ΚΕΦΑΛΑΙΟ 10

ΔΕΙΓΜΑΤΙΚΕΣ ΚΑΤΑΝΟΜΕΣ

(Sampling Distributions)

Ένα χαρακτηριστικό των επιστημονικών μελετών στις οποίες απαιτείται η χρήση των διαδικασιών της Στατιστικής Συμπερασματολογίας είναι η ύπαρξη *τυχαιότητας* ή *δειγματοληπτικής μεταβλητότητας*. Ο όρος αυτός χρησιμοποιείται για να αποδώσει το γεγονός ότι επαναληπτική δειγματοληψία από ένα πληθυσμό οδηγεί σε δείγματα διαφορετικής σύνθεσης. Ένα κλασικό παράδειγμα στην περιοχή των τυχερών παιχνιδιών είναι το εξής: Τα χαρακτηριστικά μίας δεσμίδας τραπουλόχαρτων και ο μηχανισμός μοιράσματος των χαρτιών μπορεί να είναι γνωστά, αλλά η σύνθεση μιας συγκεκριμένης μοιρασιάς μεταβάλλεται μ' έναν τρόπο ο οποίος καθιστά την πρόβλεψη της σύνθεσής της αδύνατη.

Στα επιστημονικά πειράματα, η δειγματοληπτική μεταβλητότητα τείνει να παραλλάσσει τα χαρακτηριστικά του πληθυσμού τα οποία ενδιαφέρουν τον ερευνητή. Ένας στόχος κεντρικής σημασίας για την επαγωγική στατιστική είναι να διερευνήσει αν οποιαδήποτε διαφορά μεταξύ του θεωρητικού μοντέλου και των δεδομένων μπορεί να εξηγηθεί ή να αποδοθεί σε δειγματοληπτική μεταβλητότητα και, γενικότερα, να ποσοτικοποιήσει την αβεβαιότητα που εισάγει η δειγματοληπτική μεταβλητότητα.

Το πρώτο βήμα είναι να προσδιορισθεί η μορφή της ανεξήγητης μεταβλητότητας που παρατηρείται στην μέτρηση των χαρακτηριστικών των μελών του πληθυσμού ή στα αποτελέσματα της πειραματικής διαδικασίας. Αυτό, όπως έχουμε δει, μπορεί να γίνει στο πλαίσιο του στατιστικού μοντέλου μέσω του ορισμού της κατανομής συχνότητας ή της κατανομής πιθανότητας.

Το δεύτερο βήμα είναι ο προσδιορισμός της σχέσης μεταξύ αυτής της περιγραφής της μεταβλητότητας στον πληθυσμό και του σχήματος της αναμενόμενης μεταβλητότητας στο δείγμα. Λόγω των επιδράσεων της μεταβλητότητας που οφείλεται στην τυχαιότητα, η μέτρηση του χαρακτηριστικού που μας ενδιαφέρει σε μια μονάδα του

δείγματος η οποία έχει επιλεγεί από τον πληθυσμό δεν μπορεί, εν γένει, να προβλεφθεί. Παρ' όλα αυτά, είναι δυνατόν να ποσοτικοποιηθούν οι σχετικές πιθανότητες των διαφορετικών δυνατών αποτελεσμάτων. Αυτό οδηγεί στον ορισμό της *δειγματικής κατανομής (sampling distribution)* και της *δειγματικής κατανομής δείγματος (sampling distribution of a sample)*.

Το τελευταίο βήμα είναι η κατασκευή της *δειγματικής κατανομής μιας στατιστικής συνάρτησης δείγματος (sampling distribution of a statistic)*, η οποία αποτελεί την σύνδεση μεταξύ των θεωρητικών στατιστικών αποτελεσμάτων και της επιστημονικής ερμηνείας αυτών.

Η Έννοια μιας Δειγματικής Κατανομής

Μία προϋπόθεση της εφαρμογής της επαγωγικής στατιστικής είναι η λήψη δείγματος από τον πληθυσμό που ενδιαφέρει να μελετήσουμε τα στοιχεία του οποίου θα χρησιμοποιηθούν για την στατιστική ανάλυση. Τα χαρακτηριστικά των δεδομένων του δείγματος προσδιορίζονται εν μέρει από τα χαρακτηριστικά του πληθυσμού και εν μέρει από την μέθοδο δειγματοληψίας. Ο συνδυασμός της συνεισφοράς των δύο αυτών τύπων χαρακτηριστικών προσδιορίζεται από μία *δειγματική κατανομή*. Για μία κατηγορική ή διακριτή μεταβλητή, αυτή η κατανομή ορίζει την πιθανότητα με την οποία η τυχαία μεταβλητή που περιγράφει τον πληθυσμό θα έχει μία συγκεκριμένη τιμή για μία συγκεκριμένη μονάδα του δείγματος. Έτσι, αν υπάρχουν k δυνατές τιμές, η δειγματική κατανομή θα είναι η εξής:

Κατηγορία:	1	2	…	k
Πιθανότητα:	π_1	π_2	…	π_k

Αν η μεταβλητή Y είναι συνεχής, τότε υπάρχει μια δειγματική κατανομή με συνάρτηση πυκνότητας πιθανότητας π(.), η οποία μπορεί να χρησιμοποιηθεί για τον ορισμό της πιθανότητας με την οποία η μεταβλητή Y θα πάρει μία τιμή μικρότερη ή ίση της τιμής y_0 για μία συγκεκριμένη μονάδα του δείγματος, σύμφωνα με τον τύπο

$$P(Y \leq y_0) = \int_{-\infty}^{y_0} \pi(y)dy$$

Ας υποθέσουμε, για παράδειγμα, ότι ένας φοιτητής επιλέγεται από τον πληθυσμό των φοιτητών ενός πανεπιστημίου. Η πιθανότητα ότι ο φοιτητής αυτός προέρχεται από ένα συγκεκριμένο τμήμα του πανεπιστημίου εξαρτάται από την κατανομή συχνότητας του πληθυσμού των φοιτητών στα διάφορα τμήματα και από τον τρόπο με τον οποίο η επιλογή του φοιτητή έγινε. Αν η επιλογή έγινε μεταξύ των φοιτητών ενός Τμήματος Α, τότε η πιθανότητα να έχει επιλεγεί ένας φοιτητής από ένα Τμήμα Β είναι 0. Εναλλακτικά, αν ακολουθήθηκε τυχαία δειγματοληψία με βάση δειγματοληπτικό πλαίσιο που καλύπτει ολόκληρο τον φοιτητικό πληθυσμό του συγκεκριμένου πανεπιστημίου, η πιθανότητα να επιλεγεί ένας φοιτητής από το Τμήμα Β είναι ακριβώς ίση με το ποσοστό των φοιτητών του Πανεπιστημίου αυτού που είναι εγγεγραμμένοι στο Τμήμα Β.

Αυτή η συγκεκριμένη ιδιότητα της τυχαίας δειγματοληψίας, δηλαδή η εξίσωση της δειγματικής κατανομής με την κατανομή συχνότητας του πληθυσμού από τον οποίο λαμβάνεται το δείγμα, είναι εκείνη η οποία καθιστά την τυχαία δειγματοληψία τόσο σημαντική.

Δειγματοληπτική Κατανομή Δείγματος

Για τις πρακτικές εφαρμογές, είναι αναγκαίο να επεκταθεί η ιδέα της δειγματικής κατανομής στην *δειγματική κατανομή ενός δείγματος*. Αυτό απαιτεί τον προσδιορισμό μιας δειγματικής κατανομής με συνάρτηση (πυκνότητας) πιθανότητας π_s, η οποία αντιστοιχεί μία αριθμητική τιμή σε κάθε δυνατό αποτέλεσμα οριζόμενο ως μία ακολουθία παρατηρήσεων $(y_1, y_2, \ldots, y_n)$ συνδεομένων με τις μονάδες ενός δείγματος.

Στον πίνακα που ακολουθεί δίνεται η δειγματική κατανομή ενός δείγματος μεγέθους 3 για μία κατηγορική μεταβλητή, η οποία μπορεί να πάρει μόνο δύο δυνατές τιμές Ε (ελαττωματικό) και $\not{E}$ (μη ελαττωματικό) στο πλαίσιο ενός προβλήματος ποιοτικού ελέγχου των μονάδων ενός προϊόντος, η γραμμή παραγωγής του οποίου παρουσιάζει μία συχνότητα ελαττωματικών μονάδων ίση με 1%.

Πίνακας: Παράδειγμα δειγματικής κατανομής ενός δείγματος τριών μονάδων από την γραμμή παραγωγής ενός προϊόντος, όπου τα αποτελέσματα είναι Ε (ελαττωματικό) ή Έ (μη ελαττωματικό) και τα ενδεχόμενα είναι ανεξάρτητα.

Δείγμα	ΕΕΕ	ΕΕΈ	ΕΈΕ	ΈΕΕ
Πιθανότητα	0.000001	.000099	0.000099	0.000099

Δείγμα	ΕΈΈ	ΈΕΈ	ΈΈΕ	ΈΈΈ
Πιθανότητα	0.009801	0.009801	0.009801	0.970299

Με την προϋπόθεση ότι θα υιοθετηθεί μία αξιόπιστη μέθοδος δειγματοληψίας, η πιθανότητα με την οποία μία μονάδα που επιλέγεται θα είναι ελαττωματική είναι 0.01 και η πιθανότητα με την οποία αυτή θα είναι μη ελαττωματική είναι 0.99.

Δειγματική Κατανομή Στατιστικής Συνάρτησης

Όπως έχει ήδη λεχθεί, η στατιστική συμπερασματολογία είναι η διαδικασία συναγωγής συμπερασμάτων για ένα πληθυσμό με βάση τις πληροφορίες οι οποίες περιέχονται σε ένα δείγμα από τον πληθυσμό αυτό. Επειδή οι πληροφορίες σχετικά με τους πληθυσμούς περιγράφονται συνήθως μέσω των παραμέτρων των πληθυσμών, οι στατιστικές τεχνικές που χρησιμοποιούνται συνίστανται στην συναγωγή συμπερασμάτων για τις παραμέτρους του πληθυσμού με βάση κατάλληλες στατιστικές συναρτήσεις. (Υπενθυμίζεται ότι η παράμετρος ενός πληθυσμού είναι μία συνοπτική μέτρηση για τον πληθυσμό και η στατιστική συνάρτηση είναι μία συνοπτική μέτρηση για το δείγμα). Προκειμένου λοιπόν να μελετηθεί ο συνολικός πληθυσμός ως προς το χαρακτηριστικό που η συγκεκριμένη παράμετρος εκφράζει, επιλέγεται ένα τυχαίο δείγμα και υπολογίζεται η τιμή της αντίστοιχης στατιστικής συνάρτησης. Παρά το ότι η τιμή αυτής της στατιστικής συνάρτησης και η τιμή της αντίστοιχης παραμέτρου του πληθυσμού έχουν ελάχιστη πιθανότητα να συμπίπτουν, περιμένουμε ότι δεν θα διαφέρουν πάρα πολύ. Χρειαζόμαστε, επομένως, να μπορούμε να μετρήσουμε πόσο κοντά είναι πιθανό να βρίσκεται η τιμή της στατιστικής συνάρτησης με

αυτήν της παραμέτρου του πληθυσμού. Η δειγματική κατανομή της στατιστικής συνάρτησης μας δίνει αυτή την δυνατότητα. Για τον λόγο αυτό, η δειγματική κατανομή μιας στατιστικής συνάρτησης παίζει σημαντικό ρόλο στην στατιστική, γιατί το μέτρο της εγγύτητας που παρέχει είναι κεντρικής σημασίας για την στατιστική συμπερασματολογία. Επανερχόμενοι στο προηγούμενο παράδειγμα των μονάδων του προϊόντος που παράγονται από την συγκεκριμένη γραμμή παραγωγής, μία δυνατή στατιστική συνάρτηση θα μπορούσε να είναι ο αριθμός Τ των ελαττωματικών μονάδων στο δείγμα. Η στατιστική αυτή συνάρτηση θα αντιστοιχεί στο ενδεχόμενο ΕΕΕ την τιμή 1. Γενικότερα, η σύνδεση μεταξύ των δυνατών αποτελεσμάτων (ενδεχομένων) του δείγματος και των δυνατών τιμών της στατιστικής συνάρτησης συνοψίζεται στον πίνακα που ακολουθεί:

Ενδεχόμενο	ΕΕΕ	ΕΕΕ	ΕΕΕ	ΕΕΕ	ΕΕΕ	ΕΕΕ	ΕΕΕ	ΕΕΕ
Τιμή	0	1	1	1	2	2	2	3

Είναι προφανές από τον παραπάνω πίνακα ότι μία ελαττωματική μονάδα σε ένα σύνολο τριών μονάδων του προϊόντος μπορεί να προκύψει αν το αποτέλεσμα της λήψης ενός δείγματος τριών μονάδων είναι ΕΕΕ ή ΕΕΕ ή ΕΕΕ. Επομένως, η πιθανότητα μιας ελαττωματικής μονάδας σε ένα δείγμα τριών μονάδων του προϊόντος είναι ίση με την πιθανότητα εμφάνισης ενός από τα παραπάνω τρία δυνατά αποτελέσματα. Χρησιμοποιώντας τα αξιώματα των πιθανοτήτων, είναι δυνατόν να κατασκευασθεί η δειγματική κατανομή της στατιστικής συνάρτησης από την δειγματική κατανομή του δείγματος, όπως φαίνεται στον πίνακα που ακολουθεί.

Δειγματική Κατανομή Δείγματος		Δειγματική Κατανομή της Στατιστικής Συνάρτησης T	
Δείγμα	Πιθανότητα (π_s)	Τιμή της Στατιστικής Συνάρτησης (t_i) (Αριθμός Ελαττωματικών)	Πιθανότητα (π_T)
ΕΕΕ	0.970299	0	0.970299
ΕΕΕ	0.009801		
ΕΕΕ	0.009801	1	0.029403
ΕΕΕ	0.009801		
ΕΕΕ	0.000099		
ΕΕΕ	0.000099	2	0.000297
ΕΕΕ	0.000099		
ΕΕΕ	0.000001	3	0.000001

Είναι προφανές ότι

$$\pi_T(t_i) = \sum \pi_s(y_1, y_2, ..., y_n) \quad \text{για} \quad i = 1, ..., k$$

όπου το άθροισμα στο δεξί μέλος της παραπάνω σχέσης εκτείνεται σε όλα τα δυνατά αποτελέσματα στα οποία αντιστοιχείται η τιμή t_i από την στατιστική συνάρτηση T.

Στην συνέχεια, θα εξετάσουμε την δειγματική κατανομή του μέσου $\bar{X}$ ενός δείγματος n ανεξαρτήτων παρατηρήσεων από μία οποιαδήποτε κατανομή καθώς και άλλες δειγματικές κατανομές στατιστικών συναρτήσεων που συχνά χρησιμοποιούνται στις πρακτικές εφαρμογές, όπως η διασπορά ενός δείγματος παρατηρήσεων ή ο λόγος των διασπορών δύο ανεξάρτητων δειγμάτων παρατηρήσεων.

Η ΚΑΝΟΝΙΚΗ ΚΑΤΑΝΟΜΗ ΩΣ ΔΕΙΓΜΑΤΙΚΗ ΚΑΤΑΝΟΜΗ

Όπως ήδη αναφέρθηκε, η κανονική κατανομή χρησιμοποιείται για την περιγραφή πολλών ποσοτικών φαινομένων. Αποτελεί όμως χρήσιμο εργαλείο και στην πειραματική έρευνα. Όπως αναφέρθηκε στο προηγούμενο κεφάλαιο, μέσω της χρήσης του κεντρικού οριακού

θεωρήματος, η κανονική κατανομή μπορεί να χρησιμοποιηθεί για τη συναγωγή συμπερασμάτων όσο αφορά την ακρίβεια με την οποία ο μέσος $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ ενός δείγματος n ανεξαρτήτων παρατηρήσεων $X_1, X_2, ..., X_n$ από μια οποιαδήποτε κατανομή εκτιμά την μέση της τιμή μ της κατανομής αυτής.

Συγκεκριμένα, η χρήση της κανονικής κατανομής κάνει εφικτό τον προσδιορισμό της πιθανότητας $P(-\varepsilon \leq \overline{X}-\mu \leq \varepsilon)$, δηλαδή της πιθανότητας με την οποία η εκτίμηση που παρέχει ο $\overline{X}$ δεν θα απέχει από την πραγματική αλλά άγνωστη τιμή της μέσης τιμής μ περισσότερο από ε.

Από τον ορισμό του, ο μέσος $\overline{X}$ είναι μια τυχαία μεταβλητή η οποία ακολουθεί μια κατανομή που εν γένει είναι άγνωστη. Η κατανομή αυτή ονομάζεται *δειγματική κατανομή του μέσου*. Ο καθορισμός της μορφής της κατανομής αυτής είναι προφανώς απαραίτητος για τον υπολογισμό της πιθανότητας $P(-\varepsilon \leq \overline{X}-\mu \leq \varepsilon)$ που αναφέραμε παραπάνω και γίνεται δυνατός, όπως είδαμε στο προηγούμενο κεφάλαιο, με την χρήση του κεντρικού οριακού θεωρήματος.

Πιο συγκεκριμένα στην περιοχή της στατιστικής συμπερασματολογίας, συχνά ενδιαφερόμαστε να εκτιμήσουμε την άγνωστη μέση τιμή μ μιας τυχαίας μεταβλητής Χ. Έστω ότι για το σκοπό αυτό πραγματοποιήσαμε n ανεξάρτητες παρατηρήσεις $X_1, X_2, \ldots, X_n$, πάνω στην τυχαία μεταβλητή Χ. (Υποθέτουμε δηλαδή ότι ελήφθησαν παρατηρήσεις από n ανεξάρτητες επαναλήψεις του τυχαίου πειράματος που παρήγαγε την τυχαία μεταβλητή Χ). Ένας φυσικός και προφανής τρόπος εκτίμησης της μέσης τιμής μ είναι με την χρήση του μέσου των παρατηρήσεων, που συνήθως ονομάζεται δειγματικός μέσος:

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

Επειδή από τον ορισμό του ο δειγματικός μέσος είναι μία τυχαία μεταβλητή, δεν θα μπορούσε να περιμένει κανείς ότι η τιμή του θα είναι πάντα ίση με την άγνωστη τιμή μ, αφού κάθε σύνολο n

ανεξάρτητων παρατηρήσεων πάνω στην X θα οδηγεί σε διαφορετική τιμή του $\overline{X}$. Μπορούμε όμως να "μετρήσουμε" πόσο κοντά γύρω από την άγνωστη τιμή μ κυμαίνονται οι τιμές του $\overline{X}$ με τον υπολογισμό της πιθανότητας $P(-\varepsilon \le \overline{X} - \mu \le \varepsilon)$. Η κατανομή του $\overline{X}$, επομένως, και η ζητούμενη πιθανότητα μπορεί να είναι δύσκολο να προσδιορισθούν κυρίως όταν η κατανομή της τυχαίας μεταβλητής X δεν είναι απλής μορφής. Από το κεντρικό οριακό θεώρημα προκύπτει ότι, ανεξάρτητα από την μορφή της κατανομής της τυχαίας μεταβλητής X, καθώς ο αριθμός n ανεξαρτήτων δοκιμών αυξάνει, η παραπάνω πιθανότητα μπορεί να προσεγγισθεί από μια πιθανότητα που υπολογίζεται από την τυποποιημένη κανονική κατανομή:

$$P(-\varepsilon \le \overline{X} - \mu \le \varepsilon) \cong P(\frac{-\varepsilon\sqrt{n}}{\sigma} \le Z \le \frac{\varepsilon\sqrt{n}}{\sigma})$$

$$= 2\,\Phi(\frac{\varepsilon\sqrt{n}}{\sigma}) - 1,$$

όπου Z είναι μία τυποποιημένη κανονική μεταβλητή, σ^2 είναι η διασπορά της X. Γενικότερα, ισχύει ότι

$$P(\alpha \le \overline{X} - \mu \le \beta) \cong \Phi(\frac{\beta\sqrt{n}}{\sigma}) - \Phi(\frac{\alpha\sqrt{n}}{\sigma})$$

Παράδειγμα: Ενας αστρονόμος σχεδίαζει να πραγματοποιήσει ανεξάρτητες μετρήσεις $X_1, X_2, \ldots, X_n$, της πραγματικής απόστασης D σε έτη φωτός μεταξύ του αστεροσκοπείου του και ενός απομακρυσμένου αστέρα. Οι μετρήσεις αυτές είναι γνωστό ότι έχουν μια κοινή κατανομή (είναι ισόνομες) με μέση τιμή $\mu = D$ και διασπορά $\sigma^2 = 4$. Για να εκτιμηθεί η τιμή D με μία ακρίβεια ± 0.25 ετών φωτός, ο αστρονόμος ίσως χρησιμοποιήσει τον μέσο $\overline{X} = \sum_{i=1}^{100} X_i/100$ εκατό παρατηρήσεων. Τότε, θέτοντας $\varepsilon = 0.25$ και $n = 100$, η πιθανότητα να επιτευχθεί η επιθυμητή ακρίβεια είναι

$$P(-0.25 \le \overline{X} - D \le 0.25) \cong 2\Phi(\frac{0.25\sqrt{100}}{\sqrt{4}}) - 1$$

$$= 2\Phi(1.25) - 1 = 2(0.89435) - 1 = 0.7888$$

Εάν ο αστρονόμος χρησιμοποιήσει n = 400 παρατηρήσεις, η πιθανότητα αυτή αυξάνει. Πράγματι

$$P(-0.25 \le \overline{X} - D \le 0.25) \cong 2\Phi\left(\frac{0.25\sqrt{400}}{\sqrt{4}}\right) - 1$$
$$= 2\Phi(2.5) - 1 = 2(0.9938) - 1 = 0.9876.$$

Παρατήρηση: Είναι σαφές ότι μόνο σε μία περίπτωση η τυχαία μεταβλητή Z_n (όπως και η $\overline{X}$ και η S_n) έχει ακριβώς την κανονική κατανομή ανεξάρτητα από το πόσες παρατηρήσεις έχουν ληφθεί πάνω στην τυχαία μεταβλητή X. Αυτή είναι η περίπτωση όπου οι $X_1, X_2, \ldots, X_n$, έχουν μια κανονική κατανομή με μέση τιμή μ και διασπορά σ^2.

Παράδειγμα: Για να μελετήσουμε την συμπεριφορά μαθητών σχολείων που παρουσιάζουν μία δυσκολία να συγκεντρωθούν (disruptive school children), ένας ψυχολόγος θέλει να εκτιμήσει τον μέσο δείκτη νοημοσύνης μ αυτών των παιδιών με μία ακρίβεια ± 5. Με βάση την μελέτη του Terman, υποθέτει ότι οι μετρήσεις του δείκτη νοημοσύνης X για τα παιδιά αυτά ακολουθούν μια κανονική κατανομή με μέση τιμή μ και διασπορά σ^2 =263.66. Ο δειγματικός μέσος $\overline{X} = \sum_{i=1}^{20} X_i/20$ ενός δείγματος 20 παρατηρήσεων $X_1, X_2, \ldots, X_n$, που ελήφθησαν ανεξάρτητα για 20 από τα παιδιά χρησιμοποιείται για την εκτίμηση της μέσης τιμής μ. Επειδή οι παρατηρήσεις $X_1, X_2, \ldots, X_{20}$, είναι ανεξάρτητες και ισόνομες κανονικές μεταβλητές, ο μέσος $\overline{X}$ έχει ακριβώς την κανονική κατανομή με μέση τιμή μ και διασπορά ίση με 263.66/20, παρά το γεγονός ότι ο αριθμός των 20 παρατηρήσεων που ελήφθησαν δεν είναι πολύ μεγάλος. Έτσι,

$$P(-5 \le \overline{X} - \mu \le 5) = 2\Phi\left[\frac{5\sqrt{20}}{\sqrt{263.66}}\right] - 1$$

$$= 2\Phi(1.38) - 1 = 0.8324$$

Δηλαδή με ένα δείγμα 20 παρατηρήσεων η επιθυμητή ακρίβεια θα επιτευχθεί με πιθανότητα 83.24%. Όπως όμως παρατηρούμε στον

πίνακα της τυποποιημένης κανονικής κατανομής 2Φ(2.00)–1=0.9545. Αν δηλαδή επιθυμούμε να επιτύχουμε μια εκτίμηση της μέσης τιμής μ η οποία να βρίσκεται σε διάστημα ± 5 γύρω από την πραγματική τιμή με πιθανότητα 95.45%, τότε θα πρέπει να διαλέξουμε το μέγεθος του δείγματος ώστε να ικανοποιεί την εξίσωση

$$\frac{5\sqrt{n}}{\sqrt{263.66}} = 2.00$$

Δηλαδή θα πρέπει να μαζέψουμε στοιχεία για n = 42.19 ≅ 43 παιδιά για να εξασφαλίσουμε την επιθυμητή ακρίβεια.

Η ΚΑΤΑΝΟΜΗ X^2

Ορισμός: Θα λέμε ότι η τυχαία μεταβλητή X ακολουθεί *την κατανομή* X^2 *με* r *βαθμούς ελευθερίας* και θα συμβολίζουμε με $X \sim X^2_r$ αν

$$f(x) = \frac{1}{\Gamma\left(\frac{r}{2}\right)2^{r/2}} e^{-\frac{x^2}{2}} x^{\frac{r}{2}-1}, \quad 0 \le x < \infty, \text{ r θετικός ακέραιος.}$$

Ιδιότητες:

$$E(X) = \alpha = r, \qquad \Delta(X) = \alpha =^2 2r.$$

Υπολογισμός πιθανοτήτων της κατανομής X^2

Υπάρχουν πίνακες που δίνουν την συνάρτηση κατανομής της X. (Βλέπε παράρτημα).

Η κατανομή X^2 παρουσιάζει ιδιαίτερο ενδιαφέρον στην Στατιστική γιατί συνδέεται με την δειγματική κατανομή της εκτιμήτριας S^2 της διακύμανσης σ^2 ενός κανονικού πληθυσμού.
Αυτό προκύπτει από το θεώρημα που ακολουθεί.

Το θεώρημα που ακολουθεί αναφέρεται στην σχέση της μέσης τιμής και της διασποράς ενός δείγματος n ανεξαρτήτων παρατηρήσεων $X_1, X_2, \ldots, X_n$ (n τυχαίων μεταβλητών) από ένα πληθυσμό που ακολουθεί την κανονική κατανομή. Η σχέση αυτή είναι θεμελιώδους σημασίας στην Στατιστική Συμπερασματολογία.

Θεώρημα: Έστω $X_1, X_2, ..., X_n$ είναι ένα τυχαίο δείγμα μεγέθους n από ένα πληθυσμό $N(\mu,\sigma^2)$ με δειγματικό μέσο $\overline{X}$ και δειγματική διασπορά S^2 όπου

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n} \quad \text{και} \quad S^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n}$$

Τότε, για τις τυχαίες μεταβλητές $\overline{X}$ και S^2 ισχύει ότι,

α) $\overline{X}$ και S^2 είναι ανεξάρτητες τυχαίες μεταβλητές και

β) $\frac{n}{\sigma^2} S^2 \sim X^2_{n-1}$.

Απόδειξη: Η απόδειξη είναι αρκετά πολύπλοκη και βρίσκεται πέρα από τους σκοπούς του βιβλίου αυτού.

Παράδειγμα: Οι αφίξεις αυτοκινήτων σε ένα σταθμό διοδίων ακολουθούν την κατανομή Poisson με μέσο ρυθμό 5 αυτοκινήτων ανά δεκάλεπτο. Να υπολογισθεί η πιθανότητα με την οποία ο υπάλληλος των διοδίων θα χρειασθεί να περιμένει περισσότερο από 25 λεπτά και 30 δευτερόλεπτα μέχρις ότου περάσουν 8 αυτοκίνητα.

Λύση: Έστω X ο χρόνος αναμονής (σε λεπτά). Είναι γνωστό ότι το X ακολουθεί την κατανομή γάμμα με α=n=8 και θ=2 (διότι λ=1/2 μια και έχουμε 5 αυτοκίνητα ανά δεκάλεπτο). Επομένως

$$X \sim X^2_{16}$$

και κατά συνέπεια $P(X>25.50) = 1-P(X\leq 25.50) \cong 1-(0.95) = 0.05$.

Η ΚΑΤΑΝΟΜΗ t

Έστω X μια τυχαία μεταβλητή με συνάρτηση πυκνότητας πιθανότητας,

$$f(x) = \frac{\Gamma\left(\frac{r+1}{2}\right)}{\sqrt{\pi r}\,\Gamma(r/2)\,\left(1 + x^2/r\right)^{(r+1)/2}}, \quad -\infty < x < +\infty, \; r=1, 2, ...$$

όπου $\Gamma(r)$ είναι η συνάρτηση γάμμα.

Αν η τυχαία μεταβλητή X ακολουθεί μια κατανομή με συνάρτηση πυκνότητας πιθανότητας όπως η παραπάνω θα λέμε

ότι η X *θα ακολουθεί την κατανομή* t *(ή κατανομή Student) με* r *βαθμούς ελευθερίας και θα συμβολίζουμε με* $X \sim t_r$.

Ιδιότητες της κατανομής t

1. Η κατανομή t είναι συμμετρική γύρω από το μηδέν, επομένως,
$E(X) = 0$, αν η μέση αυτή τιμή υπάρχει.
(Αποδεικνύεται ότι η μέση αυτή τιμή υπάρχει για $r \geq 2$).
2. Επίσης, για τη διακύμανση έχουμε,
$V(X) = E(X^2) = r/(r-2)$ $r \geq 3$ (για r = 1,2 η V(X) δεν υπάρχει).

Σημείωση: Υπάρχουν πίνακες που δίνουν τις πιθανότητες για τις διάφορες τιμές των βαθμών ελευθερίας. (Βλέπε παράρτημα).

Για παράδειγμα για r = 10

$$P(X_{10} < 1.812) = F(1.812) = 0.95$$
$$P(1.372 < X_{10} < 2.228) = F(2.228) - F(1.372)$$
$$= 0.975 - 0.90$$
$$= 0.075$$

Θεώρημα: Αν $Z \sim N(0,1)$ και $U \sim X^2_r$ και οι τυχαίες μεταβλητές Z και U είναι ανεξάρτητες τότε η τυχαία μεταβλητή

$$T = \frac{Z}{\sqrt{U/r}} \sim t_r,$$

δηλαδή η T ακολουθεί την t κατανομή με r βαθμούς ελευθερίας.

Σημείωση: Η κατανομή t είναι ιδιαίτερα χρήσιμη στη Στατιστική γιατί εκφράζει την κατανομή του τυποποιημένου δειγματικού μέσου ενός κανονικού πληθυσμού όταν για την τυποποίηση χρησιμοποιείται η εκτιμήτρια S^2 του σ^2. Αυτό φαίνεται στο θεώρημα που ακολουθεί.

Θεώρημα: Έστω $X_1, X_2, ..., X_n$ ένα τυχαίο δείγμα μεγέθους n από μια κατανομή $N(\mu,\sigma^2)$. Τότε η στατιστική συνάρτηση,

$$\frac{\overline{X}-\mu}{S/\sqrt{n-1}} \equiv \frac{\overline{X}-\mu}{S^*/\sqrt{n}} \sim t_{n-1}$$

όπου S^* είναι η θετική τετραγωνική ρίζα της ποσότητας $S^{*2} = \sum_{i=1}^{n}(X_i - \overline{X})/(n-1).$

Απόδειξη: Δοθέντος ότι,

$$\frac{\overline{X}-\mu}{S/\sqrt{n-1}} \equiv \frac{\overline{X}-\mu}{S^*/\sqrt{n}} \sim N(0,1)$$

και ότι

$$\frac{n}{\sigma^2} S^2 \sim X^2_{n-1}$$

έχουμε, από το προηγούμενο θεώρημα (δοθέντος ότι $\overline{X}$ και S^2 είναι ανεξάρτητες τυχαίες μεταβλητές),

$$\frac{\dfrac{\overline{X}-\mu}{\sigma/\sqrt{n}}}{\sqrt{\dfrac{n}{\sigma^2}S^2/(n-1)}} = \frac{\overline{X}-\mu}{S/\sqrt{(n-1)}} \equiv \frac{\overline{X}-\mu}{S^*/\sqrt{n}} \sim t_{n-1}.$$

Η ΚΑΤΑΝΟΜΗ F

Έστω W μια τυχαία μεταβλητή με συνάρτηση πυκνότητας πιθανότητας,

$$h(w) = \frac{\Gamma\left[\dfrac{r_1+r_2}{2}\right]\left(\dfrac{r_1}{r_2}\right)^{r_1/2} w^{(r_1/2)-1}}{\Gamma\left[\dfrac{r_1}{2}\right]\Gamma\left[\dfrac{r_2}{2}\right]\left[1+\dfrac{r_1 w}{r_2}\right]^{(r_1+r_2)/2}}, \quad 0 < w < \infty$$

όπου Γ(α) είναι η συνάρτηση γάμμα.

Μια τυχαία μεταβλητή που έχει αυτή τη συνάρτηση πυκνότητας πιθανότητας λέμε ότι *ακολουθεί την κατανομή* F *με* r_1 *και* r_2 *βαθμούς ελευθερίας και συμβολίζεται με* $X \sim F_{r_1, r_2}$.

Σημείωση: Μπορεί να αποδειχθεί ότι αν W ακολουθεί μια κατανομή F με r_1 και r_2 βαθμούς ελευθερίας τότε,

$$E(W)=\frac{r_2}{r_2-2}$$

και

$$V(W)=\frac{2r_2^2(r_1+r_2-2)}{r_1(r_2-2)^2(r_2-4)}$$

Η κατανομή F είναι σημαντική στη Στατιστική λόγω του θεωρήματος που ακολουθεί.

Θεώρημα: Αν $U \sim X^2_{r_1}$ και $V \sim X^2_{r_2}$ και U, V είναι ανεξάρτητες τυχαίες μεταβλητές τότε η τυχαία μεταβλητή,

$$F=\frac{U/r_1}{V/r_2}$$

ακολουθεί μια κατανομή F με r_1 και r_2 βαθμούς ελευθερίας.

Η κατανομή F είναι σημαντική για την στατιστική συμπερασματολογία διότι, όπως γνωρίζουμε, για τις δειγματικές διακυμάνσεις ισχύει ότι, αν $X_1, X_2, ...,X_n$ είναι ένα τυχαίο δείγμα μεγέθους n από έναν κανονικό πληθυσμό $N(\mu_X, \sigma^2_X)$ και $Y_1,Y_2,...,Y_m$ είναι ένα άλλο τυχαίο δείγμα μεγέθους m από έναν άλλο κανονικό πληθυσμό $N(\mu_Y, \sigma^2_Y)$ ο οποίος είναι ανεξάρτητος από τον πρώτο τότε,

$$\frac{(n-1)S^{*2}_X}{\sigma^2_X} \equiv \frac{nS^2_X}{\sigma^2_X} \sim X^2_{n-1}$$

και

$$\frac{(m-1)S^{*2}_Y}{\sigma^2_Y} \equiv \frac{m\,S^2_Y}{\sigma^2_Y} \sim X^2_{m-1}$$

Επομένως σύμφωνα με το προηγούμενο θεώρημα,

$$\frac{\frac{S_X^{*2}}{\sigma_X^2}}{\frac{S_Y^{*2}}{\sigma_Y^2}} \equiv \frac{\frac{nS_X^2}{(n-1)\sigma_X^2}}{\frac{mS_Y^2}{(m-1)\sigma_y^2}} \sim F_{n-1,\, m-1}$$

(S_X^2, S_Y^2 είναι ανεξάρτητες τυχαίες μεταβλητές δεδομένου ότι και X, Y είναι ανεξάρτητες τυχαίες μεταβλητές).