“Πιθανότητες και Αρχές Στατιστικής” (10η Διάλεξη)

Σωτήρης Νικολετσέας, καθηγητής

Τμήμα Μηχανικών Η/Υ & Πληροφορικής, Πανεπιστήμιο Πατρών

Ακαδημαϊκό Έτος 2023 - 2024

Σύνοψη

Α. Διαστήματα Εμπιστοσύνης (εισαγωγή)

Β. Ειδικές Κατανομές που προέρχονται από την Κανονική

- Η Κατανομή χ^2 (chi-square)
- Η Κατανομή t του Student
- Η F-Κατανομή του Fisher

Γ. Διαστήματα Εμπιστοσύνης για τη μέση τιμή Κανονικού δείγματος

Δ. Προσεγγιστικά διαστήματα εμπιστοσύνης για αρκετά μεγάλα δείγματα από οποιαδήποτε κατανομή

Ε. Διαστήματα Εμπιστοσύνης για τη διασπορά Κανονικού δείγματος

ΣΤ. Διαστήματα Εμπιστοσύνης για τη διαφορά των μέσων τιμών δύο κανονικών δειγμάτων

Α. Διαστήματα Εμπιστοσύνης (εισαγωγή)

- Ας υποθέσουμε ότι έχουμε βρεί μία συνεπή εκτιμήτρια για κάποια άγνωστη παράμετρο π.χ. για ένα δείγμα από την κανονική κατανομή $N(\mu, \sigma^2)$, ο δειγματικός μέσος $\bar{X} = \bar{X}_\nu = \frac{1}{\nu}\sum_{i=1}^{\nu} X_i$ είναι συνεπής εκτιμήτρια συνάρτηση για την άγνωστη μέση τιμή μ. Άρα αν το δείγμα είναι “αρκετά μεγάλο” η τιμή του $\bar{X}$ θα είναι “αρκετά κοντά” στην πραγματική άγνωστη τιμή μ.
- Δηλαδή, η σημειακή εκτίμηση δεν είναι ίση με την πραγματική τιμή αλλά οι πραγματικές τιμές είναι “κοντά” στην εκτίμηση.
- Επομένως, αντί για μία σημειακή εκτίμηση (δηλαδή την εκτίμηση μιας μοναδικής τιμής) θα ήταν συχνά περισσότερο χρήσιμο να προσδιορίσουμε ένα διάστημα τιμών που η παράμετρος παίρνει με “αρκετά μεγάλη” βεβαιότητα (π.χ. με πιθανότητα 95%).

Διάστημα Εμπιστοσύνης - ορισμός

- Ορισμός: Θεωρούμε το δείγμα $X_1, X_2, \ldots, X_\nu$ από κατανομή $F(x;\theta)$, όπου θ άγνωστη παράμετρος προς εκτίμηση, και σταθερά $a \in (0,1)$, που συνήθως παίρνει αρκετά μικρές τιμές $a = 0.05(5\%)$ ή $a = 0.01(1\%)$.
  Έστω οι εκτιμήτριες:
  $$L = T_1(X_1, \ldots, X_\nu) \text{ και } U = T_2(X_1, \ldots, X_\nu)$$
  για τις οποίες είναι:
  (i) $\Pr\{L \leq U\} = 1$
  (ii) $\Pr\{L \leq \theta \leq U\} = 1 - a$
  Τότε το διάστημα $[L, U]$ καλείται $100(1-a)\%$ διάστημα εμπιστοσύνης για το θ (και η πιθανότητα $1 - a$ καλείται συντελεστής εμπιστοσύνης του διαστήματος)
- π.χ. αν $a = 0.05$ τότε το διάστημα $[L, U]$ καλείται 95% διάστημα εμπιστοσύνης (και επίσης λέμε ότι το εκτιμώμενο διάστημα $[L, U]$ έχει συντελεστή εμπιστοσύνης 95%, με άλλα λόγια το 95% των τιμών αναμένεται να είναι μεταξύ L και U).

Παράδειγμα

- Ένα σήμα με τιμή μ μεταδίδεται από την θέση A και λαμβάνεται στην θέση B με τιμή κανονικά κατανεμημένη με μέση τιμή μ και διασπορά 4. Δηλαδή, αν στείλουμε τιμή μ η τιμή που λαμβάνουμε είναι $\mu + N$, όπου το N είναι θόρυβος, με κανονική κατανομή με μέση τιμή 0 και διασπορά 4. Για να μειώσουμε το λάθος, ας υποθέσουμε οτι στέλνουμε την ίδια τιμή 9 φορές και οι ληφθείσες τιμές είναι 5, 8.5, 12, 15, 7, 9 7.5, 6.5, 10.5. Τι μπορούμε να συμπεράνουμε για την αποσταλείσα τιμή μ;
- Μια σημειακή εκτίμηση για την μ είναι ο δειγματικός μέσος των ληφθεισών τιμών:

$$\bar{X} = \frac{5+8.5+\cdots+10.5}{9} = \frac{81}{9} = 9$$

- Στη συνέχεια θα υπολογίσουμε ένα 95% διάστημα εμπιστοσύνης, χρησιμοποιώντας τη γνώση της κατανομής του δειγματικού μέσου $\bar{X}$.

Β. Ειδικές κατανομές - Η κατανομή Γ

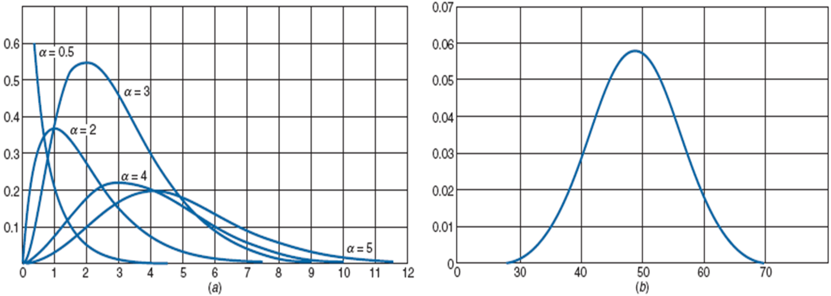
- Ορισμός: Μια τ.μ. ακολουθεί την κατανομή Γ με παραμέτρους $(a, \lambda), \lambda > 0, a > 0$, αν έχει σ.π.π.:
$$f(x) = \begin{cases} \frac{\lambda e^{-\lambda x}(\lambda x)^{a-1}}{\Gamma(a)}, & \text{αν } x \geq 0 \\ 0, & \text{αν } x < 0 \end{cases}$$
όπου $\Gamma(a) = \int_0^\infty \lambda e^{-\lambda x}(\lambda x)^{a-1}\,dx$
Είναι: $\Gamma(a) = (a-1)\Gamma(a-1)$
και για ακέραια a (π.χ. $a = n$) είναι $\Gamma(n) = (n-1)!$. Η συνάρτηση $\Gamma(a)$ καλείται συνάρτηση Γ (gamma function).

- Παρατήρηση: Αν $a = 1$, η κατανομή Γ ανάγεται στην εκθετική με παράμετρο λ.

- Η ροπογεννήτρια είναι $\Phi(t) = \left(\frac{\lambda}{\lambda - t}\right)^a$

- Η μέση τιμή είναι $E[X] = \Phi'(0) = \frac{a}{\lambda}$ και η διασπορά είναι $\mathrm{Var}(X) = E(X^2) - E^2(X) = \Phi''(0) - \left(\frac{a}{\lambda}\right)^2 = \frac{a}{\lambda^2}$

Η κατανομή Γ



Graphs of the gamma $(\alpha, 1)$ *density for (a)* $\alpha = .5, 2, 3, 4, 5$ *and (b)* $\alpha = 50$.

Παρατήρηση: Όσο μεγαλώνει η παράμετρος a, η καμπύλη της *pdf* αρχίζει να μοιάζει με την κανονική (η εξήγηση βασίζεται στο Κεντρικό Οριακό Θεώρημα)

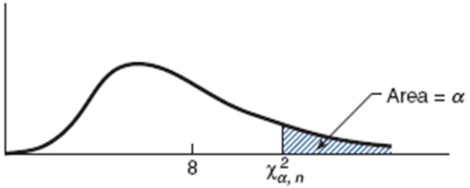
Σημαντικές ιδιότητες της κατανομής Γ

- Αν οι X_1, X_2 είναι ανεξάρτητες τ.μ. με κατανομή Γ με παραμέτρους (a_1, λ) και (a_2, λ), τότε η $X_1 + X_2$ ακολουθεί την κατανομή Γ με παραμέτρους $(a_1 + a_2, \lambda)$.
- Επειδή για παραμέτρους $(1, \lambda)$ η Γ ανάγεται στην εκθετική με παράμετρο λ, ένα άθροισμα $\sum_{i=1}^{n} X_i$ από n ανεξάρτητες εκθετικές κατανομές ακολουθεί την κατανομή Γ με παραμέτρους (n, λ).
- <u>Παράδειγμα:</u> Ο χρόνος ζωής μιας μπαταρίας ακολουθεί εκθετική κατανομή με παράμετρο λ. Αν μια συσκευή με n μπαταρίες χρειάζεται τουλάχιστον 1 (μη άδεια) μπαταρία για να λειτουργεί, ο χρόνος λειτουργίας της ακολουθεί την κατανομή Γ με παραμέτρους (n, λ).

Η κατανομή χ^2 (chi-square)

- Ορισμός: Αν $Z_1, Z_2, \ldots, Z_n$ ανεξάρτητες τυπικές κανονικές τ.μ., τότε το άθροισμα
  $X = Z_1^2 + Z_2^2 + \cdots + Z_n^2$
  ακολουθεί την κατανομή χ^2 με n βαθμούς ελευθερίας.
  Συμβολίζουμε με $X \sim \chi_n^2$
- Αν $X \sim \chi_n^2$ τότε για κάθε $a \in (0,1)$ η ποσότητα $\chi_{a,n}^2$ ορίζεται ώς εξής:

$$\Pr\{X \geq \chi_{a,n}^2\} = a$$



Οι τιμές $\chi_{a,n}^2$ υπάρχουν σε πίνακες και θεωρούνται γνωστές.

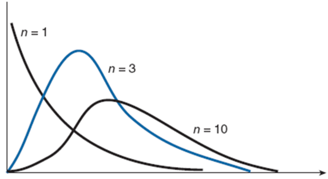
Ιδιότητες της κατανομής χ^2

- Αθροιστική ιδιότητα. Για ανεξάρτητες τ.μ. $X_1 \sim \chi^2_{n_1}$ και $X_2 \sim \chi^2_{n_2}$ το άθροισμά τους είναι $X \sim \chi^2_{n_1+n_2}$
- Παράδειγμα: Έστω ότι θέλουμε να εντοπίσουμε έναν στόχο στον τρισδιάστατο χώρο, και τα λάθη (σε μέτρα) στις 3 συντεταγμένες ενός τυχαία επιλεγμένου σημείου είναι ανεξάρτητες κανονικές τ.μ. με μέση τιμή 0 και τυπική απόκλιση 2. Να βρεθεί η πιθανότητα η απόσταση του επιλεγμένου σημείου από το στόχο να ξεπερνά τα 3 μέτρα.

  Λύση: Έστω X_i το λάθος στην συντεταγμένη i (i=1, 2, 3), άρα οι $\frac{X_i}{2}$ είναι τυπικές κανονικές δηλαδή $\frac{X_i}{2} = Z_i \sim N(0,1)$. Αν D η ζητούμενη απόσταση, είναι $D^2 = X_1^2 + X_2^2 + X_3^2$ και
  $\Pr\{D > 3\} = \Pr\{D^2 > 9\} = \Pr\{Z_1^2 + Z_2^2 + Z_3^2 > \frac{9}{4}\} =$
  $= \Pr\{\chi_3^2 > \frac{9}{4}\} = 0.5222.$

Η σχέση της χ^2 με την κατανομή Γ

- Μέσω των αντίστοιχων ροπογεννητριών, μπορούμε να δείξουμε ότι η κατανομή χ^2 με n βαθμούς ελευθερίας ανάγεται στην κατανομή Γ με παραμέτρους $\left(\frac{n}{2}, \frac{1}{2}\right)$.
- Άρα η *pdf* της χ^2 είναι: $f(x) = \dfrac{\frac{1}{2} e^{-\frac{x}{2}} \left(\frac{x}{2}\right)^{\left(\frac{n}{2}\right)-1}}{\Gamma\left(\frac{n}{2}\right)}, x > 0$
- Πάλι παρατηρούμε σύγκλιση στην κανονική (για πολλούς βαθμούς ελευθερίας):



The chi-square density function with n degrees of freedom.

- Από την ίδια σχέση προκύπτει ότι, αν $X \sim \chi^2_n$ τότε $E(X) = n$, $\mathrm{Var}(X) = 2n$

Παράδειγμα

Παίρνουμε πάλι το πρόβλημα της ακρίβειας στόχευσης, αλλά αυτή τη φορά στο επίπεδο (2 συντεταγμένες - βαθμοί ελευθερίας). Αν η ζητούμενη απόσταση είναι D και τα λάθη στις 2 συντεταγμένες X_i (i=1, 2) είναι κανονικά με μέση τιμή 0 και τυπική απόκλιση 2, οι Z_i είναι τυπικές κανονικές $N(0,1)$. Άρα:

$$\Pr\{D^2 > 9\} = \Pr\{Z_1^2 + Z_2^2 > \tfrac{9}{4}\} = \Pr\{\chi_2^2 > \tfrac{9}{4}\}$$

Αλλά η χ_n^2 ανάγεται στην Γ με παραμέτρους $a = \frac{n}{2} = 1, \lambda = \frac{1}{2}$.

Άρα $Pr\{\chi_2^2 > \frac{9}{4}\} = \Pr\{X > \frac{9}{4}\} = 1 - \Pr\{X < \frac{9}{4}\} =$

$$= 1 - (1 - e^{-\frac{1}{2}\frac{9}{4}}) = e^{-\frac{9}{8}} \simeq 0.3247$$

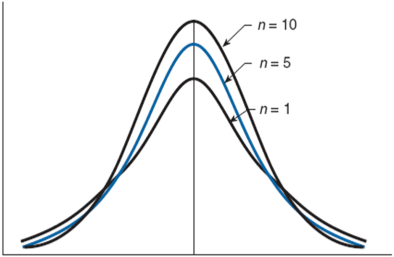
(δηλαδή η ακρίβεια είναι μεγαλύτερη). Χρησιμοποιήθηκε η αναγωγή της $\Gamma(1,\lambda)$ στην εκθετική με παράμετρο λ.

Η κατανομή t_n του Student

- Ορισμός: Αν $Z \sim N(0,1)$ και χ_n^2 ανεξάρτητες τ.μ., τότε η τ.μ.

$$T_n = \frac{Z}{\sqrt{\frac{\chi_n^2}{n}}}$$

καλείται t-κατανομή με n βαθμούς ελευθερίας.



Density function of T_n.

- Όπως και η τυπική κανονική, η t_n είναι συμμετρική περί το 0, και όσο το n μεγαλώνει η t_n μοιάζει με την τυπική κανονική.

Η κατανομή t_n (συνέχεια)

Αυτό οφείλεται στο ότι, όπως είδαμε, η χ_n^2 μπορεί να εκφραστεί σαν άθροισμα των τετραγώνων n ανεξάρτητων τυπικών κανονικών, οπότε

$$\frac{\chi_n^2}{n} = \frac{Z_1^2+Z_2^2+\cdots+Z_n^2}{n}$$

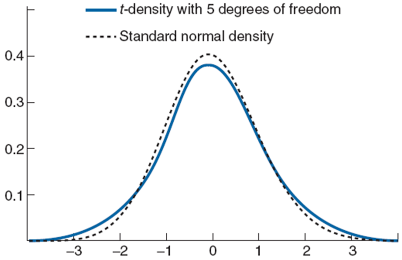
Επειδή $Z_i \sim N(0,1)$ είναι (άσκηση) $E(Z_i^2)=1$, επομένως από τον ασθενή νόμο των μεγάλων αριθμών, με πιθανότητα που τείνει στο 1, είναι:

$$\frac{Z_1^2+Z_2^2+\cdots+Z_n^2}{n} \simeq E(Z_i^2)$$

Δηλαδή $\frac{\chi_n^2}{n} \simeq 1$ επομένως η T_n έχει προσεγγιστικά την ίδια κατανομή με την Z.

Ιδιότητες

- Η σχέση της T_n με την τυπική κανονική φαίνεται εδώ:

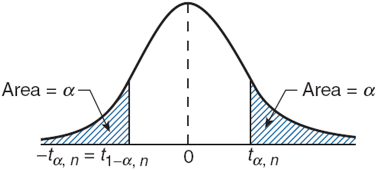


Comparing standard normal density with the density of T_5.

- Επίσης, είναι $E(T_n) = 0$, $n > 1$ και
$$\mathrm{Var}(T_n) = \frac{n}{n-2},\ n > 2$$
και επομένως η διασπορά συγκλίνει μειούμενη στο 1 (τη διασπορά της τυπικής κανονικής) καθώς το n μεγαλώνει.

Ιδιότητες (II)

- Για a: $0 < a < 1$, έστω $t_{a,n}$:
  $\Pr\{T_n \geq t_{a,n}\} = a$
  όπως φαίνεται στην ακόλουθη καμπύλη:



$t_{1-\alpha,n} = -t_{\alpha,n}$.

Από τη συμμετρία περί το 0, προφανώς η $-T_n$ έχει την ίδια κατανομή με την T_n, άρα

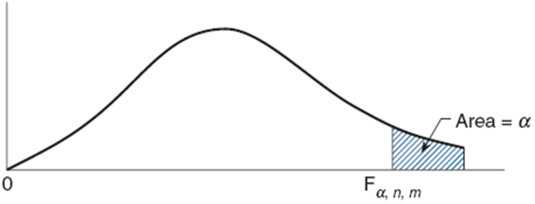
$$\Pr\{T_n \geq -t_{a,n}\} = 1 - a \qquad \text{και επομένως}$$
$$-t_{a,n} = -t_{1-a,n}$$

Οι τιμές $t_{a,n}$ υπάρχουν σε πίνακες και θεωρούνται γνωστές.

Η F-Κατανομή του Fisher

- Ορισμός: Αν χ_n^2 και χ_m^2 ανεξάρτητες τ.μ. με n και m βαθμούς ελευθερίας αντίστοιχα, λέμε ότι η η τ.μ. $F_{n,m}$:
$$F_{n,m} = \frac{\chi_n^2/n}{\chi_m^2/m}$$
ακολουθεί την F-κατανομή με n και m βαθμούς ελευθερίας.

- Για $a \in (0,1)$, ορίζουμε την τιμή $F_{a,n,m}$ ως εξής:
$$\Pr\{F_{n,m} > F_{a,n,m}\} = a$$



Density function of $F_{n,m}$.

Ιδιότητες

- Οι τιμές $F_{a,n,m}$ για $a \leq \frac{1}{2}$ υπάρχουν σε πίνακες και θεωρούνται γνωστές. Για $a > \frac{1}{2}$ είναι:

$$a = \Pr\left\{\frac{\chi_n^2/n}{\chi_m^2/m} > F_{a,n,m}\right\} = \Pr\left\{\frac{\chi_m^2/m}{\chi_n^2/n} < \frac{1}{F_{a,n,m}}\right\}$$
$$= 1 - \Pr\left\{\frac{\chi_m^2/m}{\chi_n^2/n} \geq \frac{1}{F_{a,n,m}}\right\} \text{ και άρα}$$
$$\Pr\left\{\frac{\chi_m^2/m}{\chi_n^2/n} \geq \frac{1}{F_{a,n,m}}\right\} = 1 - a$$

Αλλά η $(\chi_m^2/m)/(\chi_n^2/n)$ προφανώς έχει κατανομή F με m και n βαθμούς ελευθερίας, άρα

$$1 - a = \Pr\left\{\frac{\chi_m^2/m}{\chi_n^2/n} \geq F_{1-a,m,n}\right\}$$

και τελικά $\qquad \dfrac{1}{F_{a,n,m}} = F_{1-a,m,n}$

π.χ. $\qquad F_{0.9,5,7} = \dfrac{1}{F_{0.1,7,5}}$

Γ. Διαστήματα Εμπιστοσύνης για τη μέση τιμή Κανονικού δείγματος

- Σε αυτή τη διάλεξη εστιάζουμε στην εκτίμηση <u>του πληθυσμιακού μέσου (μ)</u> και <u>της πληθυσμιακής διασποράς (σ^2)</u>, κυρίως για τυχαία δείγματα από την <u>κανονική κατανομή $N(\mu, \sigma^2)$</u>.

- Η <u>γενίκευση</u> σε τυχαία δείγματα από <u>οποιαδήποτε κατανομή</u> βασίζεται στο Κεντρικό Οριακό Θεώρημα αλλά και στους νόμους των μεγάλων αριθμών.

Ένα βασικό θεώρημα για κανονικά δείγματα

- Θεώρημα: Έστω τυχαίο δείγμα από κανονική κατανομή, δηλαδή $X_1, X_2, \ldots, X_\nu \sim N(\mu, \sigma^2)$. Είναι:

  α) Ο δειγματικός μέσος ακολουθεί την κανονική κατανομή, δηλαδή για την τ.μ. $\bar{X} = \frac{1}{\nu}\sum_{i=1}^{\nu} X_\nu$ είναι

  $$\bar{X} \sim N(\mu, \tfrac{\sigma^2}{\nu})$$

  β) Η τ.μ. $\frac{(\nu-1)S^2}{\sigma^2}$ (όπου S^2 η δειγματική διασπορά) ακολουθεί την κατανομή $\chi^2_{\nu-1}$, δηλαδή

  $$\frac{(\nu-1)S^2}{\sigma^2} \sim \chi^2_{\nu-1} \equiv \Gamma\left(\tfrac{\nu-1}{2}, \tfrac{1}{2}\right)$$

  γ) Οι τ.μ. $\bar{X}$ και S^2 είναι ανεξάρτητες.

Παράδειγμα

- Ένα σήμα με τιμή μ μεταδίδεται από την θέση A και λαμβάνεται στην θέση B με τιμή κανονικά κατανεμημένη με μέση τιμή μ και διασπορά 4. Δηλαδή, αν στείλουμε τιμή μ η τιμή που λαμβάνουμε είναι $\mu + N$, όπου το N είναι θόρυβος, με κανονική κατανομή με μέση τιμή 0 και διασπορά 4. Για να μειώσουμε το λάθος, ας υποθέσουμε οτι στέλνουμε την ίδια τιμή 9 φορές και οι ληφθείσες τιμές είναι 5, 8.5, 12, 15, 7, 9 7.5, 6.5, 10.5. Τι μπορούμε να συμπεράνουμε για την αποσταλείσα τιμή μ;
- Μια σημειακή εκτίμηση για την μ είναι ο δειγματικός μέσος των ληφθέντων τιμών:

$$\bar{X} = \frac{5+8.5+\cdots+10.5}{9} = \frac{81}{9} = 9$$

- Στη συνέχεια θα υπολογίσουμε ένα 95% διάστημα εμπιστοσύνης, χρησιμοποιώντας τη γνώση της κατανομής του δειγματικού μέσου $\bar{X}$.

Παράδειγμα (συνέχεια)

- Σύμφωνα με το προηγούμενο βασικό θεώρημα, ο δειγματικός μέσος $\bar{X}$ ακολουθεί κανονική κατανομή $N(\mu, \frac{\sigma^2}{n})$. Επομένως η τ.μ.

$$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} = \sqrt{n}\frac{\bar{X}-\mu}{\sigma}$$

ακολουθεί τυπική κανονική κατανομή $N(0,1)$ και άρα (από τους πίνακες)

$$\Pr\{-1.96 < \sqrt{n}\frac{\bar{X}-\mu}{\sigma} < 1.96\} = 0.95$$
$$\Rightarrow \Pr\{\bar{X} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\} = 0.95$$

Επομένως, με "βεβαιότητα" (ή "εμπιστοσύνη") 95% η πραγματική τιμή της μ θα είναι σε εύρος $1.96\,\sigma/\sqrt{n}$ γύρω από τον δειγματικό μέσο. Αν λοιπόν ο δειγματικός μέσος είναι $\bar{X} = \bar{x}$ το 95% διάστημα εμπιστοσύνης για την μ είναι το $(\bar{x} - 1.96\frac{\sigma}{\sqrt{n}}, \bar{x} + 1.96\frac{\sigma}{\sqrt{n}})$.

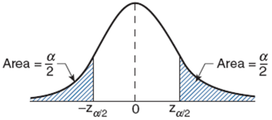
- Στο συγκεκριμένο παράδειγμα το 95% διάστημα εμπιστοσύνης είναι

$$(9 - 1.96\tfrac{\sigma}{3}, 9 + 1.96\tfrac{\sigma}{3}) = (7.69, 10.31)$$

Με άλλα λόγια, είμαστε "κατά 95% βέβαιοι" ότι η πραγματική τιμή θα βρίσκεται μεταξύ 7.69 και 10.31.

Γενίκευση

- Με παρόμοιο τρόπο υπολογίζουμε οποιοδήποτε διάστημα εμπιστοσύνης, από τη σχέση
$$\Pr\{Z > z_a\} = a$$
για Z τυπική κανονική, η οποία οδηγεί στην:
$$\Pr\{-z_{a/2} < Z < z_{a/2}\} = 1 - a$$
και επομένως το $100(1-a)$ διάστημα εμπιστοσύνης για την μ είναι:
$$(\bar{x} - z_{a/2}\frac{\sigma}{\sqrt{n}}, \bar{x} + z_{a/2}\frac{\sigma}{\sqrt{n}})$$



$P\{-z_{\alpha/2} < Z < z_{\alpha/2}\} = 1 - \alpha.$

- π.χ. στο παράδειγμά μας είναι $z_{0.005} = 2.58$ και $2.58\frac{\sigma}{\sqrt{n}} = \frac{5.16}{3} = 1.72$ και επομένως το 99% διάστημα εμπιστοσύνης είναι $\quad 9 \pm 1.72 = (7.28, 10.72)$

Μονόπλευρα διαστήματα εμπιστοσύνης

- Παρομοίως, ξέροντας ότι η $Z = \sqrt{n}\frac{\bar{X}-\mu}{\sigma}$ είναι τυπική κανονική, χρησιμοποιούμε τισ σχέσεις:

$$\Pr\{Z > z_a\} = a \quad \text{και} \quad \Pr\{Z < -z_a\} = a$$

και βρίσκουμε τα εξής μονόπλευρα $100(1-a)$ διαστήματα εμπιστοσύνης:

$$(\bar{x} - z_a \frac{\sigma}{\sqrt{n}}, \infty) \quad \text{και} \quad (-\infty, \bar{x} + z_a \frac{\sigma}{\sqrt{n}})$$

- Παράδειγμα: Είναι $z_{0.01} = 2.33$ και $2.33\frac{\sigma}{\sqrt{n}} = 2.33\frac{2}{3} \simeq 1.55$ και επομένως το 99% άνω διάστημα εμπιστοσύνης είναι

$$(9 - 1.555, \infty) \simeq (7.45, \infty)$$

ενώ το 99% κάτω διάστημα εμπιστοσύνης είναι

$$(-\infty, 9 + 1.555) \simeq (-\infty, 10.55)$$

Το κατάλληλο μέγεθος δείγματος

- Πολλές φορές μας ενδιαφέρει το “αντίστροφο” πρόβλημα, δηλαδή πόσο μεγάλο δείγμα απαιτείται για να έχουμε ένα ορισμένου εύρους διάστημα εμπιστοσύνης.

- π.χ.είναι $z_{0.005} = 2.58$ άρα το 99% διάστημα εμπιστοσύνης για την μ είναι:

$$(\bar{x} - 2.58\frac{\sigma}{\sqrt{n}}, \bar{x} + 2.58\frac{\sigma}{\sqrt{n}})$$

  Άρα, για εύρος διαστήματος π.χ. 0.1, συμπεραίνουμε ότι απαιτείται

$$5.16\frac{\sigma}{\sqrt{n}} = 0.1$$

  και επομένως το απαιτούμενο μέγεθος του δείγματος είναι

$$n = (51.6\sigma)^2$$

Δ. Διάστημα Εμπιστοσύνης Κανονικού Μέσου με Άγνωστη Διασπορά

- Όταν δεν ξέρουμε την πληθυσμιακή διασπορά, δεν μπορούμε να χρησιμοποιήσουμε στην πράξη ότι η $\sqrt{n}(\bar{X} - \mu)/\sigma$ είναι τυπική κανονική. Ωστόσο ξέρουμε από το βασικό θεώρημα ότι $\bar{X}, S^2$ ανεξάρτητες και ότι η τ.μ.:

$$\frac{(n-1)S^2}{\sigma^2}$$

(όπου S^2 η δειγματική διασπορά) ακολουθεί την χ^2 κατανομή με $(n-1)$ βαθμούς ελευθερίας. Από τον ορισμό της t-κατανομής, η τ.μ.

$$\frac{\frac{\sqrt{n}(\bar{X}-\mu)}{\sigma}}{\sqrt{\frac{\frac{(n-1)S^2}{\sigma^2}}{n-1}}} = \sqrt{n}\frac{\bar{X}-\mu}{S}$$

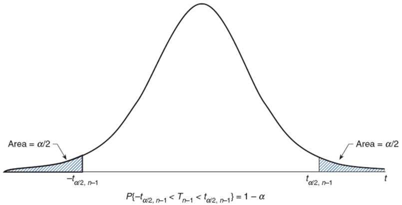
ακολουθεί την κατανομή t με $n-1$ βαθμούς ελευθερίας.

Δ. Διάστημα Εμπιστοσύνης Κανονικού Μέσου με Άγνωστη Διασπορά

- Επομένως, χρησιμοποιούμε τις ιδιότητες της t-κατανομής, και από τη συμμετρία της είναι

$$\Pr\{-t_{a/2,n-1} < \sqrt{n}\frac{\bar{X}-\mu}{S} < t_{a/2,n-1}\} = 1-a$$

$$\Rightarrow \Pr\{\bar{X} - t_{a/2,n-1}\frac{S}{\sqrt{n}} < \mu < \bar{X} + t_{a/2,n-1}\frac{S}{\sqrt{n}}\} = 1-a$$



- Άρα αν ο δειγματικός μέσος όρος και η δειγματική διασπορά είναι $\bar{X} = \bar{x}$ και $S = s$, τότε το $100(1-a)$ διάστημα εμπιστοσύνης για την μ είναι

$$\mu \in (\bar{x} - t_{a/2,n-1}\frac{s}{\sqrt{n}}, \bar{x} + t_{a/2,n-1}\frac{s}{\sqrt{n}})$$

Δ. Γενίκευση σε αρκετά μεγάλα, μη κανονικά δείγματα

- Η μέθοδός μας βασίζεται στο ότι ο πληθυσμός είναι κανονικός. Ακόμα όμως και όταν αυτό δεν ισχύει και το δείγμα ακολουθεί μια οποιαδήποτε κατανομή, το κεντρικό οριακό θεώρημα μας εξασφαλίζει ότι η τ.μ.

$$\sqrt{n}\frac{\bar{X}-\mu}{\sigma}$$

είναι προσεγγιστικά κανονική και επομένως πάλι η τ.μ.

$$\sqrt{n}\frac{\bar{X}-\mu}{S}$$

ακολουθεί προσεγγιστικά την t-κατανομή, και άρα παίρνουμε (για αρκετά μεγάλα δείγματα) μια καλή προσέγγιση.

Ε. Διάστημα Εμπιστοσύνης για τη Διασπορά Κανονικών Πληθυσμών

- Αν $X_1, X_2, \ldots, X_n$ είναι δείγμα από την κανονική κατανομή $N(\mu, \sigma^2)$ κατασκευάζουμε διαστήματα εμπιστοσύνης (όταν η μέση τιμή μ είναι άγνωστη) για την διασπορά σ^2 με βάση το γεγονός ότι:

$$(n-1)\frac{S^2}{\sigma^2} \sim \chi^2_{n-1}$$

- Επομένως, είναι

$$\Pr\{\chi^2_{1-a/2,n-1} \leq (n-1)\frac{S^2}{\sigma^2} \leq \chi^2_{a/2,n-1}\} = 1-a$$

$$\Rightarrow \Pr\left\{\frac{(n-1)S^2}{\chi^2_{a/2,n-1}} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi^2_{1-a/2,n-1}}\right\} = 1-a$$

και αν η δειγματική διασπορά είναι $S^2 = s^2$, το $100(1-a)$ δ.ε. για την σ^2 είναι:

$$\left\{\frac{(n-1)s^2}{\chi^2_{a/2,n-1}}, \frac{(n-1)s^2}{\chi^2_{1-a/2,n-1}}\right\}$$

Σύνοψη

$100(1-\alpha)$ *Percent Confidence Intervals*

$$X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$$

$$\overline{X} = \sum_{i=1}^{n} X_i / n, \qquad S = \sqrt{\sum_{i=1}^{n} (X_i - \overline{X})^2 / (n-1)}$$

Assumption	Parameter	Confidence Interval	Lower Interval	Upper Interval
σ^2 known	μ	$\overline{X} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$	$\left(-\infty, \overline{X} + z_\alpha \frac{\sigma}{\sqrt{n}}\right)$	$\left(\overline{X} + z_\alpha \frac{\sigma}{\sqrt{n}}, \infty\right)$
σ^2 unknown	μ	$\overline{X} \pm t_{\alpha/2, n-1} \frac{S}{\sqrt{n}}$	$\left(-\infty, \overline{X} + t_{\alpha, n-1} \frac{S}{\sqrt{n}}\right)$	$\left(\overline{X} - t_{\alpha, n-1} \frac{S}{\sqrt{n}}, \infty\right)$
μ unknown	σ^2	$\left(\frac{(n-1)S^2}{\chi^2_{\alpha/2, n-1}}, \frac{(n-1)S^2}{\chi^2_{1-\alpha/2, n-1}}\right)$	$\left(0, \frac{(n-1)S^2}{\chi^2_{1-\alpha, n-1}}\right)$	$\left(\frac{(n-1)S^2}{\chi^2_{\alpha, n-1}}, \infty\right)$

ΣΤ. Διαστήματα Εμπιστοσύνης για τη διαφορά των μέσων τιμών δύο κανονικών δειγμάτων

- Έστω $X_1, \ldots, X_n$ δείγμα μεγέθους n από κανονική κατανομή $N(\mu_1, \sigma_1^2)$ και $Y_1, \ldots, Y_m$ δείγμα μεγέθους m από μία άλλη κανονική κατανομή $N(\mu_2, \sigma_2^2)$. Θεωρούμε τα δείγματα ανεξάρτητα. Προσπαθούμε να εκτιμήσουμε τη διαφορά $\mu_1 - \mu_2$ των μέσων τιμών των δύο δειγμάτων.
- Είναι πολύ εύκολο να δείξουμε ότι η διαφορά των δειγματικών μέσων $\bar{X} - \bar{Y}$ είναι συνεπής εκτιμήτρια για την $\mu_1 - \mu_2$, την οποία και χρησιμοποιούμε ως σημειακή εκτίμηση. Για να βρούμε τα διαστή- ματα εμπιστοσύνης χρειαζόμαστε την κατανομή της τ.μ. $\bar{X} - \bar{Y}$.
- Επειδή το άθροισμα κανονικών κατανομών είναι κανονικό, έχουμε

$$\bar{X} \sim N(\mu_1, \frac{\sigma_1^2}{n})$$
$$\bar{Y} \sim N(\mu_2, \frac{\sigma_2^2}{m})$$
$$\Rightarrow \bar{X} - \bar{Y} \sim N(\mu_1 - \mu_2, \frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m})$$
$$\Rightarrow \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}} \sim N(0, 1)$$

ΣΤ. Διαστήματα Εμπιστοσύνης για τη διαφορά των μέσων τιμών δύο κανονικών δειγμάτων (ΙΙ)

- Αν οι πληθυσμιακές διασπορές σ_1^2 και σ_2^2 είναι γνωστές, παίρνουμε επομένως (όπως προηγουμένως) ότι αν $\bar{X} = \bar{x}$ και $\bar{Y} = \bar{y}$ οι δειγματικοί μέσοι, τότε ένα $100(1-a)$ διάστημα εμπιστοσύνης για την $\mu_1 - \mu_2$ είναι

$$\mu_1 - \mu_2 \in \left(\bar{x} - \bar{y} - z_{a/2}\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}, \bar{x} - \bar{y} + z_{a/2}\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}} \right)$$

- Αν οι διασπορές σ_1^2, σ_2^2 είναι άγνωστες τότε (όπως προηγουμένως) προσφεύγουμε στις δειγματικές διασπορές:

$$S_1^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$
$$S_2^2 = \frac{1}{m-1}\sum_{i=1}^{m}(Y_i - \bar{Y})^2$$

και βασίζουμε την κατασκευή του δ.ε. στην τ.μ.:

$$\frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n} + \frac{s_2^2}{m}}}$$

ΣΤ. Διαστήματα Εμπιστοσύνης για τη διαφορά των μέσων τιμών δύο κανονικών δειγμάτων (ΙΙΙ)

- Αυτή ωστόσο η τ.μ. έχει πολύπλοκη κατανομή και στην πραγματικότητα μπορούμε να κατασκευάσουμε δ.ε. μόνο στην περίπτωση που οι πληθυσμιακές διασπορές είναι ίσες ($\sigma_1^2 = \sigma_2^2$). Έστω επομένως ότι οι πληθυσμιακές διασπορές, αν και άγνωστες, είναι ίσες και έστω σ^2 η άγνωστη τιμή τους. Επομένως (όπως προηγουμένως) είναι:

$$(n-1)\frac{S_1^2}{\sigma^2} \sim \chi^2_{n-1}$$

$$(m-1)\frac{S_2^2}{\sigma^2} \sim \chi^2_{m-1}$$

Επειδή τα δείγματα είναι ανεξάρτητα, είναι ανεξάρτητες και οι δύο αυτές χ^2 μεταβλητές, και από την προσθετική ιδιότητα της χ^2 είναι:

$$(n-1)\frac{S_1^2}{\sigma^2} + (m-1)\frac{S_2^2}{\sigma^2} \sim \chi^2_{n+m-2}$$

Επίσης, επειδή $\bar{X} - \bar{Y} \sim N(\mu_1 - \mu_2, \frac{\sigma^2}{n} + \frac{\sigma^2}{m})$ είναι

$$\frac{\bar{X}-\bar{Y}-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma^2}{n}+\frac{\sigma^2}{m}}} \sim N(0,1)$$

ΣΤ. Διαστήματα Εμπιστοσύνης για τη διαφορά των μέσων τιμών δύο κανονικών δειγμάτων (IV)

- Αλλά ξέρουμε από το βασικό θεώρημα ότι οι $\bar{X}$ και S^2 είναι ανεξάρτητες, επομένως είναι ανεξάρτητες και οι $\bar{X}$, $\bar{Y}$, S_1^2, S_2^2. Άρα, από τον ορισμό της t-κατανομής (ως πηλίκο μιας τυπικής κανονικής με παρανομαστή την τετραγωνική ρίζα μιας, ανεξάρτητής της, χ^2 κατανομής προς τον αριθμό του βαθμού ελευθερίας της) ότι αν:

$$S_p^2 = \frac{(n-1)S_1^2+(m-1)S_2^2}{n+m-2}$$

τότε η τ.μ.

$$\frac{\bar{X}-\bar{Y}-(\mu_1-\mu_2)}{\sqrt{\sigma^2\left(\frac{1}{n}+\frac{1}{m}\right)}} \div \sqrt{\frac{S_p^2}{\sigma^2}} = \frac{\bar{X}-\bar{Y}-(\mu_1-\mu_2)}{\sqrt{S_p^2\left(\frac{1}{n}+\frac{1}{m}\right)}}$$

έχει κατανομή t με $n+m-2$ βαθμούς ελευθερίας. Επομένως, αν οι δειγματικοί μέσοι είναι $\bar{X}=\bar{x}$, $\bar{Y}=\bar{y}$, $S_p=s_p$ τότε το $100(1-a)$ διάστημα εμπιστοσύνης για την $\mu_1-\mu_2$ είναι:

$$\left(\bar{x}-\bar{y}-t_{a/2,n+m-2}\cdot s_p\sqrt{\frac{1}{n}+\frac{1}{m}}\ ,\ \bar{x}-\bar{y}+t_{a/2,n+m-2}\cdot s_p\sqrt{\frac{1}{n}+\frac{1}{m}}\right).$$

Σύνοψη (V)

$100(1-\sigma)$ *Percent Confidence Intervals for* $\mu_1 - \mu_2$

$$X_1, \ldots, X_n \sim N(\mu_1, \sigma_1^2)$$
$$Y_1, \ldots, Y_m \sim N(\mu_2, \sigma_2^2)$$

$$\overline{X} = \sum_{i=1}^{n} X_i/n, \qquad S_1^2 = \sum_{i=1}^{n} (X_i - \overline{X})^2/(n-1)$$

$$\overline{Y} = \sum_{i=1}^{m} Y_i/n, \qquad S_2^2 = \sum_{i=1}^{m} (Y_i - \overline{Y})^2/(m-1)$$

Assumption	Confidence Interval
σ_1, σ_2 known	$\overline{X} - \overline{Y} \pm z_{\alpha/2}\sqrt{\sigma_1^2/n + \sigma_2^2/m}$
σ_1, σ_2 unknown but equal	$\overline{X} - \overline{Y} \pm t_{\alpha/2, n+m-2}\sqrt{\left(\frac{1}{n} + \frac{1}{m}\right)\frac{(n-1)S_1^2 + (m-1)S_2^2}{n+m-2}}$

Assumption	Lower Confidence Interval
σ_1, σ_2 known	$(-\infty, \overline{X} - \overline{Y} + z_{\alpha}\sqrt{\sigma_1^2/n + \sigma_2^2/m})$
σ_1, σ_2 unknown but equal	$\left(-\infty, \overline{X} - \overline{Y} + t_{\alpha, n+m-2}\sqrt{\left(\frac{1}{n} + \frac{1}{m}\right)\frac{(n-1)S_1^2 + (m-1)S_2^2}{n+m-2}}\right)$

Note: Upper confidence intervals for $\mu_1 - \mu_2$ *are obtained from lower confidence intervals for* $\mu_2 - \mu_1$.