζει με το σχήμα της κανονικής κατανομής (καμπάνα Gauss) τότε:

1. το 68% περίπου των παρατηρήσεων βρίσκεται στο διάστημα με άκρα $\overline{x} \pm s$
2. το 95 % περίπου των παρατηρήσεων βρίσκεται στο διάστημα με άκρα $\overline{x} \pm 2s$

3. το 99% περίπου των παρατηρήσεων βρίσκεται στο διάστημα με άκρα $\overline{x} \pm 3s$
4. ισχύει προσεγγιστικά η σχέση $R \approx 4s$

Ανεξάρτητα πάντως από το αν τα δεδομένα ακολουθούν ή όχι την κανονική κατανομή, το ποσοστό μεταξύ ± n τυπικών αποκλίσεων από τη μέση τιμή είναι τουλάχιστον $(1-\frac{1}{n^2})100\%$.

Παρατήρηση

Τα μέτρα διασποράς αποτελούν, κυρίως για οικονομικές και επιχειρηματικές εφαρμογές, σημαντικό συμπλήρωμα των μελετών που πραγματοποιούνται. Με τη βοήθεια των μέτρων διασπορών, επιτυγχάνεται μια πληρέστερη εικόνα της καταστάσεως που εξετάζεται και κατά συνέπεια οδηγούμαστε σε ασφαλέστερη λήψη ορθών επιχειρηματικών αποφάσεων. Αν για παράδειγμα ένας οικονομολόγος που έχει αναλάβει μια μελέτη για κάποια επιχείρηση και παρατηρήσει ότι η διασπορά των δεδομένων γύρω από το μ είναι μεγάλη, τότε θα πρέπει να εξετάσει πολύ προσεκτικά τις πληροφορίες που λαμβάνονται από το μ και να προβληματιστεί αρκετά για τη λήψη κάποιας απόφασης.

Εφαρμογές στην Οικονομία

Απόδοση Επένδυσης

Η απόδοση μιας επένδυσης υπολογίζεται διαιρώντας το κέρδος δια την αξία της επένδυσης.

Αν η επένδυση έχει απώλειες, τότε εκφράζονται ως αρνητικό κέρδος και η απόδοση είναι αρνητική.

Π.χ. , αν μια επένδυση έχει αξία 100 ευρώ και μετά από ένα χρόνο έχει αξία 108 ευρώ, αυτό σημαίνει ότι η απόδοση της είναι 8%.

Αν η αξία της επένδυσης μετά από ένα χρόνο έχει πέσει στα 80 ευρώ, τότε η απόδοση της είναι -20%.

Η απόδοση , για πολλές επενδύσεις όπως απλές μετοχές και χαρτοφυλάκια μετοχών, είναι μια μεταβλητή. Οπότε ο επενδυτής δεν μπορεί να γνωρίζει την απόδοση της επένδυσής του τη στιγμή της αγοράς. Αν η απόδοση είναι αρνητική, ο επενδυτής θα χάσει ένα μέρος από τα χρήματά του.

Οι επενδυτές έχουν δυο στόχους:

1. τη μεγιστοποίηση της απόδοσης
2. την ελαχιστοποίηση του κινδύνου μιας μετοχής

Από την κατασκευή του ιστογράμματος των αποδόσεων μιας επένδυσης μπορούμε να πάρουμε τις εξής πληροφορίες:

1. η κεντρική θέση της κατανομής μας δίνει πληροφορία για την απόδοση που μπορούμε να αναμένουμε
2. το εύρος, η μεταβλητότητα της κατανομής μας δίνει πληροφορία για τον βαθμό κινδύνου

Αν η μεταβλητότητα είναι μικρή, ο επενδυτής μπορεί να πιστέψει ότι η απόδοση θα είναι κάπου κοντά στην κεντρική θέση.

Αντιθέτως, αν η μεταβλητότητα είναι μεγάλη, η απόδοση γίνεται απρόβλεπτη και η πιθανότητα κινδύνου είναι αυξημένη.

Σύγκριση των αποδόσεων δυο επενδύσεων

Ένας επενδυτής θέλει να επιλέξει μεταξύ δυο επενδύσεων. Για να μπορέσει να επιλέξει βρίσκει όλες τις προηγούμενες αποδόσεις (σε ποσοστά %) των δυο επενδύσεων, που είναι συγκεντρωμένες στους παρακάτω πίνακες.

Επένδυση Α

30	-15,83	8,47	22,92	-5,29
-2,13	0,63	36,08	20,95	-7,04
4,30	38	-21,95	43,71	-12,11
25	6,93	10,33	-12,83	12,89
12,89	-13,24	12,68	0,52	63
-20,24	-18,95	13,09	61	-19,27
1,20	-9,43	13,77	-11,96	-9,22
-2,59	1,21	22,42	1,94	-17
33	31,76	34,40	28,45	17,30
14,26	11,07	49,87	-8,55	52

Επένδυση Β

30,33	8,29	39,04	1,53	-10,01
-30,37	61	24,76	17,61	35,24
-5,61	-20,44	15,28	1,20	40,70
29	-34,75	34,21	9,94	22,18
-26,01	54,19	52	-33,39	3,24
0,46	44	-32,17	58,67	25,10
2,07	-20,23	30,31	0,25	-24,24

29,44	4,16	6,06	5,23	-38,47
11	10,03	14,73	66	13,44
-25,93	10,51	36,13	24,30	68

Από δεδομένα αυτά ο επενδυτής θα ήθελε να μάθει ποιο είναι το αναμενόμενο ύψος της απόδοσης και ποιος είναι ο κίνδυνος της κάθε επένδυσης**.**

Για παράδειγμα, αν η μεταβλητότητα των τιμών είναι μεγάλη αυτό σημαίνει ότι η επένδυση δεν παρουσιάζει σταθερότητα και είναι πιθανό μετά από μια υψηλή απόδοση να ακολουθήσει μια χαμηλή ή ακόμη και αρνητική.

Ο επενδυτής αποφάσισε να κατασκευάσει ένα ιστόγραμμα για κάθε πίνακα και να συγκρίνει τα χαρακτηριστικά των δυο επενδύσεων.

Παρατηρώντας τα ιστογράμματα διαπιστώνουμε τα εξής:

1. Οι κορυφές και των δυο αποδόσεων βρίσκονται στην ίδια τάξη με αποδόσεις μεταξύ 0-15%
2. Μόνο η επένδυση Β παρουσιάζει στο αριστερό άκρο τιμές μικρότερες από το -30
3. Τα δυο ιστογράμματα εμφανίζουν ελαφρά θετική ασυμμετρία με αυτό της επένδυσης Β να είναι μετατοπισμένο προς τα δεξιά (υψηλότερες τιμές). Με εξαίρεση την τάξη στο αριστερό άκρο που δείχνει μια πιθανότητα για υψηλές αρνητικές αποδόσεις

Από τις παραπάνω παρατηρήσεις διαπιστώνεται ότι η επένδυση Β έχει συνολικά καλύτερες αποδόσεις, αλλά περιέχει και έναν όχι ασήμαντο κίνδυνο για σχετικά υψηλές ζημιές.

Ο επενδυτής έχοντας αυτά υπόψη του, δύναται να διαλέξει την επένδυση που του ταιριάζει καλύτερα ανάλογα με την ιδιοσυγκρασία του και την οικονομική του κατάσταση.

Η ερμηνεία των γραφικών παραστάσεων είναι υποκειμενική. Για περισσότερα συμπεράσματα καλό είναι να χρησιμοποιούνται οι αριθμητικοί δείκτες (που θα αναλύσουμε παρακάτω).

Σύγκριση αποδόσεων δυο επενδύσεων

Όπως αναφέραμε σε προηγούμενη άσκηση θέλουμε να συγκρίνουμε δυο επενδύσεις και να επιλέξουμε την καλύτερη. Μια επένδυση κρίνεται από την αναμενόμενη απόδοση και το βαθμό κινδύνου που περιέχει.

Σε ένα ιστόγραμμα η αναμενόμενη απόδοση φαίνεται από το κέντρο του ιστογράμματος και ο κίνδυνος από το εύρος. Επειδή το ιστόγραμμα δίνει μόνο μια οπτική εικόνα για να συγκρίνουμε δυο ιστογράμματα μπορεί οι οπτικές διαφορές να μην είναι εμφανείς.

Αντί για τις γραφικές παραστάσεις μπορούμε να χρησιμοποιήσουμε τους αντίστοιχους αριθμητικούς δείκτες. Όπου για το κέντρο των δεδομένων είναι ο αριθμητικός μέσος και η διάμεσος, ενώ για το βαθμό κινδύνου είναι η διασπορά και η τυπική απόκλιση.

Παίρνουμε τα ακόλουθα:

Επένδυση Α

$N = 50$

$\overline{x} = 10.95$

$s = 21.89$

$s^2 = 479.35$

$CV = 200.03$

$Q_1 = -7.42$

$M = 9.88$

$Q_3 = 25.86$

$Max = 63$

$Min = -21.95$

$R = 84.95$

$Kurtosis = -0.32$

Επένδυση Β

$N = 50$

$\overline{x} = 12.76$

$s = 28.05$

$s^2 = 786.62$

$CV = 219.80$

$Q_1 = -1.22$

$M = 10.76$

$Q_3 = 31.30$

$Max = 68$

$Min = -38.47$

$R = 106.47$

$Kurtosis = -0.62$

Ερμηνεία

Παρατηρούμε ότι η επένδυση Β έχει υψηλότερο αριθμητικό μέσο και διάμεσο, ενώ η επένδυση Α έχει μικρότερη διασπορά και τυπική απόκλιση. Αν ο επενδυτής ενδιαφέρεται για τοποθέτηση χαμηλού κινδύνου θα πρέπει να επιλέξει την επένδυση Α. Επανεξετάζοντας τα ιστογράμματα παρατηρούμε ότι τα αριθμητικά μέτρα προσφέρουν μεγαλύτερη ακρίβεια πληροφοριών.

Σχολιασμός και των άλλων μέτρων.

Ιδιότητες Διαμέσου

1. Αν το πλήθος των δεδομένων είναι μεγάλο και οι παρατηρήσεις δεν επαναλαμβάνονται τακτικά τότε περίπου οι μισές παρατηρήσεις είναι μικρότερες της διαμέσου και οι υπόλοιπες μισές μεγαλύτερες της διαμέσου.
2. Η διάμεσος δεν επηρεάζεται από παρατηρήσεις που απέχουν πολύ από τον κύριο όγκο των δεδομένων. Το αντίθετο συμβαίνει με τον αριθμητικό μέσο του οποίου η τιμή είναι ευαίσθητη σε τέτοιες παρατηρήσεις.

Σύγκριση Μέτρων Θέσης

	Μέσος	**Διάμεσος**	**Κορυφή**
Πόσο μεγάλη είναι η χρήση του	μέγιστη	μεγάλη	Χρησιμοποιείται μερικές φορές
Επηρεάζεται από ακραίες τιμές	ναι	όχι	Μερικές φορές δεν υπάρχει
Σχόλια	Χρησιμοποιείται σε πολλές στατιστικές μεθόδους	Εάν υπάρχουν ακραίες τιμές συνιστάται η χρήση της	Κατάλληλη για κατηγορικές μεταβλητές

Ορισμός

Παράμετρος

Οι αριθμητικές τιμές που συνοψίζουν τα δεδομένα του πληθυσμού.

Πλεονεκτήματα και Μειονεκτήματα των μέτρων κεντρικής τάσεως και διασποράς

Κορυφή

Πλεονεκτήματα

- Δείχνει την πιο συχνή τιμή της κατανομής
- Μένει ανεπηρέαστη από ακραίες τιμές
- Μπορεί να υπολογιστεί όταν οι ακραίες τιμές είναι άγνωστες
- Δίνει περισσότερες πληροφορίες από τον μέσο όρο για την κατανομή , όταν η κατανομή είναι της μορφής U

Μειονεκτήματα

- Δεν λαμβάνει υπόψη την ακριβή τιμή του κάθε στοιχείου
- Δεν μπορεί να χρησιμοποιηθεί για να υπολογιστούν οι παράμετροι του πληθυσμού
- Δεν είναι πολύ χρήσιμη για μικρό αριθμό δεδομένων, στα οποία αρκετές τιμές εμφανίζονται το ίδιο συχνά (π.χ. 1,1,2,3,3,4)
- Δεν μπορεί να χρησιμοποιηθεί με ακρίβεια όταν έχουμε ομαδοποιημένη κατανομή

Διάμεσος

Πλεονεκτήματα

- Είναι πιο εύκολο να υπολογιστεί σε σχέση με τον μέσο όρο (εκτός αν χρησιμοποιηθεί ο μαθηματικός τύπος υπολογισμού της)
- Δεν επηρεάζεται από τις ακραίες τιμές, οπότε είναι καλύτερος δείκτης κεντρικής τάσεως όταν έχουμε ασύμμετρη κατανομή
- Μπορεί να υπολογιστεί ακόμα και όταν δεν γνωρίζουμε τις ακραίες τιμές.

Μειονεκτήματα

- Δεν λαμβάνει υπόψη την ακριβή τιμή του κάθε στοιχείου
- Δεν μπορεί να χρησιμοποιηθεί για να υπολογιστούν παράμετροι του πληθυσμού
- Αν οι τιμές της κατανομής είναι λίγες, τότε η διάμεσος δεν μπορεί να τις αντιπροσωπεύσει με ακρίβεια. Π.χ. Αν έχουμε τις τιμές 2,5,8,67 και 110 η διάμεσος είναι το 8.

Μέση τιμή

Πλεονεκτήματα

- Είναι εύκολος στον υπολογισμό του.
- Αντικατοπτρίζει πιο πιστά την κεντρική τιμή της κατανομής σε σχέση με τους άλλους δείκτες
- Μπορεί να χρησιμοποιηθεί για τον υπολογισμό των παραμέτρων του πληθυσμού (παραμετρικά τεστ)

Μειονεκτήματα

- Είναι ευαίσθητη στις τιμές των δεδομένων της κατανομής. Π.χ. μια αλλαγή σε οποιαδήποτε τιμή προκαλεί διαφοροποίηση της μέσης τιμής.
- Επειδή υπολογίζεται αλγεβρικά, η τιμή της είναι πιθανόν να μην ανήκει στις τιμές της κατανομής
- Είναι πολύ ευαίσθητη στις ακραίες τιμές

Παρατήρηση

Ποιος δείκτης κεντρικής τάσεως είναι καλύτερος εξαρτάται από το σχήμα της κατανομής.

Εύρος

Πλεονεκτήματα

- Είναι πολύ εύκολο στον υπολογισμό του.
- Περιλαμβάνει και τις ακραίες τιμές της κατανομής

Μειονεκτήματα

- Αλλοιώνεται από τις ακραίες τιμές, με αποτέλεσμα να μην παρουσιάζει σε πολλές περιπτώσεις, μια αντιπροσωπευτική εικόνα της διασποράς της κατανομής
- Δεν παρέχει καμία πληροφορία σχετικά με τη διασπορά των τιμών μεταξύ των άκρων της κατανομής. Π.χ., δεν μας «λέει» τίποτα για τη διασπορά των τιμών της κατανομής γύρω από τη μέση τιμή

Ενδοτεταρτημοριακό εύρος

Πλεονεκτήματα

- Δεν επηρεάζεται από τις ακραίες τιμές
- Είναι σχετικά εύκολο στον υπολογισμό του
- Είναι αντιπροσωπευτικό των κεντρικών τιμών της κατανομής

Μειονεκτήματα

- Δεν λαμβάνει υπόψη τις ακραίες τιμές της
- Όπως και το εύρος, δεν επιτρέπει την ακριβή ερμηνεία μιας συγκεκριμένης τιμής της κατανομής
- Δεν είναι ακριβές όταν τα δεδομένα είναι ομαδοποιημένα κατά μεγάλα διαστήματα τιμών
- Όπως και η διάμεσος δεν περιγράφει καμία από τις παραμέτρους του πληθυσμού που είναι βασικές για την επαγωγική στατιστική

Τυπική Απόκλιση

Πλεονεκτήματα

- Μπορεί να χρησιμοποιηθεί για τον υπολογισμό των παραμέτρων του πληθυσμού
- Λαμβάνει υπόψη όλες τις τιμές της κατανομής
- Είναι ο πιο ευαίσθητος από τους δείκτες διασποράς

Μειονεκτήματα

- Ο υπολογισμός της είναι σχετικά πιο περίπλοκος σε σχέση με τους υπόλοιπους δείκτες
- Είναι πολύ ευαίσθητη στις ακραίες τιμές της κατανομής.

Ερωτήσεις

1. Πότε αποτελεί η διάμεσος καλύτερο μέτρο της τυπικής τιμής μιας ομάδας από ότι ο μέσος;

Η διάμεσος είναι προτιμότερη όταν στον πληθυσμό υπάρχουν μερικές πολύ μεγάλες ή πολύ μικρές τιμές.

2. Υποθέστε ότι έχετε αναλάβει τη διεξαγωγή μιας έρευνας κατά την οποία οι ερωτώμενοι καλούνται να υποδείξουν πόσο τους αρέσουν οι λουκουμάδες (χρησιμοποιώντας μια κλίμακα από το 1 ως το 10). Αν οι μισοί άνθρωποι λατρεύουν τους λουκουμάδες και οι άλλοι μισοί τους απεχθάνονται, πως θα μοιάζει η κατανομή των συχνοτήτων; Ποιο είναι το καλύτερο μέτρο της κεντρικής τάσης αυτής της κατανομής;

Αυτού του τύπου η κατανομή ονομάζεται δικόρυφη και το διάγραμμα συχνοτήτων της μοιάζει με καμήλα με 2 καμπούρες. Ούτε ο μέσος ούτε η διάμεσος αποτελεί καλή ένδειξη της τυπικής τιμής. Με άλλα λόγια, δεν υπάρχει «καλύτερο μέτρο».

3. Γιατί θα πρέπει να γνωρίζετε τη διασπορά; Των καθημερινών σας πωλήσεων αν διευθύνετε μια εμπορική επιχείρηση;

Αν τα στοιχεία των πωλήσεων έχουν σχετικά μικρή διασπορά, ο προγραμματισμός της επιχείρησης σας διευκολύνεται σημαντικά. Αν η διασπορά είναι μεγάλη, θα πρέπει να αναπτύξετε κάποιον τρόπο για να αντιμετωπίσετε αυτού του είδους την αβεβαιότητα. Π.χ., θα μπορούσατε, ίσως να φροντίσετε ώστε το επίπεδο των αποθεμάτων σας να έχει και αυτό μεγάλη διασπορά. Θα πρέπει να εξετάσετε τους δυο τύπους προβλημάτων που ενδέχεται να προκύψουν από τις ακανόνιστες πωλήσεις: α. μερικές μέρες θα σας απομένουν απούλητα εμπορεύματα και β. κάποιες άλλες μέρες μπορεί να μην σας επαρκούν τα αποθέματα των εμπορευμάτων σας. Οι αποφάσεις που θα πάρετε θα πρέπει να εξαρτηθούν από την σχετική σοβαρότητα αυτών των δύο τύπων προβλημάτων.

4. Γιατί χρησιμοποιούνται μερικές φορές ανοικτές ακραίες τάξεις;

Δεν είναι πάντα δυνατή η ύπαρξη ακριβών δεδομένων για τις ακραία υψηλές ή χαμηλές τιμές μιας κατανομής.

5. Προκαλούν προβλήματα οι ανοικτές ακραίες τάξεις κατά τον υπολογισμό του μέσου; Κατά τον υπολογισμό της διαμέσου;

Στις ανοικτές ακραίες τάξεις ο υπολογισμός του μέσου είναι δύσκολος, αλλά για τον υπολογισμό της διαμέσου δεν υπάρχει πρόβλημα.

6. Αν κατέφτανε ξαφνικά μια ομάδα πανύψηλων παικτών μπάσκετ στον πληθυσμό σας, τι θα συνέβαινε στο διάγραμμα συχνοτήτων των υψών; Ποιες

θα ήταν οι επιπτώσεις αυτής της άφιξης στο μέσο, τη διάμεσο, και τη διασπορά;

Θα «ψήλωναν» οι ράβδοι του διαγράμματος συχνοτήτων των υψών που αντιστοιχούν στα μεγαλύτερα ύψη. Ο μέσος θα αυξανόταν σημαντικά. Η διάμεσος θα αυξανόταν και αυτή. Αλλά κατά πόσο; Πιθανότατα όχι τόσο πού όσο ο μέσος και η διασπορά.

7. Πότε είναι προτιμότερο το κυκλικό διάγραμμα από το ιστόγραμμα;

Το κυκλικό διάγραμμα δίνει πιο ξεκάθαρη εικόνα της κατανομής των στοιχείων κάθε κατηγορίας από ότι το ιστόγραμμα.

8. Υποθέστε ότι κατασκευάζετε ρούχα και ότι η αγορά στόχος σας δεν είναι ούτε οι πολύ ψηλοί ούτε οι πολύ κοντοί. Αν θέλετε τα προϊόντα σας να τα διαθέσετε στο μισό του πληθυσμού, ποιο μέτρο των υψών του πληθυσμού σας θα σας ενδιέφερε περισσότερο;

Θα ήταν πολύ χρήσιμο να γνωρίζετε το ενδοτεταρτημοριακό εύρος γιατί οι μισές τιμές περιέχονται μεταξύ του τρίτου και του πρώτου τεταρτημορίου.

Πιθανότητες

4.1 Εισαγωγή

Σε αυτό το κεφάλαιο θα δούμε μερικές βασικές έννοιες των Πιθανοτήτων που είναι απαραίτητες στην ανάλυση των τυχαίων δειγμάτων και στην εξαγωγή συμπερασμάτων.

Ορισμοί

Τυχαίο πείραμα

Είναι μια διαδικασία, μια πράξη που μπορεί να επαναληφθεί όσες φορές θέλουμε και το αποτέλεσμά του δεν είναι γνωστό εκ των προτέρων. Για παράδειγμα, η γέννηση ενός παιδιού, το τράβηγμα ενός χαρτιού, η ρίψη ενός ζαριού κ.ο.κ.

Δειγματικός χώρος

Το σύνολο όλων των δυνατών αποτελεσμάτων ενός πειράματος ονομάζεται **δειγματικός χώρος ή δειγματοχώρος** και συμβολίζεται με Ω.

Έτσι, για παράδειγμα στο πείραμα της γέννησης Ω={αγόρι, κορίτσι}, στην ρίψη ενός ζαριού Ω={1,2,3,4,5,6},στη μέτρηση του ύψους ενός ατόμου, το Ω παίρνει όλες τις τιμές από 1 έως 3 μέτρα. Άρα ο δειγματικός χώρος μπορεί να είναι πεπερασμένος ή άπειρος.

Τα υποσύνολα του Ω ονομάζονται **γεγονότα**. Το γεγονός Ω καλείται **βέβαιο γεγονός.** Ενώ το κενό σύνολο (το γεγονός που δεν πραγματοποιείται ποτέ) καλείται **αβέβαιο ή αδύνατο γεγονός.**

Αν έχουμε δυο γεγονότα Α και Β ενός χώρου Ω και πραγματοποιούνται και τα δυο, τότε λέμε ότι πραγματοποιείται η τομή τους. Αν πραγματοποιείται το ένα από τα δυο, τότε λέμε ότι πραγματοποιείται η ένωσή τους. Αν δεν συμβαίνει το Α, τότε πραγματοποιείται το Α΄ που καλείται συμπλήρωμα ή αντίθετο του Α. Τα Α , Α΄ καλούνται αντίθετα ή συμπληρωματικά.

Αν δυο γεγονότα δεν έχουν κανένα κοινό στοιχείο, ονομάζονται **ξένα ή ασυμβίβαστα.** Η πραγματοποίηση του ενός αποκλείει την πραγματοποίηση του άλλου.

Συνοπτική παρουσίαση βασικών εννοιών

Πιθανοθεωρητική ερμηνεία	**Μαθηματική ερμηνεία**	**Συμβολισμός**
Δειγματικός χώρος	Βασικό σύνολο	Ω
Δυνατά αποτελέσματα ενός πειράματος	Στοιχεία του Ω	$\omega \in \Omega$
Γεγονός	Υποσύνολο του Ω	$A \subseteq \Omega$
Εμφάνιση ενός εκ των γεγονότων Α, Β, Γ,..,Κ	Ένωση των Α, Β,…,Κ	$A \cup B \cup \Gamma \cup \ldots \cup K$
Ταυτόχρονη εμφάνιση όλων των γεγονότων Α,	Τομή των Α, Β, Γ,..,Κ	$A \cap B \cap \Gamma \cap \ldots \cap K$

Β, Γ,..,Κ		
Όχι εμφάνιση του Α	Συμπλήρωμα του Α	Α΄
Αδύνατο γεγονός	Κενό σύνολο	$\varnothing$
Βέβαιο γεγονός	Βασικό σύνολο	Ω
Ασυμβίβαστα γεγονότα	Ξένα σύνολα	$A \cap B = \varnothing$

Παρόλο που η λέξη πιθανότητα χρησιμοποιείται ευρέως είναι δύσκολο να δοθεί ένας ακριβής ορισμός της. Κατά καιρούς έχουν δοθεί διάφοροι ορισμοί. Οι πιο βασικοί θεωρούνται: α. ο κλασικός ορισμός, β. η πιθανότητα σαν όριο της σχετικής συχνότητας, γ. ο αξιωματικός ορισμός της πιθανότητας και δ. υποκειμενική πιθανότητα.

Η έννοια της πιθανότητας αναφέρεται σε κάποιο γεγονός. Δηλαδή, μιλάμε για την πιθανότητα πραγματοποίησης κάποιου γεγονότος σε ένα συγκεκριμένο τυχαίο πείραμα. Συμβολίζεται με P(A).

4.2 Η έννοια της πιθανότητας

Ο κλασικός ορισμός της πιθανότητας προϋποθέτει την ομοιομορφία του δειγματικού χώρου Ω . Δηλαδή, ο Ω να είναι πεπερασμένος με κ στοιχεία και ότι η πιθανότητα κάθε απλού γεγονότος (το γεγονός που αποτελείται από ένα μόνο στοιχείο) να είναι ίση με 1/κ, δηλαδή τα

απλά γεγονότα να είναι ισοπίθανα. Με βάση αυτές τις προϋποθέσεις η πιθανότητα ορίζεται σαν το πηλίκο του πλήθους των στοιχείων του Α δια το πλήθος των στοιχείων του Ω. Δηλαδή, P(A)=ευνοϊκές περιπτώσεις / δυνατές περιπτώσεις.

Όπως ήδη έχει αναφερθεί ένα πείραμα δύναται να επαναληφθεί πολλές φορές κάτω από τις ίδιες συνθήκες. Αν λοιπόν στις N επαναλήψεις του πειράματος, ένα γεγονός Α εμφανίστηκε N_A φορές, τότε

$$P(A) = \frac{N_A}{N}$$

που αποτελεί έναν πιο αυστηρό ορισμό της πιθανότητας.

Ισχύουν οι εξής ιδιότητες:

1. $0 \leq P(A) \leq 1$
2. $P(\Omega) = 1$ και $P(\varnothing) = 0$
3. Αν $A \subseteq B$, τότε $P(A) \leq P(B)$
4. Αν Α και Β ξένα μεταξύ τους, τότε

$$P(A \cup B) = P(A) + P(B)$$

Αν δεν είναι ξένα, τότε

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

5. Αν Α και Α΄ είναι συμπληρωματικά, τότε

$$P(A) + P(A') = 1$$

Όσο αφορά τους δυο άλλους ορισμούς της πιθανότητας δεν θα αναφερθούμε με λεπτομέρειες γιατί μας ενδιαφέρει να ασχοληθούμε με τις βασικές έννοιες των Πιθανοτήτων και της Στατιστικής και όχι με τη μαθηματική θεμελίωσή τους.

Παράδειγμα

Ρίχνουμε ένα ζάρι 100 φορές και η όψη Κ εμφανίζεται 53 φορές, τότε αν ονομάσουμε Α= εμφάνιση του Κ, το P(A)=53/100=0.53

4.3 Δεσμευμένη πιθανότητα

Πολλές φορές κατά την εκτέλεση ενός πειράματος είναι γνωστή η πραγματοποίηση ενός γεγονότος Β. Είναι εύλογο το ερώτημα, πόσο επηρεάζει αυτή η πληροφορία την πιθανότητα εμφάνισης ενός άλλου γεγονότος Α.

Σε αυτές τις περιπτώσεις μιλάμε για τη δεσμευμένη πιθανότητα του Α, όταν γνωρίζουμε ότι το γεγονός Β έχει πραγματοποιηθεί. Δηλαδή,

ζητείται η πιθανότητα πραγματοποίησης του Α δοθέντος ότι το γεγονός Β έχει πραγματοποιηθεί. Συμβολίζεται με P(A|B) και ορίζεται:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}, \text{ με } P(B) \neq 0$$

Από τον παραπάνω τύπο προκύπτει εύκολα ο πολλαπλασιαστικός νόμος της πιθανοθεωρίας:

$$P(A \cap B) = P(B)P(A \mid B)$$

Παράδειγμα

Έστω ότι ζητείται να ευρεθεί η πιθανότητα να τραβήξουμε ‘ρήγα’ από την τράπουλα δοθέντος ότι ήδη έχει τραβηχτεί μια φιγούρα.

Έστω Α = το χαρτί είναι ρήγας και Β = το χαρτί είναι φιγούρα. Ζητείται το P(A|B). Το P(A ∩ B)= 4/ 52 και το P(B) = 12/52, άρα P(A|B) = 1/3.

4.4 Ανεξάρτητα γεγονότα

Ορισμός

Τα γεγονότα Α και Β λέγονται ανεξάρτητα αν ισχύει:

$$P(A \cap B) = P(A)P(B)$$

Επίσης, ισχύει $P(A \mid B) = P(A)$ και $P(B \mid A) = P(B)$.

Ο ορισμός της ανεξαρτησίας είναι ισοδύναμος με το ότι τα γεγονότα Α και Β είναι ανεξάρτητα αν το γεγονός ότι το ένα έχει συμβεί δεν επηρεάζει την πιθανότητα να συμβεί το άλλο.

Επίσης, από τον ορισμό προκύπτει ότι κάθε γεγονός είναι ανεξάρτητο από το αδύνατο και από το βέβαιο γεγονός.

4.5 Θεώρημα Ολικής Πιθανότητας-Τύπος Bayes. Εφαρμογές

Έστω ότι ένα βέβαιο γεγονός Ε, εξαρτάται από την εμφάνιση των γεγονότων $A_1,.., A_n$, που είναι ανά δυο ξένα μεταξύ τους. Οπότε ισχύει:

$$E = E \cap (A_1 + A_2 + ... + A_n) = (E \cap A_1) + ... + (E \cap A_n)$$

Επίσης ισχύει:

$$P(E) = P(E \cap A_1) + ... + P(E \cap A_n)$$

Από τον τύπο της δεσμευμένης πιθανότητας έχουμε:

$$P(E \mid A_i) = \frac{P(E \cap A_i)}{P(A_i)} \Leftrightarrow P(E \cap A_i) = P(E \mid A_i)P(A_i)$$

Οπότε

$$P(E) = P(A_1)P(E \mid A_1) + \ldots + P(A_n)P(E \mid A_n)$$

Αν χρησιμοποιήσουμε ξανά τον τύπο της δεσμευμένης πιθανότητας

$$P(A_i \mid E) = \frac{P(A_i \cap E)}{P(E)}$$

και θέσουμε $P(A_i \cap E) = P(A_i)P(E \mid A_i)$ θα έχουμε:

$$P(A_i \mid E) = \frac{P(A_i)P(E \mid A_i)}{P(E) = P(A_1)P(E \mid A_1) + \ldots + P(A_n)P(E \mid A_n)} =$$

$$= \frac{P(A_i)P(E \mid A_i)}{\sum P(A_i)P(E \mid A_i)}$$

Η πιθανότητα $P(A_i)$ ονομάζεται **εκ των υστέρων (a priori)**, ενώ η πιθανότητα $P(A_i \mid E)$ καλείται **εκ των προτέρων γνωστή πιθανότητα (a posteriori)**.

Παράδειγμα

Έστω ότι η πιθανότητα σωστής διάγνωσης με κάποιο τεστ (θετικής ή αρνητικής) ενός καρδιακού νοσήματος είναι 95%. Αν γνωρίζουμε ότι 5% του πληθυσμού πάσχει από την ασθένεια, ποια είναι η πιθανότητα

για ένα άτομο για το οποίο το τεστ είναι θετικό είναι πράγματι ασθενής.

Έστω τα ακόλουθα γεγονότα:

Α = το γεγονός ότι ένα άτομο είναι ασθενής

Θ = το τεστ είναι θετικό σε έναν ασθενή

Άρα, $P(\Theta \mid \mathrm{A}) = 0.95$ και P(Θ΄|Α΄) = 0.95. Ζητείται η P(Α|Θ), δηλαδή ποια είναι η πιθανότητα να είναι κάποιος ασθενής όταν το τεστ είναι θετικό.

Σύμφωνα με τον τύπο του Bayes έχουμε:

$$P(\mathrm{A} \mid \Theta) = \frac{P(\Theta \mid \mathrm{A})P(\mathrm{A})}{P(\Theta \mid \mathrm{A})P(\mathrm{A}) + P(\Theta \mid \mathrm{A}')P(\mathrm{A}')}$$

Από τον παραπάνω τύπο δεν ξέρουμε την πιθανότητα P(Θ|Α΄), που είναι η πιθανότητα 'λάθος διάγνωσης'. Δηλαδή, P(Θ|Α΄)=1-P(Θ|Α)=1-0.95=0.05. Επίσης, P(Α΄)=1-P(Α)=0.95. Οπότε με αντικατάσταση προκύπτει P(Α|Θ)=0.5

Ασκήσεις

1. Έστω ότι η πιθανότητα σωστής διάγνωσης του καρκίνου της μήτρας με το τεστ Παπανικολάου είναι 0.95. Αν το ποσοστό των γυναικών που έχουν καρκίνο της μήτρας είναι 0.0001, ποια η πιθανότητα για μια γυναίκα με θετικό τεστ να είναι πράγματι ασθενής;
2. Για να διαπιστωθεί αν ένα άτομο πάσχει από καρκίνο γίνεται ακτινογραφία η οποία δίνει σωστή απάντηση στις 85% των περιπτώσεων όταν το άτομο πάσχει από καρκίνο και στις 95% των περιπτώσεων όταν το άτομο δεν πάσχει από καρκίνο. Από τα άτομα που πάνε για εξέταση μόνο το 30% πάσχει από καρκίνο.

I. Σε ποιο ποσοστό των ατόμων που εξετάζονται η ακτινογραφία είναι θετική; (δηλαδή, δείχνει ότι το άτομο πάσχει από καρκίνο;)

II. Σε ποιο ποσοστό των ατόμων που η ακτινογραφία είναι αρνητική, γίνεται λάθος διάγνωση;

3. Αν τα γεγονότα Α και Β είναι ανεξάρτητα, τότε είναι επίσης είναι ανεξάρτητα και τα : α. Α΄, Β β. Α, Β΄ γ. Α΄, Β΄
4. Μια κάλπη περιέχει 5 σφαιρίδια λευκά, 4 κόκκινα και 3 μαύρα. Μια άλλη περιέχει 5 σφαιρίδια λευκά, 6 κόκκινα και 7 μαύρα. Εξάγουμε ένα σφαιρίδιο από κάθε κάλπη. Ποια είναι η πιθανότητα τα δυο σφαιρίδια που έχουν εξαχθεί να είναι του ίδιου χρώματος;
5. Ένα εργοστάσιο παραγωγής τηλεοράσεων αποτελείται από τα τμήματα Α και Β, τέτοια ώστε τα τμήματα αυτά παράγουν 65% και

35% αντίστοιχα της συνολικής παραγωγής του εργοστασίου. Αν οι μη ελαττωματικές τηλεοράσεις είναι 95% και 92% για τα τμήματα Α και Β αντίστοιχα, ποια η πιθανότητα να επιλεγεί μια μη ελαττωματική τηλεόραση, όταν αυτή εκλέγεται στην τύχη από την ετήσια συνολική παραγωγή.

6. Μεταξύ 20 φοιτητών οι 4 είναι του πρώτου έτους και οι υπόλοιποι του δεύτερου. Τρεις από αυτούς εκλέγονται στην τύχη και τους ρωτάμε για το έτος φοίτησής τους. Ποια η πιθανότητα: α. οι δυο πρώτοι να είναι του δεύτερου έτους και ο τρίτος του πρώτου, β. ο τρίτος να είναι πρωτοετής δοθέντος οι δυο πρώτοι είναι δευτεροετείς;
7. Σε μια χώρα το 7% των εγκύων γυναικών πάσχει από πυονεφρίτιδα, το 5% από βακτηριουρία και το 1% και από τα δυο. Ποιο ποσοστό των εγκύων γυναικών της χώρας αυτής πάσχει τουλάχιστον από μια ασθένεια;
8. Ας θεωρήσουμε τις στατιστικές γεννήσεων για τον πληθυσμό των ΗΠΑ το 1987. Σύμφωνα με αυτά τα δεδομένα, οι πιθανότητες για μια τυχαία επιλεγμένη γυναίκα που γέννησε το 1987, να ανήκει σε κάθε μια από τις ακόλουθες ηλιακές ομάδες είναι:

Ηλικία	Πιθανότητα
<15	0,0027

15-19	0,1214
20-24	0,2824
25-29	0,3192
30-34	0,1997
35-39	0,0651
40-44	0,0091
45-49	0,0004
Σύνολο	1

Α. Ποια είναι η πιθανότητα μια γυναίκα που γέννησε το 1987 να είναι 24 ετών ή νεότερη;

Β. Ποια η πιθανότητα να είναι 40 ετών ή μεγαλύτερη;

Γ. Δεδομένου ότι η μητέρα ενός συγκεκριμένου βρέφους είναι κάτω των 30 ετών, ποια η πιθανότητα να μην είναι ακόμα 20 ετών;

Κατανομές

2.1 Εισαγωγή

Στο προηγούμενο κεφάλαιο αναφέραμε βασικές έννοιες των πιθανοτήτων καθώς επίσης και κάποιες μεθόδους για τον υπολογισμό της πιθανότητας ενός γεγονότος. Στο κεφάλαιο αυτό θα επεκτείνουμε τη μελέτη σε πιο πολύπλοκες περιπτώσεις. Αρχικά θα αναφέρουμε κάποιες χρήσιμες έννοιες.

Κάθε χαρακτηριστικό που μπορεί να μετρηθεί ή να ταξινομηθεί σε μια κατηγορία καλείται **μεταβλητή (variable).** Εάν μια μεταβλητή

μπορεί να πάρει κάποιον αριθμό διαφορετικών τιμών ούτως ώστε κάποιο συγκεκριμένο αποτέλεσμα να καθορίζεται τυχαία, καλείται **τυχαία μεταβλητή (random variable)**. Στην πραγματικότητα η τυχαία μεταβλητή είναι μια πραγματική συνάρτηση που ορίζεται στον δειγματικό χώρο ενός πειράματος. Οι τυχαίες μεταβλητές συμβολίζονται με κεφαλαία γράμματα και οι τιμές τους με τα αντίστοιχα μικρά. Διακρίνονται σε διακριτές και συνεχείς τυχαίες μεταβλητές (τ.μ.).

Μια τυχαία μεταβλητή (τ.μ) είναι **διακριτή** εάν παίρνει πεπερασμένο ή αριθμήσιμο πλήθος τιμών. Ενώ είναι **συνεχής** εάν μπορεί να πάρει οποιαδήποτε τιμή εντός ενός καθορισμένου διαστήματος.

Κάθε τ.μ έχει μια αντίστοιχη κατανομή πιθανότητας. Μια κατανομή πιθανότητας (probability distribution) εφαρμόζει τη θεωρία πιθανοτήτων για να περιγράψει τη συμπεριφορά μιας τ.μ. Στην περίπτωση των διακριτών μεταβλητών, προσδιορίζει όλα τα πιθανά αποτελέσματα της τυχαίας μεταβλητής καθώς επίσης και την πιθανότητα να συμβεί το καθένα από αυτά. Στην περίπτωση των συνεχών μεταβλητών, μας επιτρέπει να καθορίσουμε τις πιθανότητες που σχετίζονται με συγκεκριμένα εύρη τιμών.

Έστω, για παράδειγμα μια τ.μ Χ που αντιπροσωπεύει τη σειρά γέννησης κάθε παιδιού που γεννιέται ζωντανό από μια γυναίκα. Εάν είναι το πρώτο της τότε Χ=1, εάν είναι το δεύτερο Χ=2. Για να κατασκευάσουμε μια κατανομή πιθανότητας για την Χ, κατασκευάζουμε μια λίστα με όλες τις τιμές x της Χ μαζί με τις αντίστοιχες P(X=x) για κάθε τιμή. Έτσι , δημιουργείται ο ακόλουθος πίνακας:

x	**P(X=x)**
1	0,416
2	0,330
3	0,158
4	0,058
5	0,021
6	0,009
7	0,004
8 +	0,004
Σύνολο	1,000

Ο πίνακας αυτός μοιάζει με τις κατανομές συχνοτήτων που παρουσιάστηκαν στην Περιγραφική Στατιστική. Για ένα δείγμα παρατηρήσεων, μια κατανομή συχνότητας παρουσιάζει κάθε παρατηρούμενο αποτέλεσμα και την αντίστοιχη συχνότητά του στην ομάδα δεδομένων. Για μια διακριτή τ.μ, μια κατανομή πιθανότητας

παρουσιάζει κάθε πιθανό αποτέλεσμα και την αντίστοιχη πιθανότητά του. Οι πιθανότητες αντιπροσωπεύουν τη σχετική συχνότητα του να συμβεί κάθε αποτέλεσμα x σε ένα μεγάλο αριθμό προσπαθειών που επαναλαμβάνονται κάτω από τις ίδιες σχεδόν συνθήκες. Μας λένε ποιες τιμές είναι πιο πιθανές να συμβούν από άλλες. Υποθέτουμε ότι οι αριθμοί στους οποίους αναφερόμαστε είναι αρκετά μεγάλοι για να ικανοποιούν τον ορισμό της πιθανότητας βάσει συχνότητας. Λόγω του ότι όλες οι πιθανές τιμές της τ.μ λαμβάνονται υπ' όψη, τα αποτελέσματα εξαντλούν το δειγματικό χώρο. Άρα, το άθροισμα των πιθανοτήτων που σχετίζονται με αυτές πρέπει να είναι ίσο με 1.

Σε πολλές περιπτώσεις, μπορεί επίσης να παρουσιαστεί μια κατανομή πιθανότητας μέσω ενός γραφήματος ή μιας μαθηματικής εξίσωσης. Έτσι, για το παραπάνω παράδειγμα των γεννήσεων, έχουμε την ακόλουθη γραφική παράσταση:

Η επιφάνεια κάθε στήλης αντιπροσωπεύει την P(X=x), την πιθανότητα που σχετίζεται με αυτό το συγκεκριμένο αποτέλεσμα της τ.μ. Η συνολική επιφάνεια ισούται με 1.

Η κατανομή πιθανότητας μπορεί να χρησιμοποιηθεί για να κάνουμε δηλώσεις όσον αφορά τα πιθανά αποτελέσματα της τ.μ. Έστω ότι θέλουμε να μάθουμε ποια είναι η πιθανότητα ένα τυχαίο επιλεγμένο παιδί είναι το τρίτο της μητέρας του. Από τον πίνακα παρατηρούμε ότι P(X=3)=0,158. Αν ζητούσαμε ποια η πιθανότητα το παιδί να είναι το πρώτο ή το δεύτερο της μητέρας του, θα είχαμε: P(X=1 ή X=2)=P(X=1)+P(X=2)=0,416+0,330=0,746.

Συνήθως η P(X=x) (όταν X διακριτή) συμβολίζεται με f(x) και καλείται **συνάρτηση πιθανότητας (σ.π.)** και είναι μια πραγματική συνάρτηση με τις εξής ιδιότητες:

1. $0 \leq f(x) \leq 1$
2. $\sum_x f(x) = \sum_x P(X = x) = 1$

Πολλές φορές είναι πολύ βολικό να δουλεύει κανείς με την αθροιστική κατανομή μιας τ.μ, δηλαδή αντί για τις P(X=x) να χρησιμοποιούνται οι P(X ≤ x)=F(x) και καλείται **συνάρτηση κατανομής**. Γενικά ισχύει:

$$P(X \leq x) = \sum_{k \leq x} f(k)$$

Επίσης ισχύουν:

$$P(a \leq X \leq b) = F(b) - F(a)$$

$$P(a < X \leq b) = F(b) - F(a-1)$$

$$P(a \leq X < b) = F(b-1) - F(a)$$

$$P(a < X < b) = F(b-1) - F(a-1)$$

Στην περίπτωση συνεχούς τ.μ., η εύρεση πιθανοτήτων ανάγεται σε εύρεση πιθανοτήτων μεταξύ δυο σημείων α και b που ανάγονται σε

εμβαδά. Σε αυτήν την περίπτωση πρέπει να ευρεθεί η αναλυτική μορφή της f(x) που καλείται τώρα **συνάρτηση πυκνότητας πιθανότητας**, ισχύουν τα ίδια με τα αντίστοιχα για τη διακριτή, μόνο που τα αθροίσματα αντικαθίστανται από ολοκληρώματα. Άρα, ισχύει:

$$P(a \le X \le b) = \int_a^b f(t)dt$$

Κατά αναλογία με τις διακριτές τ.μ η συνάρτηση κατανομής μιας συνεχούς τ.μ ορίζεται σαν:

$$F(x) = P(X \le x) = \int_{-\infty}^{x} f(t)dt$$

Επίσης ισχύει:

$$P(a \le X \le b) = P(a < X \le b) = P(a \le X < b) = P(a < X < b)$$
$$= F(b) - F(a)$$

Εάν μια τυχαία μεταβλητή μπορεί να πάρει ένα μεγάλο αριθμό τιμών, τότε μια κατανομή πιθανότητας μπορεί να μην είναι χρήσιμος τρόπος για να περιγραφεί περιληπτικά η συμπεριφορά της. Όμως, όπως και με την κατανομή ομαδοποιημένων δεδομένων, δύναται να περιγραφτεί μια κατανομή πιθανότητας χρησιμοποιώντας ένα μέτρο κεντρικής τάσης και ένα μέτρο διασποράς. Η μέση τιμή μιας τ.μ. καλείται ο

μέσος του πληθυσμού (population mean) και η διασκόρπιση των τιμών γύρω από το μέσο **διασπορά του πληθυσμού (population variance)**.

Μέση τιμή

$$E(X)=\begin{cases}\sum_{x} xf(x), & \text{X διακριτη}\\ \int_{-\infty}^{+\infty} xf(x)dx, & \text{X συνεχης}\end{cases}$$

Διασπορά

$$V(X)=E(x-\mu)^2=\begin{cases}\sum (x-\mu)^2 f(x)\\ \int_{-\infty}^{+\infty} (x-\mu)^2 f(x)dx\end{cases}$$

Επίσης, ισχύει:

$$V(X)=E(X^2)-[E(X)]^2$$

Στην συνέχεια θα εξεταστούν κάποιες βασικές διακριτές και συνεχείς κατανομές.

Ιδιότητες της αναμενόμενης τιμής και διασποράς

Μέσης Τιμής

$1.E(X) = c$
$2.E(X + c) = E(X) + c$
$3.E(cX) = cE(X)$

Διασποράς

$1.V(c) = 0$
$2.V(X + c) = V(X)$
$3.V(cX) = c^2V(X)$

Παράδειγμα

Ένας πωλητής που εργάζεται για λογαριασμό ενός αμοιβαίου κεφαλαίου έχει κανονίσει τρεις συναντήσεις με πελάτες για αύριο. Από την εμπειρία του γνωρίζει ότι σε κάθε συνάντηση έχει πιθανότητα 20% να πετύχει μια πώληση. Να υπολογίσετε την κατανομή πιθανοτήτων που θα πραγματοποιήσει ο πωλητής στις τρεις συναντήσεις.

Λύση

X = αριθμό πωλήσεων που όπως διαπιστώνεται από το ακόλουθο δέντρο πιθανοτήτων παίρνει τιμές 0,1,2,3

Το δέντρο καταλήγει σε οκτώ διαφορετικούς κλάδους που αντιπροσωπεύουν οκτώ διαφορετικούς συνδυασμούς αποτελεσμάτων.

Ένας μόνο κλάδος αντιστοιχεί σε μηδέν πωλήσεις με πιθανότητα 0,512

Τρεις κλάδοι αντιστοιχούν σε μια πώληση με πιθανότητα $P(1) = 0.128 + 0.128 + 0.128 = 0.384$

Παρομοίως, τρεις κλάδοι αντιστοιχούν σε δυο πωλήσεις με πιθανότητα $P(2) = 0.032 + 0.032 + 0.032 = 0.096$

Τέλος, μόνο ένας κλάδος αντιστοιχεί σε τρεις πωλήσεις με πιθανότητα 0,008

Οπότε προκύπτει η ακόλουθη κατανομή πιθανότητας:

x	P(x)
0	0.512
1	0.384
2	0.096
3	0.008

Αν θέλαμε να συνεχίσουμε το παράδειγμα και να υπολογίσουμε τον αριθμητικό μέσο, διασπορά και τυπική απόκλιση.

$$E(X) = \mu = \sum xP(x) = 0 \cdot 0.512 + 1 \cdot 0.384 + 2 \cdot 0.096 + 3 \cdot 0.008 = 0.6$$

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

$$\sum x^2 P(x) = 0 \cdot 0.512^2 + 1 \cdot 0.384^2 + 2 \cdot 0.096^2 + 3 \cdot 0.008^2 - 0.6^2 = 0.067$$

$$\sigma = \sqrt{V(X)} = 0.259$$

Περίπτωση δυο τυχαίων μεταβλητών

Όσα έχουν αναφερθεί στην περίπτωση μιας τυχαίας μεταβλητής μπορούν να επεκταθούν και στην περίπτωση δυο τυχαίων μεταβλητών.

Διακριτές Μεταβλητές

Ορισμός

Αν X και Y **δυο διακριτές τυχαίες μεταβλητές, τότε ορίζεται μια συνάρτηση** $f(x, y)$ **με πεδίο τιμών το** R^2

και πεδίο τιμών ένα υποσύνολο του διαστήματος $[0,1]$**,**

που ορίζεται από τη σχέση

$$f(x,y) = P(X = x, Y = y), \forall x, y \in R^2$$

και καλείται **κοινή συνάρτηση πιθανότητας** των μεταβλητών X και Y.

Η από κοινού συνάρτηση πιθανότητας ικανοποιεί τις εξής ιδιότητες:

$$1. f(x,y) \geq 0, \forall x, y \in R^2$$
$$2. \sum_x \sum_y f(x,y) = 1$$

Ορισμός

Οι σ.π. των X και Y που ορίζονται ξεχωριστά από τις παρακάτω σχέσεις

$$f_x(x) = \sum_y f(x,y), x \in R$$
$$f_y(x) = \sum_x f(x,y), y \in R$$

καλούνται **περιθώριες συναρτήσεις πιθανότητας** της X και Y αντίστοιχα.

Ορισμός

Η **δεσμευμένη σ.π**. της X όταν $Y = y$ συμβολίζεται με $f_{X|Y=y}(x)$ και δίνεται από την ακόλουθη σχέση:

$$f_{X|Y=y}(x) = \frac{f(x,y)}{f_y(y)}, x \in R$$

Αντίστοιχα,

Η **δεσμευμένη σ.π**. της Y όταν $X = x$ συμβολίζεται με $f_{Y|X=x}(y)$ και δίνεται από την ακόλουθη σχέση:

$$f_{Y|X=x}(y) = \frac{f(x,y)}{f_x(x)}, y \in R$$

Ορισμός

Αν X, Y έχουν $f(x,y)$ από κοινού σ.π. και $g(x,y)$ μια συνάρτηση των μεταβλητών αυτών, τότε ορίζεται η **μαθηματική ελπίδα** ως εξής:

$$E[g(x,y)] = \sum_y \sum_x g(x,y) f(x,y)$$

Συμμεταβλητότητα ή Συνδιασπορά (Covariance)

Για τη σχέση μεταξύ δυο μεταβλητών μπορούμε να υπολογίσουμε δυο παραμέτρους: **συμμεταβλητότητα και ο συντελεστής συσχέτισης (coefficient of correlation).**

Συνεχείς Μεταβλητές

Ορισμός

Αν X και Y **δυο συνεχείς τυχαίες μεταβλητές, τότε ορίζεται μια συνάρτηση** $f(x,y)$ **με πεδίο τιμών το** R^2 **και πεδίο τιμών ένα υποσύνολο του διαστήματος** $[0,1]$**, που ορίζεται από τη σχέση**

$$f(x,y) = P(X = x, Y = y), \forall x, y \in R^2$$

και καλείται **κοινή συνάρτηση πιθανότητας** των μεταβλητών X και Y.

Η από κοινού συνάρτηση πιθανότητας ικανοποιεί τις εξής ιδιότητες:

$$1. f(x,y) \geq 0, \forall x, y \in R^2$$

$$2. \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f(x,y) = 1$$

Ορισμός

Οι σ.π. των X και Y που ορίζονται ξεχωριστά από τις παρακάτω σχέσεις

$$f_x(x) = \int_{-\infty}^{+\infty} f(x,y)dy, x \in R$$

$$f_y(x) = \int_{-\infty}^{+\infty} f(x,y)dx, y \in R$$

καλούνται **περιθώριες συναρτήσεις πιθανότητας** της X και Y αντίστοιχα.

Ορισμός

Η **δεσμευμένη σ.π**. της X όταν $Y = y$ συμβολίζεται με $f_{X|Y=y}(x)$ και δίνεται από την ακόλουθη σχέση:

$$f_{X|Y=y}(x) = \frac{f(x,y)}{f_y(y)}, x \in R$$

Αντίστοιχα,

Η **δεσμευμένη σ.π**. της Y όταν $X = x$ συμβολίζεται με $f_{Y|X=x}(y)$ και δίνεται από την ακόλουθη σχέση:

$$f_{Y|X=x}(y) = \frac{f(x,y)}{f_x(x)}, y \in R$$

Ορισμός

Αν X, Y έχουν $f(x,y)$ από κοινού σ.π. και $g(x,y)$ μια συνάρτηση των μεταβλητών αυτών, τότε ορίζεται η **μαθηματική ελπίδα** ως εξής:

$$E[g(x,y)] = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} g(x,y) f(x,y) dxdy$$

Ανεξαρτησία τυχαίων μεταβλητών

Δυο τ.μ. (διακριτές ή συνεχείς καλούνται **ανεξάρτητες** αν ισχύει

$$f(x,y) = f_x(x) f_y(y)$$

Επίσης,

$$f_{X|Y=y} = f_x(x)$$
$$f_{Y|X=x} = f_y(y)$$

Επίσης,

$$E(XY) = E(X)E(Y)$$

2.2 Διακριτές Κατανομές

Πείραμα Bernoulli

Πριν αναφέρουμε τις διακριτές κατανομές θα περιγράψουμε ένα βασικό πείραμα, το οποίο καλείται πείραμα Bernoulli και από το οποίο προέρχεται μια βασική διακριτή κατανομή. Ένα πείραμα Bernoulli είναι ένα τυχαίο πείραμα με δυο μόνο δυνατά αποτελέσματα. Για παράδειγμα, άρρωστος-υγιής, αγόρι-κορίτσι, γράμματα-κορώνα, αριθμός ελαττωματικών προϊόντων κ.ο.κ.

Σε κάθε τέτοιο πείραμα, έχουμε μόνο δυο δυνατά αποτελέσματα που συμβολίζονται τυχαία με 1 και 0, και ονομάζονται **επιτυχία** και **αποτυχία** αντίστοιχα.

Η πιθανότητα να έχουμε επιτυχία είναι p και 1-p να έχουμε αποτυχία. Η πιθανότητα αυτή παραμένει σταθερή όσες φορές και αν

επαναληφθεί το πείραμα. Και φυσικά, είναι εμφανές ότι ισχύει $p+q=1$

Κατανομή Bernoulli

Έστω X ο αριθμός των επιτυχιών σε ένα πείραμα Bernoulli, με πιθανότητα επιτυχίας p και πιθανότητα αποτυχίας q. Η κατανομή της τ.μ. X (δίτιμη τ.μ., παίρνει δυο τιμές, 1 επιτυχία, 0 αποτυχία) καλείται **Κατανομή Bernoulli** με παράμετρο p.

Διωνυμική Κατανομή

Έστω ότι ένα πείραμα Bernoulli το επαναλαμβάνουμε n φορές (όπου n είναι προκαθορισμένος αριθμός). Επίσης, θεωρούμε τα εξής:

1. Οι δοκιμές είναι ίδιες ακριβώς και κάθε δοκιμή μπορεί να έχει δυο δυνατά αποτελέσματα: επιτυχία ή αποτυχία
2. Οι δοκιμές είναι ανεξάρτητες, έτσι ώστε το αποτέλεσμα μιας ορισμένης δοκιμής δεν επηρεάζει το αποτέλεσμα οποιασδήποτε άλλης δοκιμής
3. Η πιθανότητα επιτυχίας είναι σταθερή από δοκιμή σε δοκιμή και ίση με p, η δε πιθανότητα αποτυχίας είναι ίση με $q=1-p$

Το πείραμα αυτό καλείται διωνυμικό. Και έστω ότι καταγράφουμε των αριθμών των επιτυχιών στις n επαναλήψεις (δοκιμές) του πειράματος.

Και έστω X ο αριθμός των επιτυχιών σε n δοκιμές Bernoulli με σταθερή πιθανότητα επιτυχίας p και πιθανότητα αποτυχίας q, σε όλες τις επαναλήψεις. Τότε λέμε ότι η X ακολουθεί τη *Διωνυμική Κατανομή (Binomial Distribution) με παραμέτρους* n *και* p.

Συμβολισμός: $X \sim B(n, p)$

Συνάρτηση πιθανότητας:

$$f(x) = \binom{n}{x} p^x q^{n-x}, x = 0,1,2,...,n \ , \ 0 \le p \le 1$$

$E(X) = np$ Μέση Τιμή της τ.μ.
$V(X) = npq$ Διασπορά της τ.μ.

Γεωμετρική Κατανομή

Έστω Χ ο αριθμός των δοκιμών μέχρι την εμφάνιση της πρώτης επιτυχίας, σε n δοκιμές Bernoulli με πιθανότητα επιτυχίας p. Η κατανομή της τ.μ. Χ καλείται *Γεωμετρική Κατανομή με παράμετρο* p.

Συμβολισμός: $X \sim G(p)$

$$f(x) = P(X = x) = q^{x-1} p, \quad x = 1, 2, \ldots$$
$$E(X) = 1 / p$$
$$V(X) = q / p^2$$

Αρνητική Διωνυμική Κατανομή

Έστω ότι έχουμε n δοκιμές Bernoulli με πιθανότητα επιτυχίας p και έστω Χ ο αριθμός των δοκιμών μέχρι την εμφάνιση της r επιτυχίας. Η κατανομή της τ.μ. Χ καλείται *Αρνητική Διωνυμική Κατανομή με παραμέτρους r και p.*

Συμβολισμός: $X \sim NB(r, p)$

$$f(X) = P(X = x) = \binom{x-1}{r-1} p^r q^{x-r} \quad , x = r, r+1, \ldots.$$
$$E(X) = r / p$$
$$V(X) = rq / p^2$$

Υπεργεωμετρική Κατανομή

Θεωρούμε ότι έχουμε μια κάλπη που περιέχει α λευκά σφαιρίδια και β μαύρα σφαιρίδια και εξάγουμε διαδοχικά, το ένα μετά το άλλο, n σφαιρίδια, χωρίς επανατοποθέτηση. Έστω Χ ο αριθμός των λευκών

σφαιριδίων που περιέχονται στο δείγμα. Η κατανομή της τ.μ. Χ καλείται *υπεργεωμετρική κατανομή με παραμέτρους n, α, β.*

Συμβολισμός: $X \sim H(n,\alpha,\beta)$

$$f(x) = P(X = x) = \frac{\binom{a}{x}\binom{\beta}{n-x}}{\binom{a+\beta}{n}}, x = \max(0, n-\beta), ..., \min(n, a)$$

$$E(X) = n\frac{a}{a+\beta}$$

$$V(X) = n\frac{a}{a+\beta}\frac{\beta}{a+\beta}(1-\frac{n-1}{a+\beta-1})$$

Η κατανομή αυτή χρησιμοποιείται όταν τα άτομα ενός πεπερασμένου πληθυσμού ταξινομούνται σε δυο κατηγορίες, ανάλογα με τις τιμές κάποιου χαρακτηριστικού που επιθυμούμε να εξετάσουμε. Επιλέγεται τυχαία ένα δείγμα από αυτόν τον πληθυσμό μεγέθους n και θέλουμε να μελετήσουμε το πλήθος των ατόμων (μονάδων) που ανήκουν σε κάποια από τις δυο κατηγορίες.

Παραδείγματα

- Υπέρβαρο – μη υπέρβαρο
- Καπνιστές – μη καπνιστές

- Ελαττωματικό – μη ελαττωματικό αντικείμενο μιας γραμμής παραγωγής

Κ.ο.κ

Poisson Κατανομή

Η κατανομή αυτή χρησιμοποιείται για εύρεση πιθανοτήτων κάποιων γεγονότων που πραγματοποιούνται σε δοσμένα χρονικά διαστήματα. Δηλαδή, η κατανομή αυτή βρίσκει την πιθανότητα επιτυχίας σε n επαναλήψεις ανά μονάδα χρόνου.

Έστω Χ ο αριθμός των γεγονότων που πραγματοποιούνται σε ορισμένο χρονικό διάστημα. λ ο ρυθμός πραγματοποίησης των γεγονότων, $\lambda > 0$. Η κατανομή της τ.μ Χ καλείται *Poisson κατανομή με παράμετρο λ.*

Συμβολισμός: $X \sim P(\lambda)$

$$f(x) = P(X = x) = \frac{e^{-\lambda}\lambda^{x}}{x!}, x = 0,1,2,... \text{ και } \lambda > 0$$

$$E(X) = \lambda$$
$$V(X) = \lambda$$

Παρατήρηση

- e ο νεπέριος αριθμός και ισούται 2,71828

Μοντέλα που ακολουθούν την Poisson Κατανομή

- Αριθμός τηλεφωνικών κλήσεων που καταφτάνουν σε ένα τηλεφωνικό κέντρο σε δοσμένο χρονικό διάστημα
- Αριθμός ατυχημάτων που συμβαίνουν σε μια διασταύρωση σε δοσμένο χρονικό διάστημα
- Αριθμός επειγόντων περιστατικών που καταφτάνουν σε μια νύχτα σε ένα νοσοκομείο
- Αριθμός πελατών που καταφτάνουν σε ένα συγκεκριμένο κατάστημα σε δοσμένο χρονικό διάστημα
- Αριθμός ελαττωματικών προϊόντων που παράγονται σε ένα δοσμένο χρονικό διάστημα και από συγκεκριμένη γραμμή παραγωγής

Κ.ο.κ.

Συνεχείς Κατανομές

Όπως ήδη έχει αναφερθεί μια συνεχής τ.μ. είναι αυτή που μπορεί να πάρει οποιαδήποτε τιμή x σε ένα διάστημα πραγματικών αριθμών. Έτσι μια συνεχή τ.μ. δεν μπορούμε να εκτιμήσουμε την πιθανότητα να

πάρουμε μια συγκεκριμένη τιμή, αλλά έχουμε διαστήματα τιμών. Τέτοιες μεταβλητές είναι το βάρος, το ύψος, ο χρόνος, το μήκος, κ.ο.κ.

Ομοιόμορφη Κατανομή

Έστω μια συνεχής τ.μ. X με σ.π.π

$$f(x)=\begin{cases}\dfrac{1}{b-a}, a\le x\le b\\ 0, \quad \text{αλλού}\end{cases}$$

Η κατανομή της τ.μ. ονομάζεται *ομοιόμορφη κατανομή στο* $[a,b]$.

Συμβολισμός: $X\sim U(a,b)$

$$E(X)=\frac{a+b}{2}$$

$$V(X)\frac{(b-a)^2}{12}$$

Κανονική Κατανομή

Είναι η πιο σημαντική κατανομή της θεωρίας των Πιθανοτήτων και της Στατιστικής. Χρησιμοποιείται σε πάρα πολλές εφαρμογές.

Πολλά χαρακτηριστικά όπως το βάρος, το ύψος, βαθμολογία σε εξετάσεις, κ.ο.κ. περιγράφονται από την κανονική κατανομή. Επίσης, σε διάφορες μετρήσεις εμφανίζονται τυχαία σφάλματα που προσεγγιστικά ακολουθούν την κανονική κατανομή.

Είναι μια συμμετρική κατανομή σε μορφή καμπάνας, που καλείται καμπύλη κανονικής κατανομής ή Καμπύλη του Gauss.

Η σ.π.π της τ.μ. $X \sim N(\mu, \sigma^2)$ είναι:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Όπου $f(x)$ το ύψος της κατανομής, μ ο μέσος, σ^2 η διασπορά, σ η τυπική απόκλιση, $e = 2,7183$ και $\pi = 3,1416$

$$E(X) = \mu$$
$$V(X) = \sigma^2$$

Ιδιότητες της κανονικής κατανομής

- Είναι συμμετρική γύρω από το μ και ισχύει φυσικά $\mu = \mathrm{M} = \mathrm{M}_o$. Η μέση τιμή είναι το πιο ψηλό σημείο της καμπύλης

- Το εμβαδόν που περικλείεται μεταξύ της καμπύλης και του οριζόντιου άξονα ισούται με 1
- Τα σημεία $\mu \pm \sigma$ αποτελούν σημεία καμπής
- Η απόσταση των σημείων καμπής από την κάθετο στο μ, ισούται με μια τυπική απόκλιση σ
- Μια κανονική κατανομή προσδιορίζεται πλήρως από τις παραμέτρους μ και σ. Όσο μεγαλύτερη είναι η μέση τιμή τόσο δεξιότερα προς τον οριζόντια άξονα βρίσκεται η καμπύλη. Ενώ όσο μεγαλύτερη είναι η τυπική απόκλιση τόσο πιο «απλωμένη» είναι η καμπύλη
- $$X \sim N(\mu, \sigma^2) \Rightarrow$$
$$P(\mu - \sigma \le X \le \mu + \sigma) = 0,6827 (\square\ 68\%)$$
$$P(\mu - 2\sigma \le X \le \mu + 2\sigma) = 0,9545 (\square\ 95\%)$$
$$P(\mu - 3\sigma \le X \le \mu + 3\sigma) = 0,9973 (\square\ 99\%)$$

Παρατήρηση

Συνήθως τα προβλήματα που αφορούν την κανονική κατανομή είναι προβλήματα εύρεσης της πιθανότητας η τ.μ. X να βρίσκεται μεταξύ δυο αριθμών έστω x_1, x_2 ή κάτω από τον x_1 ή πάνω από τον x_2. Δηλαδή, η

εύρεση αυτών των πιθανοτήτων ανάγεται σε υπολογισμό ενός εμβαδού που είναι φυσικά μέρος του συνολικού εμβαδού. Η ολοκλήρωση όμως της σ.π.π είναι αρκετά κοπιαστική, με αποτέλεσμα να δημιουργούνται αρκετά προβλήματα στον υπολογισμό αυτών των πιθανοτήτων. Γι' αυτό η μόνη λύση είναι να τυποποιήσουμε τα δεδομένα. Κάνοντας αυτόν τον μετασχηματισμό παίρνουμε την τυπική κανονική κατανομή.

Τυπική Κανονική Κατανομή

Η τυπική κατανομή έχει $\mu = 0$ και $\sigma^2 = 1$. Ο μετασχηματισμός γίνεται με χρήση του τύπου:

$$Z = \frac{X - \mu}{\sigma}$$

$$Z \sim N(0,1)$$

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}$$

$$E(z) = 0$$

$$V(z) = 1$$

Η Z είναι ανεξάρτητη από μονάδες μέτρησης της τ.μ., αφού δεν εκφράζεται σε καμιά μονάδα μέτρησης.

Με τον μετασχηματισμό αυτόν καταφέραμε να μετατρέψουμε τα δεδομένα μιας μεταβλητής που κατανέμεται κανονικά σε τυποποιημένη μορφή και να υπολογίσουμε τις πιθανότητες με χρήση των πινάκων της τυποποιημένης κανονικής κατανομής.

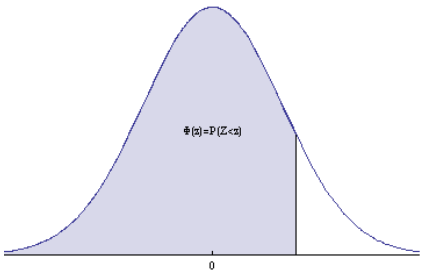
$\Phi(\alpha)$ είναι η συνάρτηση αθροιστικής κατανομής της Z δηλαδή $\Phi(a) = P(Z < a)$ με τις παρακάτω ιδιότητες:

$\Phi(0) = 0.5$
$\Phi(-\alpha) = 1 - \Phi(\alpha)$
$P(Z > b) = 1 - \Phi(b)$
$P(a < Z < b) = \Phi(b) - \Phi(a)$

Επίσης

$P(a < X < b) = \Phi(\frac{b-\mu}{\sigma}) - \Phi(\frac{a-\mu}{\sigma})$ όπου $X \sim N(\mu, \sigma^2)$
$P(X \le b) = \Phi(\frac{b-\mu}{\sigma})$
$P(X > a) = 1 - \Phi(\frac{a-\mu}{\sigma})$

Τιμές των πιθανοτήτων $\Phi(z) = P(Z \le z) = P(Z < z)$ της τυποποιημένης κανονικής κατανομής $N(0,1)$ για $z \ge 0$. Για $z < 0$ ισχύει $\Phi(z) = 1 - \Phi(-z)$.



z	**0.00**	**0.01**	**0.02**	**0.03**	**0.04**	**0.05**	**0.06**	**0.07**	**0.08**	**0.09**
0.0	0.50000	0.50399	0.50798	0.51197	0.51595	0.51994	0.52392	0.52790	0.53188	0.53586
0.1	0.53983	0.54380	0.54776	0.55172	0.55567	0.55962	0.56356	0.56749	0.57142	0.57535
0.2	0.57926	0.58317	0.58706	0.59095	0.59483	0.59871	0.60257	0.60642	0.61026	0.61409
0.3	0.61791	0.62172	0.62552	0.62930	0.63307	0.63683	0.64058	0.64431	0.64803	0.65173
0.4	0.65542	0.65910	0.66276	0.66640	0.67003	0.67364	0.67724	0.68082	0.68439	0.68793
0.5	0.69146	0.69497	0.69847	0.70194	0.70540	0.70884	0.71226	0.71566	0.71904	0.72240
0.6	0.72575	0.72907	0.73237	0.73565	0.73891	0.74215	0.74537	0.74857	0.75175	0.75490
0.7	0.75804	0.76115	0.76424	0.76730	0.77035	0.77337	0.77637	0.77935	0.78230	0.78524
0.8	0.78814	0.79103	0.79389	0.79673	0.79955	0.80234	0.80511	0.80785	0.81057	0.81327
0.9	0.81594	0.81859	0.82121	0.82381	0.82639	0.82894	0.83147	0.83398	0.83646	0.83891
1.0	0.84134	0.84375	0.84614	0.84850	0.85083	0.85314	0.85543	0.85769	0.85993	0.86214
1.1	0.86433	0.86650	0.86864	0.87076	0.87286	0.87493	0.87698	0.87900	0.88100	0.88298
1.2	0.88493	0.88686	0.88877	0.89065	0.89251	0.89435	0.89617	0.89796	0.89973	0.90147
1.3	0.90320	0.90490	0.90658	0.90824	0.90988	0.91149	0.91309	0.91466	0.91621	0.91774
1.4	0.91924	0.92073	0.92220	0.92364	0.92507	0.92647	0.92786	0.92922	0.93056	0.93189
1.5	0.93319	0.93448	0.93574	0.93699	0.93822	0.93943	0.94062	0.94179	0.94295	0.94408
1.6	0.94520	0.94630	0.94738	0.94845	0.94950	0.95053	0.95154	0.95254	0.95352	0.95449
1.7	0.95543	0.95637	0.95728	0.95818	0.95907	0.95994	0.96080	0.96164	0.96246	0.96327
1.8	0.96407	0.96485	0.96562	0.96638	0.96712	0.96784	0.96856	0.96926	0.96995	0.97062
1.9	0.97128	0.97193	0.97257	0.97320	0.97381	0.97441	0.97500	0.97558	0.97615	0.97670
2.0	0.97725	0.97778	0.97831	0.97882	0.97932	0.97982	0.98030	0.98077	0.98124	0.98169

2.1	0.98214	0.98257	0.98300	0.98341	0.98382	0.98422	0.98461	0.98500	0.98537	0.98574
2.2	0.98610	0.98645	0.98679	0.98713	0.98745	0.98778	0.98809	0.98840	0.98870	0.98899
2.3	0.98928	0.98956	0.98983	0.99010	0.99036	0.99061	0.99086	0.99111	0.99134	0.99158
2.4	0.99180	0.99202	0.99224	0.99245	0.99266	0.99286	0.99305	0.99324	0.99343	0.99361
2.5	0.99379	0.99396	0.99413	0.99430	0.99446	0.99461	0.99477	0.99492	0.99506	0.99520
2.6	0.99534	0.99547	0.99560	0.99573	0.99585	0.99598	0.99609	0.99621	0.99632	0.99643
2.7	0.99653	0.99664	0.99674	0.99683	0.99693	0.99702	0.99711	0.99720	0.99728	0.99736
2.8	0.99744	0.99752	0.99760	0.99767	0.99774	0.99781	0.99788	0.99795	0.99801	0.99807
2.9	0.99813	0.99819	0.99825	0.99831	0.99836	0.99841	0.99846	0.99851	0.99856	0.99861
3.0	0.99865	0.99869	0.99874	0.99878	0.99882	0.99886	0.99889	0.99893	0.99897	0.99900

ΣΥΝΟΠΤΙΚΕΣ ΣΗΜΕΙΩΣΕΙΣ ΣΤΑΤΙΣΤΙΚΗΣ

ΠΕΡΙΠΤΩΣΗ ΔΥΟ ΤΥΧΑΙΩΝ ΜΕΤΑΒΛΗΤΩΝ

Περίπτωση δυο τυχαίων μεταβλητών

Όσα έχουν αναφερθεί στην περίπτωση μιας τυχαίας μεταβλητής μπορούν να επεκταθούν και στην περίπτωση δυο τυχαίων μεταβλητών.

Διακριτές Μεταβλητές

Ορισμός

Αν X και Y **δυο διακριτές τυχαίες μεταβλητές, τότε ορίζεται μια συνάρτηση** $f(x,y)$ **με πεδίο τιμών το** R^2 **και πεδίο τιμών ένα υποσύνολο του διαστήματος** $[0,1]$**, που ορίζεται από τη σχέση**

$$f(x,y) = P(X = x, Y = y), \forall x, y \in R^2$$

και καλείται **κοινή συνάρτηση πιθανότητας** των μεταβλητών X και Y.

Η από κοινού συνάρτηση πιθανότητας ικανοποιεί τις εξής ιδιότητες:

$$1. f(x,y) \geq 0, \forall x, y \in R^2$$
$$2. \sum_x \sum_y f(x,y) = 1$$

Ορισμός

Οι σ.π. των X και Y που ορίζονται ξεχωριστά από τις παρακάτω σχέσεις

$$f_x(x) = \sum_y f(x,y), x \in R$$
$$f_y(x) = \sum_x f(x,y), y \in R$$

καλούνται **περιθώριες συναρτήσεις πιθανότητας** της X και Y αντίστοιχα.

Ορισμός

Η **δεσμευμένη σ.π**. της X όταν $Y = y$ συμβολίζεται με $f_{X|Y=y}(x)$ και δίνεται από την ακόλουθη σχέση:

$$f_{X|Y=y}(x) = \frac{f(x,y)}{f_y(y)}, x \in R$$

Αντίστοιχα,

Η **δεσμευμένη σ.π**. της Y όταν $X = x$ συμβολίζεται με $f_{Y|X=x}(y)$ και δίνεται από την ακόλουθη σχέση:

$$f_{Y|X=x}(y) = \frac{f(x,y)}{f_x(x)}, y \in R$$

Ορισμός

Αν X, Y έχουν $f(x,y)$ από κοινού σ.π. και $g(x,y)$ μια συνάρτηση των μεταβλητών αυτών, τότε ορίζεται η **μαθηματική ελπίδα** ως εξής:

$$E[g(x,y)] = \sum_y \sum_x g(x,y) f(x,y)$$

Συνεχείς Μεταβλητές

Ορισμός

Αν X και Y **δυο συνεχείς τυχαίες μεταβλητές, τότε ορίζεται μια συνάρτηση** $f(x,y)$ **με πεδίο τιμών το** R^2

και πεδίο τιμών ένα υποσύνολο του διαστήματος $[0,1]$**,**

που ορίζεται από τη σχέση

$$f(x,y) = P(X = x, Y = y), \forall x, y \in R^2$$

και καλείται **κοινή συνάρτηση πιθανότητας** των μεταβλητών X και Y.

Η από κοινού συνάρτηση πιθανότητας ικανοποιεί τις εξής ιδιότητες:

$$1. f(x,y) \geq 0, \forall x, y \in R^2$$

$$2. \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f(x,y) = 1$$

Ορισμός

Οι σ.π. των X και Y που ορίζονται ξεχωριστά από τις παρακάτω σχέσεις

$$f_x(x) = \int_{-\infty}^{+\infty} f(x,y)dy, x \in R$$

$$f_y(x) = \int_{-\infty}^{+\infty} f(x,y)dx, y \in R$$

καλούνται **περιθώριες συναρτήσεις πιθανότητας** της X και Y αντίστοιχα.

Ορισμός

Η **δεσμευμένη σ.π**. της X όταν $Y = y$ συμβολίζεται με $f_{X|Y=y}(x)$ και δίνεται από την ακόλουθη σχέση:

$$f_{X|Y=y}(x) = \frac{f(x,y)}{f_y(y)}, x \in R$$

Αντίστοιχα,

Η **δεσμευμένη σ.π**. της Y όταν $X = x$ συμβολίζεται με $f_{Y|X=x}(y)$ και δίνεται από την ακόλουθη σχέση:

$$f_{Y|X=x}(y) = \frac{f(x,y)}{f_x(x)}, y \in R$$

Ορισμός

Αν X, Y έχουν $f(x,y)$ από κοινού σ.π. και $g(x,y)$ μια συνάρτηση των μεταβλητών αυτών, τότε ορίζεται η **μαθηματική ελπίδα** ως εξής:

$$E[g(x,y)] = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} g(x,y) f(x,y) dxdy$$

Ανεξαρτησία τυχαίων μεταβλητών

Δυο τ.μ. (διακριτές ή συνεχείς καλούνται **ανεξάρτητες** αν ισχύει

$$f(x,y) = f_x(x) f_y(y)$$

Επίσης,

$$f_{X|Y=y} = f_x(x)$$
$$f_{Y|X=x} = f_y(y)$$

Επίσης,

$$E(XY) = E(X)E(Y)$$

για

Στατιστική Συμπερασματολογία
Διαστήματα Εμπιστοσύνης-Έλεγχοι Υποθέσεων

Γενικά

Στα προηγούμενα κεφάλαια ασχοληθήκαμε με την παρουσίαση χαρακτηριστικών ή μετρήσεων που αναφέρονται στον πληθυσμό ή σε κάποιο τυχαίο δείγμα. Εδώ θα ασχοληθούμε με τη μελέτη τυχαίων δειγμάτων των οποίων τα συμπεράσματα επεκτείνονται σε ολόκληρο τον πληθυσμό. Όπως ήδη είναι γνωστό ο κλάδος της Στατιστικής που ασχολείται με την γενίκευση των συμπερασμάτων του δείγματος ονομάζεται **στατιστική συμπερασματολογία.**

Για παράδειγμα, η παρακολούθηση του ύψους 1000 βρεφών της χώρας, μπορεί να δώσει κάποια συμπεράσματα για την κατανομή του ύψους όλων των βρεφών της χώρας.

Επίσης, εάν μια εταιρεία ισχυρίζεται ότι κάποιο προϊόν είναι καλύτερο από κάποιο άλλο. Ο ισχυρισμός αυτός δύναται να ελεγχθεί με βάσει κάποιο τυχαίο δείγμα.

Εκτιμητική

Είναι ο κλάδος της Στατιστικής Συμπερασματολογίας που ασχολείται με μεθόδους εκτίμησης της τιμής αγνώστων παραμέτρων κάποιας κατανομής. Η εκτίμηση γίνεται με το προσδιορισμό στατιστικών συναρτήσεων, των εκτιμητών, η τιμή των οποίων «τείνει να είναι κοντά» στη τιμή των αγνώστων παραμέτρων.

Για παράδειγμα, ο δειγματικός μέσος χρησιμοποιείται ως αμερόληπτη εκτιμήτρια του μ. Δηλαδή, αν πάρουμε όλα τα δυνατά τυχαία δείγματα από τον πληθυσμό και υπολογίσουμε τους μέσους, τότε ο μέσος όρος όλων αυτών θα είναι ο μ.

Παρόμοια, για τη δειγματική διασπορά. Δεν συμβαίνει το ίδιο και για την τυπική απόκλιση. Ο s δεν είναι αμερόληπτη εκτιμήτρια του σ.

Παράδειγμα

Έστω ένα δείγμα 20 ατόμων με τα παρακάτω ύψη σε cm:

173 166 168 166 169 166 173 170 173 166 161 166 170 168 158 173 166 165 165

εύκολα βρίσκουμε ότι η μέση τιμή είναι 167.35 cm και η τυπική απόκλιση 4.23 cm. Αν τώρα θέλουμε να εκτιμήσουμε τον πραγματικό μέσο του πληθυσμού στηριζόμενοι στο δείγμα, μπορούμε να απαντήσουμε με διάφορους τρόπους μερικοί από τους οποίους είναι:

α. μια εκτίμηση για τη μέση τιμή του πληθυσμού είναι 167.35 cm

β. μια εκτίμηση για τη μέση τιμή του πληθυσμού είναι 167.35 cm και το δείγμα έχει μέγεθος 20

γ. μια εκτίμηση για τη μέση τιμή του πληθυσμού είναι 167.35 cm, τυπική απόκλιση ίση με 4.23 και μέγεθος δείγματος 20

Και οι τρεις απαντήσεις, είναι απαντήσεις βασιζόμενες στην εκτίμηση σημείου, άσχετα αν η καθεμία δίνει διαφορετικό πλήθος πληροφοριών.

Παρατηρείται ότι η σημειακή εκτίμηση δεν δίνει απάντηση της μορφής:΄η μέση τιμή είναι ίση με’, αλλά περίπου ίση με τη μέση τιμή του δείγματος. Αυτή η παρατήρηση αποτελεί και το μειονέκτημα της σημειακής εκτίμησης.

Διαστήματα Εμπιστοσύνης

Είναι ο κλάδος της Στατιστικής Συμπερασματολογίας που ασχολείται με τον προσδιορισμό διαστημάτων, περιοχών γενικότερα, που περιέχουν με μεγάλη πιθανότητα άγνωστες παραμέτρους κάποιας κατανομής.

Με εκτίμηση σε διάστημα, είμαστε σίγουροι ότι το διάστημα που υπολογίζετε περιέχει την πραγματική τιμή της εξεταζόμενης παραμέτρου. Επειδή το προτεινόμενο διάστημα συνοδεύεται με ένα **συντελεστή εμπιστοσύνης** (δηλ. με μια πιθανότητα), το διάστημα αυτό καλείται Δ**ιάστημα εμπιστοσύνης (Δ.Ε.) (Confidence Interval)** και πιο συγκεκριμένα **Δ.Ε. με βαθμό εμπιστοσύνης γ (β.ε).** Ο αριθμός γ=1-α εκφράζει την ακρίβεια με την οποία θέλουμε να γίνει η εκτίμηση, ενώ ο α (επίπεδο σημαντικότητας) εκφράζει τον βαθμό ανεκτικότητας ώστε το διάστημα να μην περιέχει την πραγματική τιμή της παραμέτρου. Για παράδειγμα, αν γ=0.95, αυτό ερμηνεύεται ως εξής: στα 100 δείγματα στα 95 το ΔΕ θα περιέχει την πραγματική τιμή.

Το ΔΕ έχει μεγαλύτερη έκταση όσο ο συντελεστής εμπιστοσύνης είναι μεγαλύτερος. Δηλαδή, ένα 99% ΔΕ έχει μεγαλύτερη έκταση από ένα 95% ΔΕ, ώστε να είμαστε σίγουροι ότι στο διάστημα αυτό βρίσκεται η παράμετρος που εκτιμάμε.

Έλεγχοι Στατιστικών Υποθέσεων

Είναι ο κλάδος της Στατιστικής Συμπερασματολογίας, που ασχολείται με τον έλεγχο ισχυρισμών, υποθέσεων, για την τιμή των αγνώστων παραμέτρων κάποιας κατανομής.

Μια στατιστική υπόθεση είναι συνήθως μια υπόθεση για έναν πληθυσμό. Για παράδειγμα, ο μέσος αριθμός βακτηρίων που σκοτώνονται από ένα συγκεκριμένο φάρμακο είναι ίσος με έναν αριθμό, ο μέσος αριθμός πωλήσεων ενός προϊόντος είναι ίσος με έναν αριθμό, το ποσοστό των ελαττωματικών προϊόντων που κατασκευάζει ένα εργοστάσιο είναι ίσο με έναν αριθμό, κ.ο.κ.

Η διαδικασία που ακολουθείται για την επαλήθευση ή όχι των παραπάνω ισχυρισμών, ονομάζεται **έλεγχος υποθέσεως** ή **στατιστικός έλεγχος**. Σύμφωνα με τη διαδικασία αυτή υπάρχει πάντα μια τιμή για την παράμετρο που μας ενδιαφέρει και εξετάζεται το ερώτημα: τι σχέση έχει η παράμετρος αυτή με τη συγκεκριμένη τιμή;

Ο έλεγχος που εφαρμόζεται στη στατιστική για την παραπάνω εξέταση ακολουθεί τα εξής βήματα:

1. Ορίζουμε τη **μηδενική** ή **αρχική υπόθεση H_0(null hypotheses).** Μηδενική υπόθεση καλείται η υπόθεση που κάνουμε για μια συγκεκριμένη τιμή της παραμέτρου και η οποία είναι η σοβαρότερη υπόθεση στον έλεγχο.
2. Ορίζουμε την **εναλλακτική υπόθεση H_1(alternative hypotheses).** Δηλαδή, την υπόθεση για την οποία ελέγχεται η μηδενική υπόθεση
3. Ορίζουμε τον έλεγχο που εφαρμόζεται για την αποδοχή ή την απόρριψη της αρχικής υπόθεσης.
4. Ορίζουμε την απορριπτική περιοχή της μηδενική υπόθεσης ή αλλιώς την κρίσιμη περιοχή του ελέγχου. Δηλαδή την περιοχή του δειγματοχώρου για την οποία απορρίπτεται η αρχική υπόθεση.
5. Βγάζουμε συμπεράσματα.

Η μεθοδολογία λήψης της απόφασης καλείται έλεγχος της μηδενικής υπόθεσης κατά την εναλλακτική υπόθεση.

Σε κάθε έλεγχο είναι δυνατόν να πραγματοποιηθούν δυο είδη σφαλμάτων:

- Σφάλμα τύπου Ι: Απόρριψη της H_0 ενώ στην πραγματικότητα είναι αληθής
- Σφάλμα τύπου ΙΙ: Αποδοχή της H_0 ενώ στην πραγματικότητα η H_1 είναι αληθής.

Η σημαντικότητα του ελέγχου έγκειται στο κατά πόσο η πιθανότητα πραγματοποίησης των σφαλμάτων τύπου Ι και ΙΙ είναι η μικρότερη δυνατή. Οι αντίστοιχες πιθανότητες πραγματοποίησης αυτών των σφαλμάτων είναι:

α(θ)=P(σφάλμα τύπου Ι)=P(απόρριψη της μηδενικής υπόθεσης ενώ αυτή είναι αληθής)

β(θ)= P(σφάλμα τύπου ΙΙ)=P(αποδοχή της μηδενικής υπόθεσης ενώ αυτή είναι ψευδής)

Το μέγιστο των α(θ) ισούται με α και καλείται επίπεδο σημαντικότητας ή μέγεθος σημαντικότητας(ε.σ). Συνήθως επιλέγεται εξ' αρχής ένα μικρό α και προσπαθούμε να αποφύγουμε την πιθανή απόρριψη της μηδενικής υπόθεσης ενώ είναι αληθής.

Ερμηνεία του α

Εάν για παράδειγμα, για έναν έλεγχο επιλέξουμε επίπεδο σημαντικότητας ίσο με 0.05 και απορρίψουμε την υπόθεση, αυτό σημαίνει ότι σε 100 όμοιες περιπτώσεις μόνο σε 5 είναι δυνατόν να σφάλουμε. Δηλαδή, να απορρίψουμε την υπόθεση ενώ είναι αληθής. Σε μια τέτοια περίπτωση λέμε ότι η υπόθεση απορρίπτεται σε επίπεδο σημαντικότητας 0.05, που σημαίνει η πιθανότητα να γίνει λάθος είναι 0.05. συχνά στην πράξη χρησιμοποιούνται οι τιμές 0.01, 0.05 για ε.σ.

p-value

Είναι φανερό ότι όσο πιο μικρό είναι το επιλεγμένο επίπεδο σημαντικότητας, τόσο πιο δύσκολη είναι η απόρριψη της

μηδενικής υπόθεσης. Συνεπώς αν το επίπεδο σημαντικότητας α_1 μειωθεί σε α_2, τότε είναι δυνατόν μια αρχικά απορριπτέα υπόθεση να μην απορριφθεί. Αντιθέτως, για κάθε άλλο επίπεδο σημαντικότητας μεγαλύτερο του α_1 η μηδενική υπόθεση θα απορρίπτεται. Το μικρότερο επίπεδο σημαντικότητας για το οποίο απορρίπτεται η μηδενική υπόθεση ονομάζεται **p-value.**

Τα περισσότερα στατιστικά πακέτα, δίνουν το **p-value** στα αποτελέσματα των περισσοτέρων στατιστικών ελέγχων. Το **p-value** χρησιμοποιείται με τρόπο παρόμοιο με αυτό του ε.σ., δηλαδή απορρίπτεται η μηδενική υπόθεση αν η τιμή του **p-value** είναι μικρή (μικρότερη του επιλεγμένου ε.σ.).

Το β δηλαδή το σφάλμα τύπου ΙΙ είναι μια συνάρτηση που ορίζεται πάνω στο χωρίο απόρριψης και ονομάζεται ισχύς του ελέγχου.

Τα α , β και το μέγεθος του δείγματος είναι αλληλοεξαρτώμενα μεγέθη, γενικά όμως μπορούμε να πούμε ότι τα α και β μπορούν

να γίνουν όσο μικρά θέλουμε αυξάνοντας το μέγεθος του δείγματος.

Συνοπτική παρουσίαση σφαλμάτων ενός ελέγχου

Υπόθεση	**Αληθής**	**Ψευδής**
Αποδοχή	Σωστή απόφαση	1-β
Απόρριψη	α	Σωστή απόφαση

Είδη ελέγχων

Αν η εναλλακτική υπόθεση αφορά τιμές της παραμέτρου μόνο μικρότερες ή μόνο μεγαλύτερες από μια ορισμένη τιμή, τότε η κρίσιμη περιοχή καθορίζεται από ένα διάστημα όπου οι τιμές στη δειγματοληπτική κατανομή της παραμέτρου είναι μικρότερες ή μεγαλύτερες αντίστοιχα από μια τιμή (κρίσιμη τιμή), η οποία εξαρτάται από το ε.σ. Σε αυτή την περίπτωση ο έλεγχος χαρακτηρίζεται ως **μονόπλευρος έλεγχος**. Αντιθέτως, αν οι τιμές της παραμέτρου είναι απλώς διάφορες από κάποια τιμή, τότε η κρίσιμη περιοχή ορίζεται από δυο διαστήματα και ο έλεγχος χαρακτηρίζεται ως **αμφίπλευρος έλεγχος.**

Κατασκευή Διαστημάτων Εμπιστοσύνης

Έστω ότι επιθυμούμε να εκτιμήσουμε μια παράμετρο θ ενός πληθυσμού. Για την εκτίμησή της βρίσκουμε με τη βοήθεια ενός τυχαίου δείγματος και με προκαθορισμένη πιθανότητα γ δυο αριθμούς θ_1 και θ_2 μεταξύ των οποίων αναμένεται να βρίσκεται η παράμετρος θ του πληθυσμού με πιθανότητα γ δηλαδή:

$$P(\theta_1 \le \theta \le \theta_2) = 1 - \alpha$$

Το διάστημα με όρια τους αριθμούς θ_1 και θ_2 καλείται ΔΕ για την παράμετρο θ του πληθυσμού. Οι αριθμοί θ_1 και θ_2 καλούνται όρια του διαστήματος, το θ_1 καλείται κατώτερο όριο και το θ_2 ανώτερο όριο του διαστήματος. Τα όρια αυτά υπολογίζονται με βάση τις παρατηρήσεις ενός συγκεκριμένου δείγματος. Επομένως, δεν είναι σταθερά, αλλά μεταβάλλονται από δείγμα σε δείγμα.

Η διαδικασία κατασκευής ΔΕ είναι η ακόλουθη:

1. Από έναν πληθυσμό του οποίου θέλουμε να εκτιμήσουμε την παράμετρο θ λαμβάνουμε ένα τυχαίο δείγμα με τιμές $x_1,...,x_n$
2. Υπολογίζουμε με τη βοήθεια του παραπάνω δείγματος μια εκτίμηση της παραμέτρου θ
3. Υπολογίζουμε μια εκτίμηση της τυπικής απόκλισης της παραμέτρου του πληθυσμού από τα στοιχεία του δείγματος
4. Καθορίζουμε την κατανομή που ακολουθεί η παράμετρος θ
5. Προσδιορίζουμε την πιθανότητα με την οποία θα υπολογίσουμε το ΔΕ.

Οι τιμές θ_1 και θ_2 υπολογίζονται από τις ακόλουθες σχέσεις:

$$\theta_1 = \hat{\theta} - \kappa_{\alpha/2}\sigma_\theta$$

$$\theta_2 = \hat{\theta} + \kappa_{\alpha/2}\sigma_\theta$$

Η τιμή $\kappa_{\alpha/2}$ καλείται κρίσιμη τιμή και υπολογίζεται από πίνακες και η τιμή της εξαρτάται από την πιθανότητα α.

Σημείωση

Όσο μικρότερο είναι το τυπικό σφάλμα εκτίμησης της τυπικής απόκλισης και όσο το μέγεθος του δείγματος μεγαλώνει, τόσο μικρότερο θα είναι το εκτιμώμενο ΔΕ. Δηλαδή, όσο μικρότερη είναι η διασπορά της κατανομής του δείγματος, τόσο καλύτερη είναι η εκτίμηση.

Για να μπορέσουμε να καθορίσουμε ένα ΔΕ, πρέπει να γνωρίζουμε την κατανομή πιθανότητας της εκτιμήτριας της παραμέτρου του πληθυσμού.

Διάστημα Εμπιστοσύνης για το μέσο ενός κανονικού δείγματος

α. Γνωστή Διακύμανση

Έστω ότι έχουμε ένα τυχαίο δείγμα $X_1,...,X_n$ που προέρχεται από έναν κανονικό πληθυσμό με γνωστή διακύμανση αλλά άγνωστη μέση τιμή (την οποία ζητάμε να εκτιμήσουμε).

Αν η κατανομή του πληθυσμού είναι κανονική, τότε ανεξάρτητα από το μέγεθος του δείγματος, η κατανομή του δειγματικού μέσου $\bar{X}$ είναι κανονική με μέση τιμή μ και διασπορά $\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n} \Rightarrow \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$. Δηλαδή, $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$ οπότε η $Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \sim \mathrm{N}(0,1)$. Επειδή γνωρίζουμε την κατανομή της μεταβλητής Z, που είναι η τυπική κανονική κατανομή,

μπορούμε να βρούμε δυο αριθμούς, που είναι αντίθετοι λόγω συμμετρίας της τυπικής κανονικής κατανομής ώστε να ισχύει:

$$P(\theta_1 \le \theta \le \theta_2) = 1-\alpha \Rightarrow P(-z_{a/2} \le Z \le z_{a/2}) = 1-\alpha \Rightarrow$$

$$P(-z_{a/2} \le \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \le z_{a/2}) = 1-\alpha \Rightarrow$$

$$P(-z_{a/2}\sigma/\sqrt{n} \le \bar{X}-\mu \le z_{a/2}\sigma/\sqrt{n}) = 1-a \Rightarrow$$

$$P(-\bar{X}-z_{a/2}\sigma/\sqrt{n} \le -\mu \le -\bar{X}+z_{a/2}\sigma/\sqrt{n}) = 1-a \Rightarrow$$

$$P(\bar{X}-z_{a/2}\sigma/\sqrt{n} \le \mu \le \bar{X}+z_{a/2}\sigma/\sqrt{n}) = 1-a$$

Η τελευταία σχέση δηλώνει ότι η πιθανότητα η μέση τιμή του πληθυσμού να βρίσκεται ανάμεσα στις τιμές $\bar{X}-z_{a/2}\sigma/\sqrt{n}$ και $\bar{X}+z_{a/2}\sigma/\sqrt{n}$ είναι 1-α. Άρα το ζητούμενο ΔΕ είναι το διάστημα που έχει όρια αυτές τις τιμές.

Πιο συγκεκριμένα, αν ζητείται να κατασκευαστεί το 95% ΔΕ, δηλαδή

1-α=0.95, τότε από τους πίνακες της τυπικής κανονικής κατανομής, βρίσκουμε ότι $z_{a/2}=1.96$ άρα

$\bar{x} - 1.96\sigma/\sqrt{n} < \mu < \bar{x} + 1.96\sigma/\sqrt{n}$. Αν ζητείται το 99%, τότε $z_{a/2} = 2.58$ άρα $\bar{x} - 2.58\sigma/\sqrt{n} < \mu < \bar{x} + 2.58\sigma/\sqrt{n}$.

Στον ακόλουθο πίνακα δίνονται μερικές κρίσιμες τιμές της τυπικής κανονικής που αντιστοιχούν σε διάφορους συντελεστές εμπιστοσύνης.

γ.	99.73 %	99 %	98 %	96 %	95.45 %	95 %	90 %	80 %	68.27 %	50%
$z_{a/2}$	3	2.58	2.33	2.05	2	1.96	1.645	1.28	1	0.6745

Σημείωση

Αν η κατανομή του πληθυσμού **δεν είναι κανονική**, τότε σύμφωνα με ένα θεώρημα (γνωστό ως Κ.Ο.Θ), η κατανομή του δειγματικού μέσου προσεγγίζει την κανονική κατανομή με μέση τιμή μ και διασπορά $\frac{\sigma^2}{n}$, με την προϋπόθεση το δείγμα να είναι μεγάλο, δηλαδή **n > 30** .

α. Άγνωστη Διακύμανση

Συνήθως η διακύμανση του πληθυσμού είναι άγνωστη και εκτιμάται από τη δειγματική διασπορά (όχι όμως αμερόληπτα). Τότε η δειγματική κατανομή της ποσότητας $\frac{\bar{X}-\mu}{s/\sqrt{n}}$ δεν είναι η τυπική κατανομή αλλά η t – κατανομή με n-1 βαθμούς ελευθερίας (β.ε). Τότε τα σημεία αντικαθίστανται από τα αντίστοιχα της t κατανομής. Οπότε το ΔΕ δίνεται:

$$\bar{x} - t_{n-1,a/2}s/\sqrt{n} < \mu < \bar{x} + t_{n-1,a/2}s/\sqrt{n}$$

Σημείωση

Αν n > 30, τότε αντί για την t κατανομή μπορούμε να χρησιμοποιήσουμε την τυπική κανονική κατανομή.

Διάστημα Εμπιστοσύνης για τη διασπορά ενός κανονικού δείγματος

α. Γνωστό μέσο

Αν η μεταβλητή X του πληθυσμού κατανέμεται κανονικά με άγνωστη διασπορά αλλά γνωστό μέσο, τότε η μεταβλητή

$\frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\sigma^2} \sim \chi_n^2$. Οι βαθμοί ελευθερίας είναι n και όχι n-1 γιατί ο μέσος είναι γνωστός. Γνωρίζοντας ποια κατανομή ακολουθεί η παραπάνω ποσότητα μας βοηθάει στον προσδιορισμό των ορίων του διαστήματος για τη διασπορά. Από τους πίνακες της χ_n^2 βρίσκουμε δυο τιμές $\chi_{n,1-a/2}^2$ και $\chi_{n,a/2}^2$ που αντιστοιχούν σε πιθανότητες 1-α/2 και α/2 αντίστοιχα. Οπότε έχουμε:

$$P(\chi_{n,1-a/2}^2 \le \frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\sigma^2} \le \chi_{n,a/2}^2) = 1-a \Leftrightarrow$$

$$P(\frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\chi_{n,a/2}^2} \le \sigma^2 \le \frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\chi_{n,1-a/2}^2}) = 1-a$$

άρα το ζητούμενο ΔΕ, με σ.ε. (1-α)100% θα είναι:

$$\left(\frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\chi_{n,a/2}^2}, \frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\chi_{n,1-a/2}^2} \right)$$

Παρατήρηση

Το ΔΕ της διασποράς θα είναι τόσο μικρό, όσο μεγαλύτερο θα είναι το μέγεθος του δείγματος, για καθορισμένο σ.ε.

Β. Άγνωστο μέσο

Αν θεωρείται άγνωστος και ο μέσος, τότε η μεταβλητή $\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}$. Οπότε

$$P(\chi^2_{n-1,1-a/2} \le \frac{(n-1)s^2}{\sigma^2} \le \chi^2_{n-1,a/2}) = 1-a \Leftrightarrow$$
$$P(\frac{(n-1)s^2}{\chi^2_{n-1,a/2}} \le \sigma^2 \le \frac{(n-1)s^2}{\chi^2_{n-1,1-a/2}}) = 1-a$$

άρα το ζητούμενο ΔΕ, με σ.ε. (1-α)100% θα είναι: $\left(\frac{(n-1)s^2}{\chi^2_{n-1,a/2}}, \frac{(n-1)s^2}{\chi^2_{n-1,1-a/2}}\right)$

Παρατήρηση

Η ποσότητα $(n-1)s^2$ είναι ίση με $\sum_{i=1}^{n}(x_i-\bar{x})^2$ (προκύπτει από τον τύπο της δειγματικής διασποράς).

Διάστημα Εμπιστοσύνης για τη διαφορά δυο πληθυσμών

α. Διασπορές γνωστές

Έστω, τώρα ότι έχουμε δυο κανονικούς πληθυσμούς και από αυτούς παίρνουμε δυο ανεξάρτητα δείγματα μεγέθους n_1 και n_2 αντίστοιχα. Ο πρώτος πληθυσμός έχει μέσο μ_1 και διασπορά σ_1^2 και ο δεύτερος μ_2 και σ_2^2. Οι διασπορές θεωρούνται γνωστές. Για τον προσδιορισμό του ΔΕ της διαφοράς των δυο μέσων ακολουθείται η εξής διαδικασία: Οι δειγματικοί μέσοι είναι αντίστοιχα $\bar{x}_1$ και $\bar{x}_2$. Τότε ο $\bar{x}_1 \sim N(\mu_1, \frac{\sigma_1^2}{n_1})$ και $\bar{x}_2 \sim N(\mu_2, \frac{\sigma_2^2}{n_2})$, οπότε η διαφορά των δυο δειγματικών μέσων είναι μια άλλη τυχαία μεταβλητή που ακολουθεί την $N(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2})$ και η μεταβλητή $Z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1)$. Συνεπώς θα έχουμε:

$$P(-z_{a/2} < \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} < z_{a/2}) = 1 - a \Leftrightarrow$$

$$P((\bar{x}_1 - \bar{x}_2) - z_{a/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} < (\mu_1 - \mu_2) < (\bar{x}_1 - \bar{x}_2) + z_{a/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}) = 1 - a$$

Οπότε:

$$\left(\overline{x}_1 - \overline{x}_2 - z_{a/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}, \overline{x}_1 - \overline{x}_2 + z_{a/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \right)$$

Παρατήρηση

Αν $\sigma_1^2 = \sigma_2^2 = \sigma^2$ τότε

$$\left(\overline{x}_1 - \overline{x}_2 - z_{a/2}\sqrt{\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}}, \overline{x}_1 - \overline{x}_2 + z_{a/2}\sqrt{\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}} \right)$$ ή ισοδύναμα

$$\left(\overline{x}_1 - \overline{x}_2 - z_{a/2}\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}, \overline{x}_1 - \overline{x}_2 + z_{a/2}\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \right)$$

β. άγνωστες διασπορές και τα δείγματα μεγάλα

Σε αυτήν την περίπτωση οι άγνωστες διασπορές θα αντικατασταθούν με τις αντίστοιχες δειγματικές. Άρα το ζητούμενο ΔΕ:

$$\left(\overline{x}_1 - \overline{x}_2 - z_{a/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}, \overline{x}_1 - \overline{x}_2 + z_{a/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \right)$$

γ. άγνωστες διασπορές και ίσες και τα δείγματα μικρά

Σε αυτήν την περίπτωση η εκτίμηση της κοινής δειγματικής διασποράς δίνεται από τον ακόλουθο τύπο:

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

Τότε η μεταβλητή $t = \dfrac{(\overline{x}_1 - \overline{x}_2) - (\mu_1 - \mu_2)}{s_p\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}} \sim t_\nu$ όπου $\nu = n_1 + n_2 - 2$.

Η κρίσιμη τιμή $t_{\nu,a/2}$ αφήνει δεξιά της επιφάνεια ίση με α/2. Το ΔΕ είναι:

$$\left(\overline{x}_1 - \overline{x}_2 - t_{\nu,a/2}s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}, \overline{x}_1 - \overline{x}_2 + t_{\nu,a/2}s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right)$$

δ. άγνωστες διασπορές και άνισες και τα δείγματα μικρά

Σε αυτήν την περίπτωση ισχύει προσεγγιστικά η *t* κατανομή και το ΔΕ είναι

$$\left(\overline{x}_1 - \overline{x}_2 - t_{\nu,a/2}s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}, \overline{x}_1 - \overline{x}_2 + t_{\nu,a/2}s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right)$$

με $\nu = \dfrac{1}{\dfrac{\left(\dfrac{s_1^2/n_1}{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}\right)^2}{n_1 - 1} + + \dfrac{\left[1 - \left(\dfrac{s_1^2/n_1}{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}\right)\right]^2}{n_2 - 1}}$ για $s_1^2 > s_2^2$

Παρατήρηση

Στην περίπτωση μη κανονικών πληθυσμών εργαζόμαστε όπως και στις αντίστοιχες περιπτώσεις κανονικών πληθυσμών, θεωρώντας ότι οι κατανομές είναι προσεγγιστικές.

Παρατήρηση

Παρατηρήσεις κατά ζεύγη

Πολλές φορές στη δειγματοληψία από δυο πληθυσμούς κάποιοι εξωτερικοί παράγοντες προκαλούν σημαντική διαφορά στους μέσους ακόμη και αν πραγματικά δεν υπάρχει. Αντιστρόφως, αυτοί οι παράγοντες μπορούν να καλύψουν κάποια υπαρκτή διαφορά. Το μειονέκτημα αυτό δύναται να ξεπεραστεί λαμβάνοντας κάθε παρατήρηση κατά ζεύγη. Δηλαδή, προσπαθούμε να δημιουργήσουμε τις παρατηρήσεις των δυο δειγμάτων (ίσου μεγέθους) κατά ζεύγη. Θεωρούμε βέβαια ότι τα δυο μέλη κάθε ζεύγους έχουν κάποια κοινά χαρακτηριστικά (να είναι το ίδιο άτομο μετρημένο σε δυο διαφορετικές καταστάσεις ή διαφορετικά άτομα μετρημένα στις ίδιες ακριβώς εξωτερικές συνθήκες). Έστω οι παρατηρήσεις $(x_1, y_1),...,(x_n, y_n)$, αν πάρουμε τις διαφορές $d_i = x_i - y_i$ θα έχουμε n παρατηρήσεις που

αποτελούνται από τις διαφορές των αρχικών μας παρατηρήσεων σε κάθε ζεύγος. Έστω ότι οι διασπορές είναι ίσες και άγνωστες, τότε σ_d^2 η οποία εκτιμάται από την ποσότητα $s_d^2 = \frac{1}{n-1}\sum(d_i - \bar{d})^2$. Οπότε το ΔΕ είναι

$$\left(\bar{d} - t_{n-1,a/2} s_d \sqrt{\frac{1}{n}}, \bar{d} + t_{n-1,a/2} s_d \sqrt{\frac{1}{n}} \right)$$

Διάστημα εμπιστοσύνης για το λόγο των διασπορών

α. Οι μέσοι γνωστοί

Στην περίπτωση αυτή το ΔΕ του λόγου των διασπορών δίνεται:

$$\frac{1}{F_{n_1,n_2,a/2}} \frac{\frac{\sum(x_i-\mu_1)^2}{n_1-1}}{\frac{\sum(y_i-\mu_2)^2}{n_2-1}} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{1}{F_{n_1,n_2,1-a/2}} \frac{\frac{\sum(x_i-\mu_1)^2}{n_1-1}}{\frac{\sum(y_i-\mu_2)^2}{n_2-1}}$$

Επίσης ισχύει: $F_{n_1,n_2,1-a/2} = \frac{1}{F_{n_2,n_1,a/2}}$

β. οι μέσοι άγνωστοι

$$\frac{1}{F_{n_1,n_2,a/2}}\frac{\frac{\sum(x_i-\bar{x})^2}{n_1-1}}{\frac{\sum(y_i-\bar{y})^2}{n_2-1}}<\frac{\sigma_1^2}{\sigma_2^2}<\frac{1}{F_{n_1,n_2,1-a/2}}\frac{\frac{\sum(x_i-\bar{x})^2}{n_1-1}}{\frac{\sum(y_i-\bar{y})^2}{n_2-1}}\Leftrightarrow\frac{1}{F_{n_1,n_2,a/2}}\frac{s_1^2}{s_2^2}<\frac{\sigma_1^2}{\sigma_2^2}<\frac{1}{F_{n_1,n_2,1-a/2}}\frac{s_1^2}{s_2^2}$$

Διάστημα Εμπιστοσύνης για αναλογίες

Με τον όρο αναλογία εννοούμε την αναλογία ή το ποσοστό *p* των ατόμων ενός πληθυσμού που έχουν κάποιο συγκεκριμένο χαρακτηριστικό.

Έστω μια μεταβλητή που παίρνει τιμές 1 ή 0 αντίστοιχα αν το άτομο έχει ή δεν έχει το χαρακτηριστικό. Τότε αν πάρουμε ένα τυχαίο δείγμα μεγέθους n από τον πληθυσμό, η συνάρτηση $T=\sum_{i=1}^{n}X_i$ δείχνει πόσα άτομα έχουν αυτό το χαρακτηριστικό, ενώ η $\bar{X}=\frac{T}{n}$ παρέχει το ποσοστό ή την αναλογία των ατόμων με το χαρακτηριστικό. Επομένως η $\bar{X}$ μπορεί να χρησιμοποιηθεί για την εκτίμηση του *p* του πληθυσμού. Είναι προφανές ότι η δειγματική κατανομή της $\bar{X}$ είναι διωνυμική με μέση τιμή και διασπορά αντίστοιχα, p και $\frac{pq}{n}$. Επομένως για τον προσδιορισμό του ΔΕ του ποσοστού, θεωρούμε ότι για μεγάλο

n, η κατανομή του $\bar{X}$ είναι προσεγγιστικά κανονική και συνεπώς η $Z = \frac{\bar{X} - p}{\sqrt{\frac{pq}{n}}} \sim N(0,1)$(Και όταν το n δεν είναι μεγάλο δεν υπάρχει σοβαρό σφάλμα). Άρα, $-z_{a/2} < Z < z_{a/2}$ και λύνοντας την ανίσωση ως προς p έχουμε: $\hat{p} - z_{a/2}\sqrt{\frac{pq}{n}} < p < \hat{p} + z_{a/2}\sqrt{\frac{pq}{n}}$

Παρατήρηση

Αν ζητείται το ΔΕ της διαφοράς $p_1 - p_2$ δυο πληθυσμών που ακολουθούν τη διωνυμική, με τη βοήθεια δυο δειγμάτων n_1 και n_2, τότε ισχύει:

$$\hat{p}_1 - \hat{p}_2 - z_{a/2}\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}} < p_1 - p_2 < \hat{p}_1 - \hat{p}_2 + z_{a/2}\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}$$

Μέγεθος δείγματος

Το σπουδαιότερο πρόβλημα που εμφανίζεται κατά το στάδιο σχεδιασμού της έρευνας είναι ο προσδιορισμός του μεγέθους του δείγματος. Το μέγεθος του δείγματος εξαρτάται από τους εξής παράγοντες:

α. το βαθμό εμπιστοσύνης

β. την κατανομή του πληθυσμού

γ. το σφάλμα δειγματοληψίας δηλαδή το βαθμό ακρίβειας με τον οποίο θέλουμε να πραγματοποιηθεί η έρευνά μας.

Το σφάλμα αυτό επηρεάζεται από το βαθμό εμπιστοσύνης, από την κατανομή του πληθυσμού και από το μέγεθος του δείγματος. Όσο μεγαλύτερο είναι το μέγεθος του δείγματος, τόσο καλύτερα τα αποτελέσματα. Το μέγεθος φυσικά εξαρτάται από το σκοπό της έρευνας και από το πόσο ακριβής θέλουμε να είναι η δειγματική εκτίμηση. Η εκτίμηση αυτή προσδιορίζεται να είναι μέσα σε κάποια όρια με βάση κάποια ορισμένη πιθανότητα (95% ή 99% συνήθως).

Έστω για παράδειγμα, ότι θέλουμε να προσδιορίσουμε το δειγματοληπτικό σφάλμα του δειγματικού μέσου από το μέσο του πληθυσμού, δηλαδή $\lambda = |\bar{X} - \mu|$. Το σφάλμα αυτό δεν μπορεί να υπερβεί την ποσότητα $z_{a/2}\frac{\sigma}{\sqrt{n}}$ με σ.ε. γ. Αν ο δειγματικός

μέσος ταυτίζεται με το μέσο του πληθυσμού τότε το σφάλμα είναι ίσο με μηδέν. Αν ο μέσος του πληθυσμού βρίσκεται στο ανώτερο σημείο του ΔΕ, τότε $\lambda = z_{a/2}\frac{\sigma}{\sqrt{n}}$ και λύνοντας αυτή την σχέση ως προς το μέγεθος του δείγματος, βρίσκουμε ότι: $n = \frac{z_{a/2}^2\sigma^2}{\lambda^2}$. Η σχέση αυτή προσδιορίζει το μέγιστο μέγεθος του δείγματος με το οποίο εξασφαλίζεται πλάτος διαστήματος λ για σ.ε. γ.

Για παράδειγμα, θέλουμε να προσδιοριστεί το μέγεθος του δείγματος ενός κανονικού πληθυσμού, ώστε το δειγματοληπτικό σφάλμα να είναι μικρότερο του 3 με γ=0.95 και τυπική απόκλιση ίση με 16. Σύμφωνα με τα ποιο πάνω έχουμε $n = \frac{z_{a/2}^2\sigma^2}{\lambda^2} = \frac{1,96^2 16^2}{3^2} = 109$. Αυτό σημαίνει ότι με βεβαιότητα 95%, εάν πάρουμε ένα δείγμα μεγέθους 109, ο δειγματικός μέσος δεν θα διαφέρει από τον πραγματικό μέσο παραπάνω από 3 μονάδες.

Παρατήρηση

Αν θέλουμε η εκτίμηση του ποσοστού να διαφέρει από το πραγματικό ποσοστό λιγότερο ενός αριθμού λ με πιθανότητα 1-α, τότε πρέπει να επιλέγει δείγμα μεγέθους n που υπολογίζεται από την σχέση: $n = \frac{z_{a/2}^2 p(1-p)}{\lambda^2}$

Παρατήρηση

Στην πράξη υπάρχουν παράγοντες που επιβάλλουν περιορισμούς στο μέγεθος του αναγκαίου δείγματος. Οι παράγοντες αυτοί είναι συνήθως τα χρήματα που διατίθονται για την έρευνα και ο χρονικός περιορισμός πραγματοποίησής της. Στο κόστος της δειγματοληψίας ανήκουν τα γενικά έξοδα και το κόστος που αντιστοιχεί σε κάθε συνέντευξη για τη συμπλήρωση του ερωτηματολογίου που είναι ανάλογο με το μέγεθος του δείγματος. Αν συμβολίσουμε με Γ τα γενικά έξοδα, Κ το συνολικό κόστος και με κ τα έξοδα για κάθε συνέντευξη, τότε ισχύει η σχέση: Κ=Γ+nκ

Έλεγχοι Υποθέσεων για τη μέση τιμή

Σε αυτές τις υποθέσεις αυτό που αρχικά μας ενδιαφέρει είναι αν ο μέσος του πληθυσμού ισούται με κάποιο προκαθορισμένο αριθμό. Οπότε σαν μηδενική υπόθεση ορίζεται η υπόθεση ο μέσος να ισούται με αυτήν την συγκεκριμένη τιμή. Ως εναλλακτική δύναται να ορισθούν τρεις διαφορετικές: α. η τιμή του μέσου να είναι διαφορετική από αυτήν την τιμή, β. να είναι μεγαλύτερη και γ. να είναι μικρότερη. Οι έλεγχοι αυτοί παρουσιάζονται συνοπτικά ως εξής:

$$H_0 : \mu = \mu_0 \quad \text{vs} \quad H_1 : \mu \neq \mu_0$$
$$H_0 : \mu = \mu_0 \quad \text{vs} \quad H_1 : \mu > \mu_0$$
$$H_0 : \mu = \mu_0 \quad \text{vs} \quad H_1 : \mu < \mu_0$$

Ο πρώτος έλεγχος χαρακτηρίζεται ως αμφίπλευρος και οι άλλοι δυο ως μονόπλευροι.

Α. Έλεγχος μέσης τιμής ενός πληθυσμού με γνωστή διασπορά (πληθυσμός κανονικός ή το μέγεθος του δείγματος μεγάλο ανεξαρτήτως της μορφής της κατανομής)

Γνωρίζουμε σε αυτήν την περίπτωση ότι η κατανομή του δειγματικού μέσου είναι κανονική και επομένως η μεταβλητή $Z = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}} \sim N(0,1)$. Η συνάρτηση αυτή εξαρτάται μόνο από την τιμή του δειγματικού μέσου και χρησιμοποιείται ως κριτήριο αποδοχής ή απόρριψης της μηδενικής υπόθεσης.

Θα εξετάσουμε αρχικά τον αμφίπλευρο έλεγχο σε ε.σ α. Η περιοχή απόρριψης για ε.σ. α ορίζεται από τις σχέσεις:

$Z = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}} > z_{a/2}$ ή $Z = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}} < -z_{a/2}$ και η υπόθεση γίνεται δεκτή, αν ισχύει: $-z_{a/2} < Z < z_{a/2}$. Τις τιμές $-z_{a/2}$ και $z_{a/2}$ βρίσκονται από τους πίνακες της κανονικής κατανομής.

Άλλο κριτήριο ισοδύναμο του προηγούμενου είναι η εύρεση δυο ακραίων τιμών c_1 και c_2 που ορίζουν την περιοχή αποδοχής και απόρριψης. Για το σκοπό αυτό χρησιμοποιείται η ακόλουθη σχέση:

$P(c_1 < \bar{X} < c_2) = 1 - a$ από την οποία βρίσκουμε ότι: $c_1 = \mu_0 - z_{a/2} \frac{\sigma}{\sqrt{n}}$ και $c_2 = \mu_0 + z_{a/2} \frac{\sigma}{\sqrt{n}}$. Αν ο δειγματικός μέσος

βρίσκεται μέσα στο διάστημα που ορίζουν αυτές οι τιμές η μηδενική υπόθεση γίνεται δεκτή, ενώ αν βρίσκεται εκτός απορρίπτεται και γίνεται αποδεκτή η εναλλακτική.

Τώρα ας εξετάσουμε το μονόπλευρο έλεγχο: $H_0 : \mu = \mu_0$ vs $H_1 : \mu > \mu_0$. Στην π4ερίπτωση αυτή μόνο πολύ μεγάλες τιμές της *Z* λόγω μεγάλης τιμής του δειγματικού μέσου οδηγούν στην απόρριψη της μηδενικής υπόθεσης. Η υπόθεση απορρίπτεται αν ισχύει: $Z = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}} > z_a$. Η κρίσιμη τιμή υπολογίζεται πάλι από πίνακες της κανονικής κατανομής. Άλλο κριτήριο αποδοχής ή απόρριψης, είναι η εύρεση μιας ακραίας τιμής που χωρίζει την περιοχή αποδοχής από την περιοχή απόρριψης. Η κρίσιμη τιμή υπολογίζεται από την σχέση $P(\bar{X} \geq c) = a \Leftrightarrow c = \mu_0 + z_a \frac{\sigma}{\sqrt{n}}$. Αν ο δειγματικός μέσος είναι μεγαλύτερος από αυτήν την τιμή απορρίπτεται η μηδενική υπόθεση, αλλιώς την αποδεχόμαστε.

Τώρα ας εξετάσουμε το μονόπλευρο έλεγχο: $H_0 : \mu = \mu_0$ vs $H_1 : \mu < \mu_0$. Η υπόθεση απορρίπτεται αν ισχύει: $Z = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}} < -z_a$. Η κρίσιμη τιμή υπολογίζεται πάλι από πίνακες της κανονικής κατανομής. Άλλο κριτήριο αποδοχής ή απόρριψης, είναι η εύρεση μιας ακραίας τιμής που χωρίζει την περιοχή αποδοχής από την περιοχή απόρριψης. Η κρίσιμη τιμή υπολογίζεται από την σχέση $P(\bar{X} < c) = a \Leftrightarrow c = \mu_0 - z_a \frac{\sigma}{\sqrt{n}}$. Αν ο δειγματικός μέσος είναι μικρότερος από αυτήν την τιμή απορρίπτεται η μηδενική υπόθεση, αλλιώς την αποδεχόμαστε.

Συνοπτικά για ε.σ. α έχουμε:

$H_0 : \mu = \mu_0$ vs $H_1 : \mu \neq \mu_0$ απορρίπτεται η H_0 αν $|z| > z_{a/2}$

$H_0 : \mu = \mu_0$ vs $H_1 : \mu > \mu_0$ απορρίπτεται η H_0 αν $z > z_a$

$H_0 : \mu = \mu_0$ vs $H_1 : \mu < \mu_0$ απορρίπτεται η H_0 αν $z < -z_a$

Ο παρακάτω πίνακας παρέχει τις κρίσιμες τιμές των κυριότερων επιπέδων σημαντικότητας της Z για δίπλευρους και μονόπλευρους ελέγχους:

ε.σ. α	**0,01**	**0,05**	**0,1**
Κρίσιμες τιμές για μονόπλευρο έλεγχο	-2,33 +2,33	-1,645 +1,645	-1,28 +1,28
Κρίσιμες τιμές για δίπλευρο έλεγχο	-2,58 +2,58	-1,96 +1,96	-1,645 +1,645

Β. έλεγχος μέσης τιμής με διασπορά άγνωστη και το μέγεθος του δείγματος μικρό

Σε αυτήν την περίπτωση έχουμε τη μεταβλητή $t = \frac{\bar{X} - \mu_0}{s/\sqrt{n}} \sim t_{n-1}$.

Εν συνεχεία η τιμή αυτή συγκρίνεται με τη κρίσιμη τιμή που προκύπτει από τους πίνακες της student κατανομής με τους αντίστοιχους β.ε. Πιο συγκεκριμένα έχουμε:

$H_0 : \mu = \mu_0$ vs $H_1 : \mu \neq \mu_0$ απορρίπτεται η H_0 αν $|t| > t_{n-1,a/2}$ ή $\overline{x} > c_2, \overline{x} < c_1$ όπου $c_1 = \mu_0 - t_{n-1,a/2} s/\sqrt{n}$ και $c_1 = \mu_0 + t_{n-1,a/2} s/\sqrt{n}$

$H_0 : \mu = \mu_0$ vs $H_1 : \mu > \mu_0$ απορρίπτεται η H_0 αν $t > t_{n-1,a}$ ή $\overline{x} > \mu_0 + t_{n-1,a} s/\sqrt{n}$

$H_0 : \mu = \mu_0$ vs $H_1 : \mu < \mu_0$ απορρίπτεται η H_0 αν $t < -t_{n-1,a}$ ή $\overline{x} < \mu_0 - t_{n-1,a} s/\sqrt{n}$

Παρατήρηση

Γνωρίζουμε ότι αν το $p-value$ μικρότερο από το ε.σ. απορρίπτεται η μηδενική υπόθεση, αν είναι μεγαλύτερο δεν απορρίπτεται και είναι πολύ μεγάλο είναι πιθανόν η μηδενική να είναι αληθής. Αντιθέτως, αν είναι μικρό είναι πιθανόν να μην είναι αληθινή. Οι ακόλουθοι ισχυρισμοί συντελούν σε μια καλή ερμηνεία του $p-value$:

Εάν είναι μικρότερο από :

1. 0.10, τότε υπάρχει ένδειξη ότι η μηδενική υπόθεση δεν είναι αληθής
2. 0.05, τότε υπάρχει ισχυρή ένδειξη ότι η μηδενική υπόθεση δεν είναι αληθής
3. 0.01, τότε υπάρχει πολύ ισχυρή ένδειξη ότι η μηδενική υπόθεση δεν είναι αληθής
4. 0.001, τότε υπάρχει παρά πολύ ισχυρή ένδειξη ότι η μηδενική υπόθεση δεν είναι αληθής

Η χρησιμοποίηση της τιμής p αντί του α στην αντιμετώπιση ενός στατιστικού ελέγχου δεν μεταβάλλει την κλασσική στατιστική μεθοδολογία. Απλά εκείνο που συμβαίνει είναι ότι ο ερευνητής αναφέρει στον ενδιαφερόμενο τη τιμή p και αφήνει στην επιλογή του αν θα απορριφθεί ή όχι η μηδενική υπόθεση. Με λίγα λόγια μεταφέρεται η επιλογή του ε.σ. από τον ερευνητή στον ενδιαφερόμενο.

Συνοπτικά έχουμε:

1. $H_1 : \mu < \mu_0, p - value = P(Z_0 \leq z_0)$
2. $H_1 : \mu > \mu_0, p - value = P(Z_0 \geq z_0)$

3. $H_1 : \mu \neq \mu_0, p - value = P(|Z_0| \geq z_0)$

για την απλή περίπτωση μηδενικής υπόθεσης. Για πιο σύνθετη ορίζονται ως οι μέγιστες πιθανότητες των ενδεχομένων για τα οποία αυτές ορίσθηκαν στην περίπτωση της απλής.

Έλεγχος υποθέσεων για τη διασπορά ενός κανονικού πληθυσμού

Οι πιο συνηθισμένοι συνδυασμοί για τον έλεγχο της διασποράς είναι οι παρακάτω:

$$H_0 : \sigma^2 = \sigma_0^2 \text{ vs } H_1 : \sigma^2 \neq \sigma_0^2$$
$$H_0 : \sigma^2 = \sigma_0^2 \text{ vs } H_1 : \sigma^2 > \sigma_0^2$$
$$H_0 : \sigma^2 = \sigma_0^2 \text{ vs } H_1 : \sigma^2 < \sigma_0^2$$

Η μηδενική υπόθεση των παραπάνω περιπτώσεων ελέγχεται με το κριτήριο:

$$\chi^2 = \frac{\sum (x_i - \mu)^2}{\sigma_0^2} \sim \chi_\nu^2, \text{ αν μ γνωστό και ν=n}$$

και

$$\chi^2 = \frac{\sum (x_i - \bar{x})^2}{\sigma_0^2} = \frac{(n-1)s^2}{\sigma_0^2} \sim \chi_\nu^2, \text{ αν μ άγνωστο όπου ν=n-1}$$

Αν ο έλεγχος αναφέρεται στην πρώτη περίπτωση και ισχύει $\chi^2 > \chi^2_{\nu,a/2}$ ή $\chi^2 < \chi^2_{\nu,1-a/2}$ τότε απορρίπτεται η μηδενική υπόθεση.

Αν ο έλεγχος αναφέρεται στην δεύτερη περίπτωση και ισχύει $\chi^2 > \chi^2_{\nu,a}$ τότε απορρίπτεται η μηδενική υπόθεση.

Αν ο έλεγχος αναφέρεται στην τρίτη περίπτωση και ισχύει $\chi^2 < \chi^2_{\nu,1-a}$ τότε απορρίπτεται η μηδενική υπόθεση.

Έλεγχος υποθέσεων της διαφοράς των μέσων δυο κανονικών πληθυσμών

Αν από δυο κανονικούς πληθυσμούς που έχουν μέσους μ_1 και μ_2 και διασπορές σ_1^2, σ_2^2 αντίστοιχα, πάρουμε δυο τυχαία και ανεξάρτητα μεταξύ τους δείγματα μεγέθους n_1, n_2 αντίστοιχα και υπολογίσουμε τους αντίστοιχους δειγματικούς μέσους, τότε μπορούμε να ελέγξουμε την υπόθεση $H_0 : \mu_1 - \mu_2$ με τους παρακάτω συνδυασμούς:

1.

$$H_0 : \mu_1 - \mu_2 = 0 \Leftrightarrow H_0 : \mu_1 = \mu_2$$
vs
$$H_1 : \mu_1 - \mu_2 \neq 0 \Leftrightarrow H_1 : \mu_1 \neq \mu_2$$

2.

$$H_0 : \mu_1 - \mu_2 = 0 \Leftrightarrow H_0 : \mu_1 = \mu_2$$
vs
$$H_1 : \mu_1 - \mu_2 > 0 \Leftrightarrow H_1 : \mu_1 > \mu_2$$

3.

$$H_0 : \mu_1 - \mu_2 = 0 \Leftrightarrow H_0 : \mu_1 = \mu_2$$
vs
$$H_1 : \mu_1 - \mu_2 < 0 \Leftrightarrow H_1 : \mu_1 < \mu_2$$

Ανάλογα με το αν οι διασπορές είναι γνωστές ή άγνωστες διακρίνουμε τις εξής περιπτώσεις:

Α. γνωστές διασπορές

Τα δυο τυχαία και ανεξάρτητα δείγματα έχουν δειγματικούς μέσους $\bar{X}_1$ και $\bar{X}_2$ αντίστοιχα, τότε η κατανομή της διαφοράς των δειγματικών μέσων είναι και αυτή κανονική με μέση τιμή $\mu_1 - \mu_2$ και διασπορά $\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$, τότε η μεταβλητή $Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1)$. Και με βάση τη μηδενική

υπόθεση αυτών των περιπτώσεων η μεταβλητή απλοποιείται ως

$Z = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1)$. Η σχέση αυτή αποτελεί κριτήριο ελέγχου

των παραπάνω υποθέσεων 1,2,3. Έτσι έχουμε:

1.

$H_0 : \mu_1 - \mu_2 = 0 \Leftrightarrow H_0 : \mu_1 = \mu_2$

vs

$H_1 : \mu_1 - \mu_2 \neq 0 \Leftrightarrow H_1 : \mu_1 \neq \mu_2$

σε αυτήν την περίπτωση η μηδενική περίπτωση απορρίπτεται όταν $Z > z_{a/2}$ ή όταν $Z < -z_{a/2}$

2.

$H_0 : \mu_1 - \mu_2 = 0 \Leftrightarrow H_0 : \mu_1 = \mu_2$

vs

$H_1 : \mu_1 - \mu_2 > 0 \Leftrightarrow H_1 : \mu_1 > \mu_2$

σε αυτήν την περίπτωση η μηδενική περίπτωση απορρίπτεται όταν $Z > z_a$

3.

$H_0 : \mu_1 - \mu_2 = 0 \Leftrightarrow H_0 : \mu_1 = \mu_2$

vs

$H_1 : \mu_1 - \mu_2 < 0 \Leftrightarrow H_1 : \mu_1 < \mu_2$

σε αυτήν την περίπτωση η μηδενική περίπτωση απορρίπτεται όταν $Z < -z_a$

β. οι διασπορές άγνωστες και άνισες και το μέγεθος του δείγματος μικρό

Τότε έχουμε το εξής κριτήριο:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

οπότε έχουμε:

1.

$H_0 : \mu_1 - \mu_2 = 0 \Leftrightarrow H_0 : \mu_1 = \mu_2$

vs

$H_1 : \mu_1 - \mu_2 \neq 0 \Leftrightarrow H_1 : \mu_1 \neq \mu_2$

σε αυτήν την περίπτωση η μηδενική περίπτωση απορρίπτεται όταν $t > t_{\nu,a/2}$ ή όταν $t < -t_{\nu,a/2}$

2.

$H_0 : \mu_1 - \mu_2 = 0 \Leftrightarrow H_0 : \mu_1 = \mu_2$
vs
$H_1 : \mu_1 - \mu_2 > 0 \Leftrightarrow H_1 : \mu_1 > \mu_2$

σε αυτήν την περίπτωση η μηδενική περίπτωση απορρίπτεται όταν $t > t_{\nu,a}$

3.

$H_0 : \mu_1 - \mu_2 = 0 \Leftrightarrow H_0 : \mu_1 = \mu_2$
vs
$H_1 : \mu_1 - \mu_2 < 0 \Leftrightarrow H_1 : \mu_1 < \mu_2$

σε αυτήν την περίπτωση η μηδενική περίπτωση απορρίπτεται όταν $t < -t_{\nu,a}$

όπου οι β.ε. $\nu = n_1 + n_2 - 2$

γ. οι διασπορές άγνωστες και ίσες και το μέγεθος του δείγματος μικρό

Τότε έχουμε το εξής κριτήριο:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{s^2(\frac{1}{n_1} + \frac{1}{n_2})}} \text{ όπου } s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

οπότε έχουμε:

1.

$H_0 : \mu_1 - \mu_2 = 0 \Leftrightarrow H_0 : \mu_1 = \mu_2$
vs
$H_1 : \mu_1 - \mu_2 \neq 0 \Leftrightarrow H_1 : \mu_1 \neq \mu_2$

σε αυτήν την περίπτωση η μηδενική περίπτωση απορρίπτεται όταν $t > t_{\nu,a/2}$ ή όταν $t < -t_{\nu,a/2}$

2.

$H_0 : \mu_1 - \mu_2 = 0 \Leftrightarrow H_0 : \mu_1 = \mu_2$
vs
$H_1 : \mu_1 - \mu_2 > 0 \Leftrightarrow H_1 : \mu_1 > \mu_2$

σε αυτήν την περίπτωση η μηδενική περίπτωση απορρίπτεται όταν $t > t_{\nu,a}$

3.

$H_0 : \mu_1 - \mu_2 = 0 \Leftrightarrow H_0 : \mu_1 = \mu_2$
vs
$H_1 : \mu_1 - \mu_2 < 0 \Leftrightarrow H_1 : \mu_1 < \mu_2$

σε αυτήν την περίπτωση η μηδενική περίπτωση απορρίπτεται όταν $t < -t_{\nu,a}$

όπου οι β.ε. $\nu = n_1 + n_2 - 2$

Παρατήρηση

Έστω, τώρα ότι έχουμε δυο τυχαία και ανεξάρτητα δείγματα, τα οποία όμως δεν προέρχονται από κανονικούς πληθυσμούς. Αν το μέγεθος των δειγμάτων είναι μεγάλο, τότε ο έλεγχος των στατιστικών υποθέσεων που αναφέρεται για τη διαφορά των δυο μέσων γίνεται όπως και στη περίπτωση που τα δείγματα προέρχονται από κανονικούς πληθυσμούς.

Έλεγχος για τη σύγκριση μέσων τιμών – Ζευγαρωτές παρατηρήσεις

Η έννοια των ζευγαρωτών παρατηρήσεων δόθηκε αρκετά αναλυτικά πιο πάνω. Σε αυτή την παράγραφο, θα περιγραφεί ο έλεγχος για τη σύγκριση των μέσων τιμών των πληθυσμών. Ισχύουν τα ακόλουθα: για το πρώτο δείγμα έχουμε τις

παρατηρήσεις $x_1, x_2, ..., x_n$ και για το δεύτερο δείγμα οι παρατηρήσεις είναι $y_1, y_2, ..., y_n$. Υπολογίζουμε τις διαφορές $d_i = x_i - y_i$, $i = 1, 2, ..., n$ που είναι ανεξάρτητες και μπορεί να υποτεθεί ότι αποτελούν ένα τυχαίο δείγμα από έναν κανονικό πληθυσμό με μέση τιμή d και διασπορά σ_d^2. Αν η μέση τιμή d είναι μηδέν, τότε τα δυο δείγματα μπορεί να θεωρηθούν ισοδύναμα. Αν d είναι θετικό, τότε η μέση τιμή του πρώτου δείγματος είναι μεγαλύτερη από τη μέση τιμή του δεύτερου δείγματος. Ενώ, αν d είναι αρνητικό η μέση τιμή του πρώτου είναι μικρότερη από του δεύτερου. Οπότε ο έλεγχος που δύναται να εφαρμοστεί είναι ισοδύναμος με τον έλεγχο μέσης τιμής που περιγράφτηκε πιο πάνω. Δηλαδή, έχουμε:

$$H_0 : d = \mu_1 - \mu_2 = d_0$$
$$H_1 : 1.\, d = \mu_1 - \mu_2 \neq d_0$$
$$2.\, d = \mu_1 - \mu_2 > d_0$$
$$3.\, d = \mu_1 - \mu_2 < d_0$$

Βρίσκουμε την ποσότητα $t = \dfrac{\bar{d} - d_0}{s_d / \sqrt{n}}$ και

1. απορρίπτουμε τη μηδενική υπόθεση όταν $|t| > t_{n-1, a/2}$

2. απορρίπτουμε τη μηδενική υπόθεση όταν $t > t_{n-1,a}$

3. απορρίπτουμε τη μηδενική υπόθεση όταν $t < t_{n-1,a}$

Έλεγχος υποθέσεων του λόγου των διασπορών δυο κανονικών πληθυσμών

Πολλές φορές από δυο κανονικούς πληθυσμούς παίρνουμε δυο δείγματα με τη βοήθεια των οποίων θέλουμε να ελεγχθεί η υπόθεση της ισότητας των δυο διασπορών. Οι υποθέσεις που αναφέρονται στην περίπτωση αυτή είναι:

$$H_0 : \frac{\sigma_1^2}{\sigma_2^2} = 1 \Leftrightarrow H_0 : \sigma_1^2 = \sigma_2^2$$

vs

$$1.H_1 : \frac{\sigma_1^2}{\sigma_2^2} \neq 1 \Leftrightarrow H_1 : \sigma_1^2 \neq \sigma_2^2$$

$$2.H_1 : \frac{\sigma_1^2}{\sigma_2^2} > 1 \Leftrightarrow H_1 : \sigma_1^2 > \sigma_2^2$$

$$3.H_1 : \frac{\sigma_1^2}{\sigma_2^2} < 1 \Leftrightarrow H_1 : \sigma_1^2 < \sigma_2^2$$

Στην περίπτωση που **οι μέσοι είναι γνωστοί** το κριτήριο είναι:

$$F = \frac{\frac{\sum (x_i - \mu_1)^2}{n_1 - 1}}{\frac{\sum (y_i - \mu_2)^2}{n_2 - 1}} = \frac{\sigma_1^2}{\sigma_2^2} \sim F_{n_1, n_2, a}$$

και

1. απορρίπτουμε τη μηδενική υπόθεση όταν $|F| > F_{n_1, n_2, a/2}$
2. απορρίπτουμε τη μηδενική υπόθεση όταν $F > F_{n_1, n_2, a}$
3. απορρίπτουμε τη μηδενική υπόθεση όταν $F < F_{n_1, n_2, a}$

ενώ αν **οι μέσοι είναι άγνωστοι** το κριτήριο είναι:

$$F = \frac{\frac{\sum (x_i - \overline{x})^2}{n_1 - 1}}{\frac{\sum (y_i - \overline{y})^2}{n_2 - 1}} = \frac{s_1^2}{s_2^2} \sim F_{n_1 - 1, n_2 - 1, a}$$

και

4. απορρίπτουμε τη μηδενική υπόθεση όταν $|F| > F_{n_1 - 1, n_2 - 1, a/2}$
5. απορρίπτουμε τη μηδενική υπόθεση όταν $F > F_{n_1 - 1, n_2 - 1, a}$
6. απορρίπτουμε τη μηδενική υπόθεση όταν $F < F_{n_1 - 1, n_2 - 1, a}$

Έλεγχος υποθέσεων για αναλογίες

Έστω ότι έχουμε *n* επαναλαμβανόμενα πειράματα στα οποία έχουμε ως αποτέλεσμα «επιτυχία» ή «αποτυχία». Αν ονομάσουμε το πλήθος των επιτυχιών *x* και *p* την πιθανότητα επιτυχίας σε *n* επαναλήψεις (*n*>30), τότε $\hat{p} = \frac{x}{n}$ θα είναι το ποσοστό των επιτυχιών στο δείγμα. Το $\hat{p} \sim N(p, \frac{p(1-p)}{n})$. Οι έλεγχοι που δύναται να εφαρμοστούν σε αυτήν την περίπτωση είναι:

1. $H_0 : p = p_0$ vs $H_1 : p \neq p_0$
2. $H_0 : p = p_0$ vs $H_1 : p > p_0$
3. $H_0 : p = p_0$ vs $H_1 : p < p_0$

Η μεταβλητή $Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \sim N(0,1)$ άρα για τους παραπάνω ελέγχους έχουμε:

1. απορρίπτουμε τη μηδενική υπόθεση όταν $|z| > z_{a/2}$
2. απορρίπτουμε τη μηδενική υπόθεση όταν $z > z_a$
3. απορρίπτουμε τη μηδενική υπόθεση όταν $z < -z_a$

Έλεγχος υποθέσεων για τη διαφορά δυο αναλογιών

Αν έχουμε δυο ανεξάρτητους διωνυμικούς πληθυσμούς, μπορούμε να έχουμε τους παρακάτω συνδυασμούς υποθέσεων:

$$1.H_0 : p_1 = p_2 \quad \text{vs} \quad H_1 : p_1 \neq p_2$$
$$2.H_0 : p_1 = p_2 \quad \text{vs} \quad H_1 : p_1 > p_2$$
$$3.H_0 : p_1 = p_2 \quad \text{vs} \quad H_1 : p_1 < p_2$$

Αν το μέγεθος των δειγμάτων μεγάλο, τότε εφαρμόζεται το κριτήριο:

$$Z = \frac{p_1 - p_2}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1} + \frac{1}{n_2})}}, \quad \text{όπου} \quad \hat{p} = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} \text{ άρα για τους}$$

παραπάνω ελέγχους έχουμε:

4. απορρίπτουμε τη μηδενική υπόθεση όταν $|z| > z_{a/2}$
5. απορρίπτουμε τη μηδενική υπόθεση όταν $z > z_a$
6. απορρίπτουμε τη μηδενική υπόθεση όταν $z < -z_a$

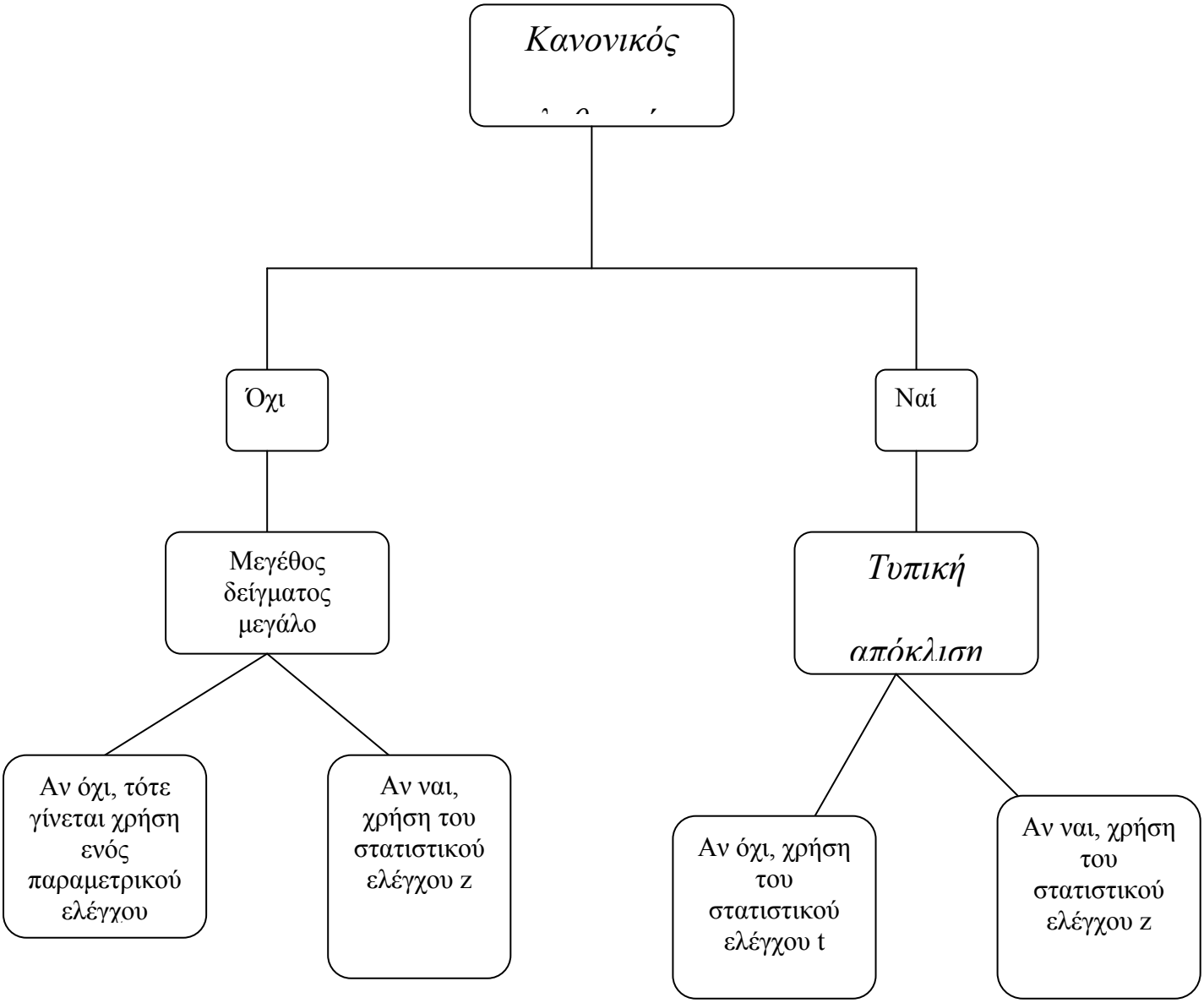
Σημείωση

Συγκρίνοντας τους τύπους των ελέγχων σημαντικότητας και τους αντίστοιχους των ΔΕ, συμπεραίνουμε ότι μπορούμε να

δεχτούμε τη μηδενική υπόθεση σε ε.σ α όταν η συγκεκριμένη τιμή που εξετάζεται ανήκει στο αντίστοιχο 100(1-α)%ΔΕ της εξεταζόμενης παραμέτρου του πληθυσμού. Ο τρόπος αυτός, επειδή μπορεί να ελέγξει συγχρόνως πολλές μηδενικές υποθέσεις θεωρείται αρκετά χρήσιμος στη στατιστική.

Διάγραμμα

Καθορισμός στατιστικού ελέγχου για τη μέση τιμή ενός πληθυσμού



Ασκήσεις

1. Οι υπεύθυνοι μιας αλυσίδας fast food ισχυρίζονται ότι ο μέσος χρόνος αναμονής των πελατών τους είναι 3 λεπτά. Προκειμένου το τμήμα ποιοτικού ελέγχου της επιχείρησης να πιστοποιήσει τον ισχυρισμό, παίρνει τυχαίο δείγμα 50 πελατών και σημειώνει τον χρόνο αναμονής κάθε πελάτη. Οι παρατηρήσεις φαίνονται στον ακόλουθο πίνακα:

4,56	2,95	2,02	3,09	3,13	1,80	2,78	4,52	1,01	2,14
3,02	3,98	3,09	3,19	3,69	4,17	2,82	1,61	0,96	3,24
5,07	3,74	1,40	3,17	3,13	2,18	1,25	3,28	1,44	1,39
3,49	2,64	1,23	3,07	3,21	2,98	3,50	1,96	2,18	3,18
2,36	3,93	3,03	2,06	2,28	3,04	2,34	2,51	1,73	2,64

Α. θα μπορούσατε να δεχθείτε ότι ο χρόνος αναμονής των πελατών είναι κανονικά κατανεμημένος;

Β. θα μπορούσατε να βγάλετε από αυτά τα δεδομένα το συμπέρασμα ότι ο μέσος χρόνος αναμονής των πελατών είναι λιγότερο από 3 λεπτά;

Γ. αν οι πέντε πρώτες στήλες είναι χρόνοι αναμονής πελατών που είναι άνδρες και οι πέντε επόμενες πελατών που είναι γυναίκες θα συμπεραίνετε το συμπέρασμα ότι ο μέσος χρόνος αναμονής των ανδρών διαφέρει από τον μέσο χρόνο αναμονής των γυναικών;

2. Ένα μεσιτικό γραφείο που ειδικεύεται στις πωλήσεις οικοπέδων έχει παρατηρήσει ότι κατά μέσο όρο τα οικόπεδα πωλούνται σε 90 μέρες από την στιγμή που θα περάσουν στη δικαιοδοσία του. Τελευταία έχει δημιουργηθεί η εντύπωση ότι τα οικόπεδα «παραμένουν» περισσότερο καιρό στο γραφείο. Για να ελεγχθεί αν συμβαίνει κάτι τέτοιο, παίρνουν ένα τυχαίο δείγμα 20 πρόσφατα πουλημένων οικοπέδων. Οι μέρες μετά από τις οποίες πουλήθηκαν αυτά ήταν:

98	133	91	138
62	99	97	125
99	109	99	87

59	93	111	94
83	107	134	107

3. Ο ιδιοκτήτης ενός ορυχείου ενδιαφέρεται να αξιολογήσει μια νέα μέθοδο παραγωγής συνθετικών διαμαντιών. Η μελέτη του κόστους που συνεπάγεται η διαδικασία κατασκευής, έχει οδηγήσει στο συμπέρασμα ότι για να είναι επικερδής η νέα αυτή μέθοδος θα πρέπει το μέσο βάρος των συνθετικών διαμαντιών να είναι γύρω στα 0.5 καράτια. Προκειμένου να αξιολογηθεί η διαδικασία κατασκευής επιλέγεται δείγμα από 6 συνθετικά διαμάντια που έχουν κατασκευαστεί με τη νέα μέθοδο παρασκευής. Το βάρος τους βρίσκεται ότι είναι: 0.46, 0.61, 0.52, 0.48, 0.57 και 0.54 καράτια αντίστοιχα. Να καθοριστεί σε ε.σ 5% με βάση τις πληροφορίες από το δείγμα αυτό αν η νέα μέθοδος είναι επικερδής.
4. Μια βιομηχανία παρασκευής και συσκευασίας καφέ χρησιμοποιεί αεροστεγείς συσκευασίες που περιέχουν 368

γραμμάρια καφέ. Όπως είναι φυσικό δεν είναι δυνατόν να επιτυγχάνεται πάντοτε συσκευασία που να περιέχει ακριβώς το περιεχόμενο αυτό. Ο υπεύθυνος της συσκευασίας προκειμένου να ελέγξει το κατά πόσο η επιδίωξη αυτή επιτυγχάνεται, επιλέγει ένα τυχαίο δείγμα 25 πακέτων που έχουν συσκευαστεί με αυτόν τον τρόπο. Μετρώντας το περιεχόμενο στις συσκευασίες αυτές διαπιστώνεται ότι η μέση ποσότητα καφέ που περιέχεται στις συσκευασίες αυτές παρέχονται στον παρακάτω πίνακα. Με βάση το τυχαίο αυτό δείγμα σε τι συμπέρασμα μπορεί να καταλήξει ο προϊστάμενος της εταιρείας όσο αφορά την επιδίωξή του;

346,807	353,706	357,446	323,653	361,489
335,197	342,793	342,306	354,780	346,435
350,148	333,668	359,834	352,557	329,726
327,246	367,542	354,729	326,554	344,710
337,528	324,331	335,434	351,072	342,789

5. Προκειμένου να γίνει σύγκριση της ποιότητας δυο ειδών λαδιών αυτοκινήτου μια εταιρεία προστασίας καταναλωτών ενδιαφέρεται να κατασκευάσει ένα 95% ΔΕ για τη διαφορά της μέσης κατανάλωσης βενζίνης (μετρούμενης σε km/lt) με τη χρησιμοποίηση των δυο διαφορετικών ειδών λαδιών. Για το λόγο αυτό χρησιμοποιούνται τέσσερα αυτοκίνητα τα οποία δοκιμάζονται σε απόσταση 1000km την πρώτη φορά χρησιμοποιώντας το λάδι μηχανής τύπου Α και τη δεύτερη φορά χρησιμοποιώντας το λάδι μηχανής τύπου Β. Οι μετρήσεις στις οποίες η εταιρεία κατέληξε δίνονται στον πίνακα που ακολουθεί:

	αυτοκίνητα			
	1	2	3	4
Λάδι Α	19,77	18,90	20,20	16,29
Λάδι Β	18,91	18,21	18,84	16,92

Επίσης να γίνει έλεγχος της ισότητας των μέσων.

6. Μια εταιρεία που κατασκευάζει πυρίτιδα έχει δημιουργήσει ένα νέο είδος πυρίτιδας της οποίας την ποιότητα θέλει να ελέγξει. Επιλέγοντας τυχαία 8 οβίδες που έχουν κατασκευαστεί με την πυρίτιδα αυτή η εταιρεία μετρά τις ταχύτητες των βλημάτων αυτών σε m/sec και βρίσκει τα εξής αποτελέσματα: 3005 2925 2935 2965 2995 3005 2937 2905. να βρεθεί ένα 95% ΔΕ για την πραγματική μέση ταχύτητα των οβίδων που κατασκευάζονται με την ταχύτητα. Να ελεγχθεί εάν η ταχύτητα των βλημάτων είναι 4000. Αρχικά, να γίνει έλεγχος κανονικότητας για την ταχύτητα των βλημάτων.
7. Έστω ότι ένας ερευνητής αγοράς ενδιαφέρεται να μελετήσει την επίδραση που έχει στις πωλήσεις ενός προϊόντος το μέρος στο οποίο το προϊόν τοποθετείται μέσα στο κατάστημα. Συγκεκριμένα ενδιαφέρεται να εξετάσει αν υπάρχει διαφορά στις πωλήσεις αν το προϊόν αυτό τοποθετείται κοντά στην έξοδο ή σε άλλο μέρος του καταστήματος δίπλα σε άλλα παρόμοια προϊόντα. Για να εξετάσει το πρόβλημα αυτό ο ερευνητής επιλέγει ένα

τυχαίο δείγμα 13 καταστημάτων ίδιου μεγέθους από τη συγκεκριμένη αλυσίδα καταστημάτων και σε 7 από αυτά τοποθετεί το προϊόν κοντά στην έξοδο ενώ στα υπόλοιπα 6 στο σημείο που έχει και άλλα παρόμοια προϊόντα. Τα αποτελέσματα των πωλήσεων, σε αριθμό πακέτων του προϊόντος που πωλούνται ανά βδομάδα δίνονται στον ακόλουθο πίνακα.

ΕΒΔΟΜΑΔΙΑΙΕΣ ΠΩΛΗΣΕΙΣ (σε αριθμό πακέτων)	
Τοποθέτηση στην έξοδο Χ	Τοποθέτηση στο εξωτερικό Υ
107	90
153	83
52	86
158	94
141	89
87	93
119	
$\overline{x} = 121$ $s_1^2 = 945$ $n_1 = 7$	$\overline{y} = 89,17$ $s_2^2 = 17,37$ $n_2 = 6$

Ας υποθέσουμε ότι οι πωλήσεις ανά βδομάδα του προϊόντος αυτού ακολουθούν την κανονική κατανομή. Δοθέντος δε ότι δεν έχουμε κάποια ένδειξη ότι οι διασπορές των δυο πληθυσμών είναι ίσες ας θεωρήσουμε την περίπτωση των άνισων διασπορών. Να ελεγχθεί η υπόθεση της ισότητας των μέσων πωλήσεων σε ε.σ. 5% και να κατασκευαστεί το αντίστοιχο ΔΕ.

8. Μια εταιρεία παρασκευής και συσκευασίας καφέ έχει δυο εργοστάσια συσκευασίας σε δυο διαφορετικές περιοχές. Στον έλεγχο ποιότητας της συσκευασίας αν διαπιστωθεί ότι η ποσότητα που τοποθετείται στα κουτιά αποκλίνει πέρα από κάποιο συγκεκριμένο σημείο τότε η εταιρεία απορρίπτει τα κουτιά αυτά. Έστω ότι ο υπεύθυνος του ελέγχου ποιότητας της εταιρείας ενδιαφέρεται να διαπιστώσει εάν υπάρχει στατιστικά σημαντική διαφορά στην αναλογία των κουτιών που χρειάζεται να απορριφθούν στα δυο εργοστάσια. Προκειμένου να γίνει αυτό επιλέγονται τυχαία δυο δείγματα από 200 κουτιά από κάθε εργοστάσιο από τα οποία προκύπτει ότι 19 από το

πρώτο και 21 από το δεύτερο χρειάζεται να απορριφθούν. Να ελεγχθεί η υπόθεση ότι η διαφορά των αναλογιών στα κουτιά που προέρχονται από τα δυο εργοστάσια και που πρέπει να απορριφθούν δεν είναι στατιστικά σημαντική.

9. Μια εταιρεία ενδιαφέρεται να υιοθετήσει μια νέα πολιτική για τους πωλητές της σύμφωνα με την οποία κάθε πωλήτης θα παίρνει κάθε μήνα ένα δώρο ανάλογα με τις πωλήσεις που θα πραγματοποιήσει. Η διεύθυνση της εταιρείας προκειμένου να διερευνήσει τις προσδοκίες των ανδρών και γυναικών εργαζομένων στην εταιρεία επιλέγει δυο τυχαία δείγματα από 40 άνδρες και 40 γυναίκες από τους οποίους ζητά να προβλέψουν τις πρόσθετες μηνιαίες αποδοχές που θα έχουν αν υιοθετηθεί το προτεινόμενο σχήμα. Τα δειγματικά δεδομένα έδωσαν τις εξής πληροφορίες: $\overline{x} = 31083, s_1 = 2312$ για τους άνδρες και για τις γυναίκες $\overline{y} = 29745, s_2 = 2569$. Παρέχουν τα δεδομένα αυτά ενδείξεις ότι υπάρχει διαφορά στο μέσο αναμενόμενο πρόσθετο εισόδημα μεταξύ ανδρών και γυναικών πωλητών με την υιοθέτηση της νέας πολιτικής; (α=0.05)

10. Μια εταιρεία δημοσκοπήσεων ενδιαφέρεται να εκτιμήσει το ποσοστό των ψηφοφόρων που προτιμούν ένα συγκεκριμένο υποψήφιο. Για το λόγο αυτό επιλέγει ένα τυχαίο δείγμα 100 ψηφοφόρων και βρίσκει ότι 55 από αυτούς προτιμούν τον συγκεκριμένο υποψήφιο. Τι μπορούμε να πούμε για την πιθανότητα του υποψηφίου αυτού να κερδίσει τις εκλογές με βάση το δείγμα αυτό;

11. Προκειμένου να διαμορφώσει τη στρατηγική της προεκλογικής του εκστρατείας ένας υποψήφιος επιθυμεί να εκτιμήσει τη διαφορά στην απήχηση που έχει στο εκλογικό σώμα μεταξύ ανδρών και γυναικών ψηφοφόρων. Προκειμένου να εκτιμήσει τη διαφορά αυτή αποφασίζει να κατασκευάσει ένα 99% ΔΕ για τη διαφορά $p_1 - p_2$ όπου p_1 η αναλογία των γυναικών του πληθυσμού που τον υποστηρίζουν και p_2 η αναλογία των ανδρών του πληθυσμού που τον υποστηρίζουν. Προκειμένου να προχωρήσει στη μελέτη αυτή ο υποψήφιος επιλέγει ένα

τυχαίο δείγμα 1000 ψηφοφόρων από κάθε μια κατηγορία ανδρών και γυναικών. Μεταξύ των ανδρών βρίσκει ότι 388 θα τον υποστήριζαν και μεταξύ των γυναικών 459 θα τον υποστήριζαν. Να καθορισθεί το 99% ΔΕ της διαφοράς των αναλογιών.

12. Μια εταιρεία εξετάζει εάν ο μισθός των νοσοκόμων είναι ψηλότερος από τον μισθό των δασκάλων. Για τη μελέτη αυτή συλλέχθηκε το ακόλουθο δείγμα. Είναι λογικό να καταλήξουμε στο συμπέρασμα ότι ο μισθός των νοσοκόμων είναι ψηλότερος από των δασκάλων; (α=0.01). Ποια είναι η τιμή του $p-value$;

Δασκάλες(ευρώ)	545	526	527	484	509	502	520	529	530	542	532
Νοσοκόμες(ευρώ)	541	590	521	471	550	559	525	529			

13. Σε μια χημική αντίδραση είναι αναγκαίο να χρησιμοποιήσουμε ένα διάλυμα με δείκτη ph ίσο με 8.30. Μια μέθοδος προσδιορισμού του δείκτη ph δίνει μετρήσεις που

προσσεγγιστικά ακολουθούν την $N(\mu,(0.02)^2)$. Έξι μετρήσεις του δείκτη ph έδωσαν τις τιμές 8.34, 8.29, 8.30, 8.31, 8.30, 8.32 για ένα συγκεκριμένο διάλυμα

I. μπορούμε να ισχυριστούμε σε ε.σ 5% ότι ο δείκτης ph είναι 8.30;

II. Αν θέλουμε να είμαστε σίγουροι 95% ότι η αρχική υπόθεση δεν γίνεται δεκτή, όταν ο πραγματικός δείκτης είναι >8.33 ή <8.27 τότε πόσες μετρήσεις θα πρέπει να γίνουν;

III.να βρεθεί το μέγεθος του σφάλματος δεύτερου είδους που διαπράττουμε στην περίπτωση I όταν μ=8.32, μ=8.26 και μ=8.31

14. Μια μελέτη διεξήχθη για να διερευνήσει εάν η βρώμη βοηθάει να ελαττωθεί το επίπεδο χοληστερόλης στον ορό σε άνδρες με υψηλή χοληστερόλη. Επελέγησαν τυχαία 14 τέτοιοι άνδρες και τοποθετήθηκαν σε μια από δυο δίαιτες. Η μια περιλάμβανε πρωινό με βρώμη και η άλλη με αραβόσιτο. Μετά δυο βδομάδες με αυτή τη δίαιτα κατεγράφησαν τα

χαμηλής πυκνότητας λιποπρωτείνης(LDL) επίπεδα χοληστερόλης για αυτά τα άτομα. Κατόπιν κάθε άνδρας ακολούθησε την άλλη δίαιτα. Μετά από άλλες δυο βδομάδες με τη νέα δίαιτα κατεγράφησαν και πάλι τα LDL επίπεδα χοληστερόλης για αυτά τα άτομα. Τα δεδομένα αυτής της μελέτης δίνονται παρακάτω.

LDL(mmol/l)

Άτομο	Αραβόσιτος	Βρώμη
1	4,61	3,84
2	6,42	5,57
3	5,40	5,85
4	4,54	4,80
5	3,98	3,68
6	3,82	2,96
7	5,01	4,41
8	4,34	3,72
9	3,80	3,49

10	4,56	3,84
11	5,35	5,26
12	3,89	3,73
13	2,25	1,84
14	4,24	4,14

Να διεξαχθεί ο κατάλληλος στατιστικός έλεγχος για α=0.05 και για α=0.02. Ποια είναι τα συμπεράσματά σας;

15. Τα ακόλουθα δεδομένα προέρχονται από μια μελέτη που εξετάζει την αποτελεσματικότητα της κοτινίνης στο σίελο ως δείκτη έκθεσης σε καπνό τσιγάρου. Σε ένα μέρος της μελέτης, από 7 άτομα, από τα οποία κανένα δεν ήταν καπνιστής που καπνίζει πολύ και όλοι απείχαν από το κάπνισμα για τουλάχιστον μια βδομάδα προ της έναρξης της μελέτης, ζητείται να καπνίσουν ένα και μόνο τσιγάρο. Δείγματα σιέλου λαμβάνονται από όλα τα άτομα 2,12, 24 και 48 ώρες αφού κάπνισαν αυτό το μοναδικό τσιγάρο. Στον πίνακα παρακάτω

παρουσιάζονται τα επίπεδα κοτινίνης στις 12 ώρες και στις 24 ώρες. Πιστεύετε ότι κατά μέσο όρο τα επίπεδα κοτινίνης μετά από 24 ώρες πρέπει να είναι μικρότερα από τα αντίστοιχα μετά 12 ώρες; Να κατασκευαστεί ένα 95% μονόπλευρο ΔΕ για την αληθινή διαφορά στους μέσους των πληθυσμών και να ελεγχθεί η μηδενική υπόθεση ότι οι μέσοι ταυτίζονται σε $\alpha=0.05$.

ΣΥΝΟΠΤΙΚΕΣ ΣΗΜΕΙΩΣΕΙΣ

ΓΙΑ ΠΑΛΙΝΔΡΟΜΗΣΗ

Εισαγωγή

Μέχρι τώρα ασχοληθήκαμε με τη μελέτη μιας μόνο μεταβλητής, δηλαδή μελετήσαμε τις στατιστικές μονάδες ενός πληθυσμού ως προς ένα μόνο χαρακτηριστικό τους.

Συχνά όμως, εμφανίζονται περιπτώσεις που ασχολούμαστε με την ταυτόχρονη μελέτη δυο ή περισσότερων μεταβλητών γιατί οι αποφάσεις στηρίζονται στην πιθανή σχέση ανάμεσά τους.

Τέτοια παραδείγματα είναι τα ακόλουθα:

- Το ύψος της μετοχής μιας εταιρείας και η οικονομική βάση αυτής
- Το βάρος και ύψος ενός ανθρώπου
- Αριθμός πωλήσεων και αριθμός δαπανών για διαφήμιση
- Αριθμός πωλήσεων και αριθμός πωλητών

Σε αυτές τις περιπτώσεις τα προβλήματα που μας απασχολούν είναι:

1. Αν υπάρχει κάποια σχέση μεταξύ των δυο μεταβλητών
2. Πως μπορούν να προβλέψουμε τις τιμές της μιας γνωρίζοντας τις τιμές της άλλης μεταβλητής

Το μεν πρώτο ερώτημα απαντιέται με τη βοήθεια της ανάλυσης της συσχέτισης που χρησιμοποιείται προκειμένου να διαπιστωθεί αν υπάρχει στατιστική σχέση μεταξύ των δυο μεταβλητών.

Αν όντως διαπιστωθεί ότι «οι δυο μεταβλητές συσχετίζονται» τότε χρησιμοποιούμε την ανάλυση παλινδρόμησης για την απάντηση του δεύτερου ερωτήματος δηλ θα εκτιμηθεί ένα μοντέλο που θα περιγράφει τη σχέση των δυο μεταβλητών.

Σε αυτές τις περιπτώσεις θεωρούμε δυο μεταβλητές: X και Y, όπου

X καλείται ανεξάρτητη μεταβλητή (independent variable)

Y καλείται εξαρτημένη μεταβλητή (dependent variable)

Έτσι τώρα οι παρατηρήσεις εμφανίζονται σε ζεύγη τιμών $(x_1, y_1),...,(x_n, y_n)$.

Το πρώτο βήμα είναι να γίνει μια γραφική παράσταση των δεδομένων έτσι ώστε να πάρουμε μια πρώτη εικόνα της σχέσης που υπάρχει ανάμεσα στις δυο μεταβλητές.

Τα ζεύγη των παρατηρήσεων παριστάνονται σε ένα ορθοκανονικό σύστημα συντεταγμένων, όπου στον άξονα x παριστάνονται οι τιμές της ανεξάρτητης μεταβλητής και στον άξονα y της εξαρτημένης. Οπότε σχηματίζεται ένα πλήθος σημείων που καλείται **νέφος σημείων ή διάγραμμα διασποράς.**

Πρέπει να διευκρινιστεί ότι με τη μελέτη αυτή θέλουμε να εξηγήσουμε τις μεταβολές της εξαρτημένης μεταβλητής. Ενώ η ανεξάρτητη μεταβλητή πιστεύουμε ότι επιδρά στην εξαρτημένη και προκαλεί τις μεταβολές και επομένως χρησιμοποιείται για να ερμηνευτεί η μεταβλητότητα που παρουσιάζει η εξαρτημένη μεταβλητή.

Η απλούστερη σχέση που μπορεί να συνδέει δυο μεταβλητές είναι η γραμμική και θα ασχοληθούμε μόνο σε αυτήν την περίπτωση, αφού πολλές άλλες μορφές σχέσεων μπορούν εύκολα με κάποιους κατάλληλους μετασχηματισμούς των μεταβλητών να αναχθούν σε γραμμικές.

Συναρτήσεις που ανάγονται σε γραμμικές

$y' = \exp(a + bx')$ θέτοντας $y = \ln y' \Rightarrow y = a + bx$

$y' = a'x'^{b}$ θέτοντας $y = \ln y', \ln a' = a, \ln x' = x \Rightarrow y = a + bx$

$z = c\exp(bx)$ θέτοντας $y = \ln z, a = \ln c \Rightarrow y = a + bx$

$y = a + b\frac{1}{z}$ θέτοντας $x = \frac{1}{z} \Rightarrow y = a + bx$

$z = \frac{1}{a + bx}$ θέτοντας $y = \frac{1}{z} \Rightarrow y = a + bx$

$z = \frac{1}{(a + bx)^2}$ θέτοντας $y = \frac{1}{z} \Rightarrow y = a + bx$

$\frac{1}{z} = a + b\frac{1}{1 + x'}$ θέτοντας $y = \frac{1}{z}, x = \frac{1}{1 + x'} \Rightarrow y = a + bx$

$y = a + b\sqrt{x'}$ θέτοντας $x = \sqrt{x'} \Rightarrow y = a + bx$

Συντελεστής Συσχέτισης

Όπως ήδη έχει αναφερθεί, το πρώτο βήμα στη μελέτη της παλινδρόμησης είναι η κατασκευή του διαγράμματος διασποράς. Στην συνέχεια η μελέτη συνεχίζεται με τον υπολογισμό του **συντελεστή συσχέτισης**.

Συντελεστής Συσχέτισης καλείται η ποσοτική μέτρηση της έντασης της (γραμμικής) σχέσης μεταξύ δυο μεταβλητών. Ο συντελεστής αυτός καλείται γραμμικός συντελεστής συσχέτισης και συμβολίζεται με ρ ενώ η αντίστοιχη εκτίμησή του με r.

Υπολογίζεται:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{[\sum_{i=1}^{n}(x_i - \bar{x})^2][\sum_{i=1}^{n}(y_i - \bar{y})^2]}} = \frac{n\sum_{i=1}^{n}xy - \sum_{i=1}^{n}x\sum_{i=1}^{n}y}{\sqrt{[n\sum_{i=1}^{n}x^2 - (\sum_{i=1}^{n}x)^2(\sum_{i=1}^{n}y)^2}} = \frac{\mathrm{cov}(X,Y)}{\sigma_X\sigma_Y}$$

Όπου

$$\mathrm{cov}(X,Y) = E(X - \bar{X})(Y - \bar{Y}) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n}\sum_{i=1}^{n}xy - \overline{x}\,\overline{y}$$

Και καλείται συνδιακύμανση των δυο μεταβλητών.

Ιδιότητες συνδικύμανσης

$1.\mathrm{cov}(X,X) = V(X)$
$2.\mathrm{cov}(a + bX, c + dY) = bd\,\mathrm{cov}(X,Y)$
$3.\mathrm{cov}(X,Y) = \mathrm{cov}(Y,X)$
$4.\mathrm{cov}(X + Y, Z) = \mathrm{cov}(X,Z) + \mathrm{cov}(Y,Z)$
$5.V(X \pm Y) = V(X) + V(Y) \pm 2\,\mathrm{cov}(X,Y)$
$6.\mathrm{cov}(X,-Y) = -\mathrm{cov}(X,Y)$

Όταν η συνδιακύμανση (συνδιασπορά) είναι θετικός αριθμός τότε οι μεταβλητές είναι θετικά συσχετισμένες και μεταβάλλονται ομόρροπα.

Αν η συνδιακύμανση είναι αρνητικός αριθμός τότε οι μεταβλητές είναι αρνητικά συσχετισμένες και μεταβάλλονται αντίρροπα.

Αν η συνδιακύμανση είναι ίση με μηδέν τότε δεν υπάρχει γραμμική συμμεταβολή των δυο μεταβλητών. Στην περίπτωση αυτή λέμε ότι οι δυο μεταβλητές είναι γραμμικά ασυσχέτιστες.

Η συνδιακύμανση εξαρτάται από τις μονάδες μέτρησης των δυο μεταβλητών δηλαδή η μονάδα μέτρησής της είναι ίση με το γινόμενο των μονάδων μέτρησης των δυο μεταβλητών. Δεν μπορεί να εκφράσει με αντικειμενικότητα το βαθμό της γραμμικής συμμεταβολής, ούτε όμως και να χρησιμοποιηθεί για τη σύγκριση του βαθμού γραμμικής συμμεταβολής διαφορετικών κατανομών.

Γι' αυτό χρησιμοποιούμε ως μέτρο της γραμμικής συμμεταβολής των δυο μεταβλητών τον γραμμικό συντελεστή συσχέτισης που είναι καθαρός αριθμός.

Ιδιότητες ρ

1. Ο γραμμικός συντελεστής συσχέτισης έχει πάντοτε το ίδιο πρόσημο με την συνδιασπορά
2. Παίρνει τιμές μεταξύ -1 και 1 δηλ. $-1 \le \rho \le 1$
3. Αν $\rho = 1$ ή $\rho = -1$ τότε οι μεταβλητές έχουν τέλεια θετική ή αρνητική αντίστοιχα, γραμμική σχέση

4. Είναι καθαρός αριθμός
5. Αν $\rho = 0$ οι μεταβλητές είναι γραμμικά ασυσχέτιστες μπορεί να έχουν οποιαδήποτε άλλου είδους σχέση.

Επίσης

- Αν $|\rho| \le 0.30$ δεν υπάρχει γραμμική συσχέτιση
- Αν $0.30 \le |\rho| \le 0.50$, ασθενής συσχέτιση
- Αν $0.50 \le |\rho| \le 0.70$, μέτρια συσχέτιση
- Αν $0.70 \le |\rho| \le 0.80$, ισχυρή συσχέτιση
- Αν $|\rho| \ge 0.80$, πολύ ισχυρή συσχέτιση
- Αν $|\rho| = 1$, τέλεια συσχέτιση