Πιθανότητες και Στατιστική

Διαστήματα εμπιστοσύνης

Κώστας Μανές

Τμήμα Πληροφορικής, Πανεπιστήμιο Πειραιώς

2021-2022

Διαστήματα εμπιστοσύνης

Είναι σχεδόν βέβαιο ότι η τιμή μιας εκτιμήτριας T μιας παραμέτρου θ δεν ταυτίζεται ακριβώς με την πραγματική τιμή της παραμέτρου θ. Επομένως, είναι πιο χρήσιμο, αντί μιας συγκεκριμένης τιμής, να δοθεί ένα διάστημα εντός του οποίου βρίσκεται η πραγματική τιμή της παραμέτρου θ, με κάποια πιθανότητα. Το διάστημα αυτό ονομάζεται **διάστημα εμπιστοσύνης** (δ.ε.) για την παράμετρο θ. Συγκεκριμένα, αν

$$P(\theta_1 \leq \theta \leq \theta_2) = 1 - a$$

τότε το διάστημα $[\theta_1, \theta_2]$ ονομάζεται $(1-a)100\%$ **διάστημα εμπιστοσύνης για την παράμετρο** θ.
Η πιθανότητα $1-a$ ονομάζεται **επίπεδο εμπιστοσύνης** και η πιθανότητα a ονομάζεται **επίπεδο σημαντικότητας**.
Συνήθως αναζητούμε διαστήματα εμπιστοσύνης τα οποία είναι συμμετρικά ως προς την σημειακή εκτίμηση T της θ, δηλαδή είναι της μορφής $[T-\epsilon, T+\epsilon]$ (διότι αποδεικνύεται ότι τα διαστήματα αυτά έχουν ελάχιστο μήκος).

Διαστήματα εμπιστοσύνης

Στον προσδιορισμό διαστημάτων εμπιστοσύνης κεντρικό ρόλο παίζει η έννοια του ποσοστιαίου σημείου μια κατανομής:

Ορισμός

Το άνω a-ποσοστιαίο σημείο, $a \in (0,1)$, μιας ΤΜ X που ακολουθεί κάποια κατανομή F συμβολίζεται με x_a και ορίζεται από τις ισοδύναμες ισότητες:

$$P(X > x_a) = a \Leftrightarrow 1 - P(X \leq x_a) = a \Leftrightarrow 1 - F(x_a) = a \Leftrightarrow F(x_a) = 1 - a$$
$$\Leftrightarrow x_a = F^{-1}(1-a)$$

Γενικά, αν η F δεν είναι αντιστρέψιμη, τότε

$$x_a = \inf\{x \in S_X : F(x) \geq 1 - a\}.$$

Παρατήρηση: Ορισμένοι συγγραφείς ορίζουν το x_a βάσει των σχέσεων $P(X \geq x_a) = a$ και $P(X \leq x_a) = 1 - a$. Ο ορισμός αυτός ταυτίζεται με τον προηγούμενο όταν η ΤΜ X είναι συνεχής. Όταν είναι διακριτή, τότε μπορεί το x_a να μην ανήκει στο S_X. Για παράδειγμα, αν η X είναι διακριτή ομοιόμορφη στο $S_X = \{1, 2, 3, 4\}$ και $a = 0.5$, τότε με τον πρώτο ορισμό είναι $x_a = 2$, ενώ με τον δεύτερο μπορεί να είναι οποιοσδήποτε $x_a \in (2, 3)$ και συνήθως θεωρείται ο $x_a = 2.5$.

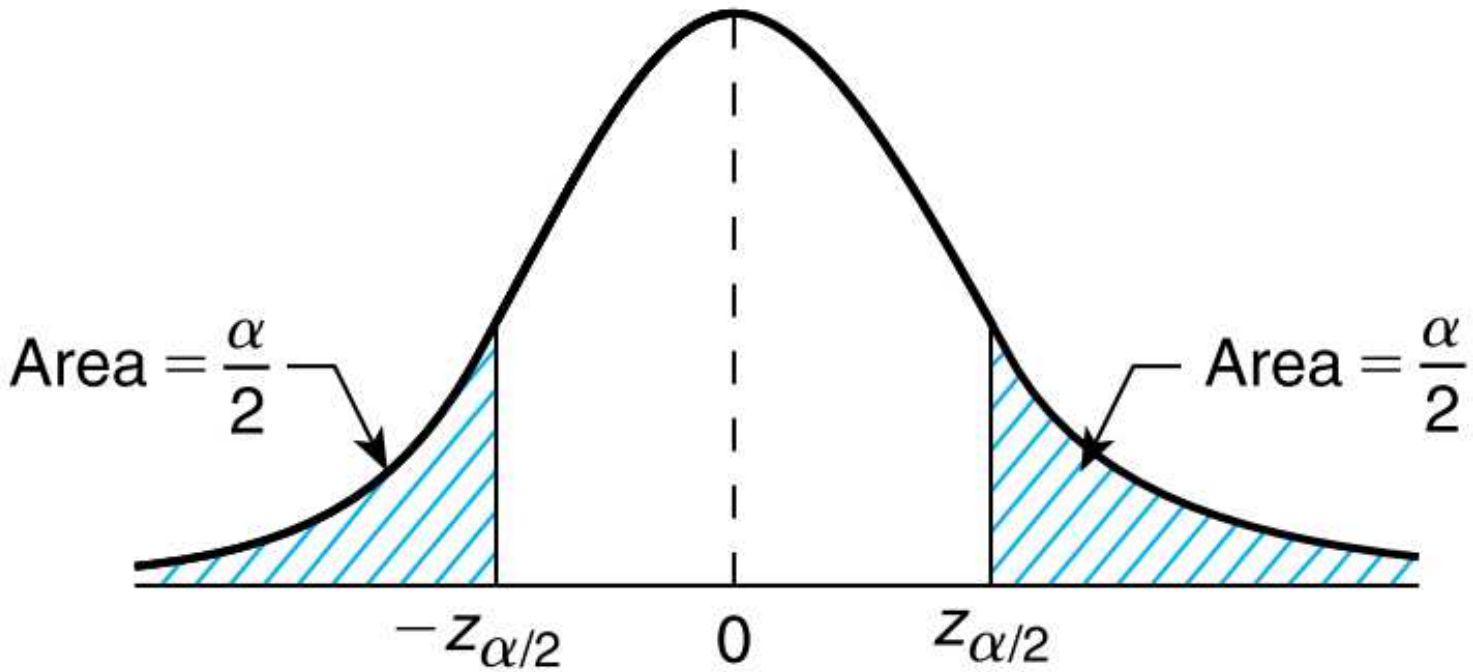
Διαστήματα εμπιστοσύνης

Για την τυπική κανονική κατανομή, το άνω a-ποσοστιαίο σημείο συμβολίζεται με z_a, δηλαδή

$$z_a := \Phi^{-1}(1-a), \quad a \in (0,1).$$

Συνηθισμένες τιμές:

a	0.0005	0.001	0.005	0.01	0.025	0.05	0.10
z_a	3.29	3.09	2.576	2.326	1.960	1.645	1.282



Διαστήματα εμπιστοσύνης

Το άνω a-ποσοστιαίο σημείο μιας κατανομής υπολογίζεται στην Python με οποιαδήποτε από τις συναρτήσεις isf(a) και ppf(1-a). Ο επόμενος κώδικας υπολογίζει τις τιμές του προηγούμενου πίνακα:

```
from scipy import stats as st
vals = [0.0005, 0.001, 0.005, 0.01, 0.025, 0.05, 0.1]
for a in vals:
    print("a = %s, z_a = %s, z_a = %s"%(a,st.norm.isf(a), st
    .norm.ppf(1-a)))
```

Δ.Ε. για τον πληθυσμιακό μέσο

Πρόταση

Αν $\overline{X}$ είναι ο δειγματικός μέσος ενός τυχαίου δείγματος μεγέθους n από ένα πληθυσμό που ακολουθεί την κανονική κατανομή $N(\mu, \sigma^2)$, όπου σ^2 γνωστό, τότε ένα $(1-a)\cdot 100\%$ διάστημα εμπιστοσύνης για τον μέσο όρο μ είναι το $[\overline{X} \pm B] := \left[\overline{X} - B, \overline{x} + B\right]$, όπου $B = z_{a/2}\dfrac{\sigma}{\sqrt{n}}$.

Απόδειξη.

Ως γνωστό, η ΤΜ $Z = \dfrac{\overline{X} - \mu}{\sigma/\sqrt{n}}$ ακολουθεί την $N(0,1)$, οπότε

$$P(|Z| \le z) \ge 1 - a \Leftrightarrow \Phi(z) - \Phi(-z) \ge 1 - a \Leftrightarrow 2\Phi(z) - 1 \ge 1 - a$$
$$\Leftrightarrow \Phi(z) \ge \frac{2-a}{2} = 1 - a/2 \Leftrightarrow z \ge z_{a/2}$$

Επομένως,

$$1 - a = P(|Z| \le z_{a/2}) = P(|\overline{X} - \mu| \le B) = P(\mu \in [\overline{X} \pm B]) \qquad \square$$

Παρατηρήσεις:

- Η τελευταία πιθανότητα $P(\mu \in [\overline{X} \pm B]) = P(\overline{X} \in [\mu \pm B]) = 1 - a$ που υπολογίσθηκε στην παραπάνω απόδειξη είναι η πιθανότητα το διάστημα $[\overline{X} \pm B]$ που κατασκευάσαμε βάσει της παρατήρησης να περιέχει το μ. Δηλαδή, η ΤΜ είναι το διάστημα και όχι ο μ, ο οποίος θεωρείται σταθερός, αν και άγνωστος.
- Από την παράσταση $B = z_{a/2}\frac{\sigma}{\sqrt{n}}$ μπορούμε να υπολογίσουμε ένα εκ των a, B, n, αν δίνονται τα άλλα δύο. Για παράδειγμα, αν δίνονται τα a και B, τότε μπορούμε να υπολογίσουμε το απαιτούμενο μέγεθος δείγματος $n = z^2_{a/2}\frac{\sigma^2}{B^2}$ ώστε να επιτύχουμε το απαιτούμενο επίπεδο σημαντικότητας και πλάτος διαστήματος.

Δ.Ε. για τον πληθυσμιακό μέσο

Παράδειγμα

Έστω ένα τυχαίο δείγμα 4 μετρήσεων $[1.2, 3.4, 0.6, 5.6]$ από ένα πληθυσμό που ακολουθεί την κανονική κατανομή $N(\mu, 9)$. Να βρεθεί ένα 90% διάστημα εμπιστοσύνης για την μέση τιμή μ.

Λύση

Βάσει της εκφώνησης, είναι $n = 4$ και $a = 0.1$, οπότε $z_{a/2} = 1.645$.
Μια σημειακή εκτίμηση για την μέση τιμή μ είναι η

$$\overline{X} = \frac{1.2 + 3.4 + 0.6 + 5.6}{4} = 2.7.$$

Βάσει της προηγούμενης πρότασης, είναι

$$B = z_{a/2}\frac{\sigma}{\sqrt{n}} = 1.645\frac{3}{2} = 2.4675$$

και το ζητούμενο διάστημα είναι το $[\overline{X} \pm B] = [0.2325, 5.1675]$.

Δ.Ε. για τον πληθυσμιακό μέσο

Παράδειγμα

Έστω ένα δείγμα μεγέθους $n = 100$ έχει δειγματικό μέσο $\overline{X} = 17$ και ο πληθυσμός απ' όπου επιλέξαμε το δείγμα ακολουθεί την κανονική κατανομή με διακύμανση $\sigma^2 = 9$. Να βρεθεί ένα διάστημα εμπιστοσύνης για την μέση τιμή μ του πληθυσμού με επίπεδο εμπιστοσύνης

$$i)\ 1 - a = 0.95, \qquad ii)\ 1 - a = 0.99$$

Λύση

$i)$ Για $a = 0.05$, είναι $z_{a/2} = 1.96$, οπότε

$$B = z_{a/2}\sigma/\sqrt{n} = 1.96 \cdot 3/10 = 0.588,$$

άρα $[\overline{X} \pm B] = [17 \pm 0.588] = [16.412, 17.588]$.

$ii)$ Για $a = 0.01$, είναι $z_{a/2} = 2.576$, οπότε

$$B = z_{a/2}\sigma/\sqrt{n} = 2.576 \cdot 3/10 = 0.7728,$$

άρα $[\overline{X} \pm B] = [17 \pm 0.7728] = [16.2272, 17.7728]$.

Δ.Ε. για τον πληθυσμιακό μέσο

Το δ.ε., για $a = 0.01$, του προηγούμενου παραδείγματος μπορεί να υπολογισθεί απευθείας με τη συνάρτηση interval(1-a) της βιβλιοθήκης scipy.stats

```
import numpy as np
import scipy.stats as st
xbar, a, sigma, N = 17, 0.01, 3, 100
se = sigma/np.sqrt(N) #standard error
c = st.norm.interval(alpha = 1-a, loc = xbar, scale = se)
print("CI (for a = %s):"%a, c)
```

Output:

```
CI (for a = 0.01): (16.22725120893533, 17.77274879106467)
```

Παράδειγμα

Ένας πληθυσμός ακολουθεί την κανονική κατανομή με μέση τιμή μ και διακύμανση $\sigma^2 = 16$. Να βρεθεί το ελάχιστο μέγεθος n ενός τυχαίου δείγματος που θα μας επιτρέψει να εκτιμήσουμε την μέση τιμή μ με

i) επίπεδο εμπιστοσύνης 0.95 και σφάλμα ± 1.

ii) επίπεδο εμπιστοσύνης 0.95 και σφάλμα ± 0.5.

iii) επίπεδο εμπιστοσύνης 0.99 και σφάλμα ± 1.

Λύση

i) Για $a = 0.05$ και $B = 1$, είναι $z_{a/2} = 1.96$ και

$$B = z_{a/2}\sigma/\sqrt{n} \leq 1 \Leftrightarrow n \geq z_{a/2}^2\sigma^2 = 1.96^2 \cdot 16 = 61.4656 \Leftrightarrow n \geq 62.$$

Λύση (συνέχεια)

ii) Για $a = 0.05$ και $B = 0.5$, είναι $z_{a/2} = 1.96$ και

$$B = z_{a/2}\sigma/\sqrt{n} \leq 0.5 \Leftrightarrow n \geq \frac{z_{a/2}^2\sigma^2}{0.5^2} = \frac{1.96^2 \cdot 16}{0.25} = 4 \cdot 61.4656 = 245.8624$$
$$\Leftrightarrow n \geq 246.$$

iii) Για $a = 0.01$ και $B = 1$, είναι $z_{a/2} = 2.576$ και

$$B = z_{a/2}\sigma/\sqrt{n} \leq 1 \Leftrightarrow n \geq \frac{z_{a/2}^2\sigma^2}{B^2} = 2.576^2 \cdot 16 = 106.172416 \Leftrightarrow n \geq 107.$$

Δ.Ε. για τον πληθυσμιακό μέσο

Παράδειγμα

Σε ένα εργοστάσιο εμφιαλώσεως νερού παρατηρήθηκε ότι η ποσότητα X του νερού σε κάθε μπουκάλι ακολουθεί την κανονική κατανομή $N(\mu, 1.5^2)$. Ένας έλεγχος σε δείγμα $n = 25$ μπουκαλιών έδωσε $\overline{x} = 499.28$. Να βρεθεί ένα 95% διάστημα εμπιστοσύνης για το μ.

Λύση

Για $a = 0.05$, είναι $z_{a/2} = 1.96$, οπότε

$$B = z_{a/2}\sigma/\sqrt{n} = 1.96 \cdot 1.5/5 = 0.588$$

και το ζητούμενο δ.ε. είναι το $[499.28 \pm 0.588]$.

Δ.Ε. για τον πληθυσμιακό μέσο

```
import numpy as np
import scipy.stats as st
xbar, a, sigma, N = 499.28, 0.05, 1.5, 25
se = sigma/np.sqrt(N) #standard error
c = st.norm.interval(alpha = 1-a, loc = xbar, scale = se)
print("CI (for a = %s):"%a, c)
```

Output:

```
CI (for a = 0.05): (498.69201080463796, 499.867989195362)
```

Δ.Ε. για τον πληθυσμιακό μέσο όταν σ^2 άγνωστη

Πρόταση (Δ.ε. μέσης τιμής όταν σ^2 άγνωστη)

Έστω τυχαίο δείγμα $(X_1, X_2, \ldots, X_n)$ από κανονικό πληθυσμό $N(\mu, \sigma^2)$ με άγνωστή αλλά πεπερασμένη διακύμανση σ^2. Η στατιστική συνάρτηση

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}}$$

ακολουθεί την κατανομή t_{n-1} και ένα $(1-a)\cdot 100\%$ διάστημα εμπιστοσύνης για την μέση τιμή μ είναι το

$$\overline{X} \pm B,$$

όπου $B = t_{n-1,a/2}\dfrac{S}{\sqrt{n}}$ και $t_{n-1,a/2}$ το $a/2$-άνω ποσοστιαίο σημείο της κατανομής t με $n-1$ βαθμούς ελευθερίας.

Απόδειξη.

Έχει ήδη αποδειχθεί ότι $T \sim t_{n-1}$ και ότι η συνάρτηση κατανομής F της T είναι συμμετρική, δηλαδή $F(t) + F(-t) = 1$. Επομένως, όπως και στην προηγούμενη απόδειξη, προκύπτει ότι

$$1 - a = P(|T| \leq t_{n-1,a/2}) = P(|\overline{X} - \mu| \leq B) = P(\mu \in [\overline{X} \pm B]) \qquad \square$$

Παρατήρηση: Για μεγάλο n ($n \geq 30$), είναι $T \to N(0,1)$ και ένα προσεγγιστικό $(1-a)\cdot 100\%$ δ.ε. για τη μέση τιμή είναι το

$$[\overline{X} \pm B], \qquad \text{όπου } B = z_{a/2}\frac{S}{\sqrt{n}}.$$

Η προσέγγιση αυτή μπορεί να χρησιμοποιηθεί για οποιαδήποτε κατανομή δείγματος, όταν το n είναι μεγάλο.

Δ.Ε. για τον πληθυσμιακό μέσο όταν σ^2 άγνωστη

Παράδειγμα

Ένας δείκτης της κυκλοφορίας οχημάτων είναι ο αριθμός χιλιομέτρων που κάνει ένα όχημα το χρόνο. Σε μια περιοχή επιλέχθηκε ένα τυχαίο δείγμα 200 αυτοκινήτων και καταγράφηκε για κάθε αυτοκίνητο ο αριθμός των χιλιομέτρων που διένυσε τον τελευταίο χρόνο και βρέθηκε ότι $\overline{X} = 14500$ και $S = 4000$.

i) Να βρεθεί ένα 95% διάστημα εμπιστοσύνης για τον μέσο αριθμό χιλιομέτρων που διανύει το χρόνο ένα αυτοκίνητο.

ii) Να βρεθεί ένα 99% διάστημα εμπιστοσύνης για τον ίδιο αριθμό και να συγκριθούν τα αποτελέσματα.

Δ.Ε. για τον πληθυσμιακό μέσο όταν σ^2 άγνωστη

Λύση

Έχουμε ότι $n = 200$, $\overline{X} = 14500$, $S = 4000$ και η διασπορά είναι άγνωστη. Θεωρούμε τη στατιστική $T = \dfrac{\overline{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$.

i) Για $a = 0.05$ είναι $t_{n-1,a/2} = 1.972$, οπότε

$$B = t_{n-1,a/2} S/\sqrt{n} = \frac{1.972 \cdot 4000}{\sqrt{200}} = \frac{1.972 \cdot 400}{\sqrt{2}} = 557.76583$$

και το ζητούμενο 95% δ.ε. είναι το

$$[14500 \pm 557.76583]$$

Αν χρησιμοποιηθεί η προσέγγιση από κανονική κατανομή, τότε είναι

$$B' = z_{a/2} S/\sqrt{n} = \frac{1.96 \cdot 400}{\sqrt{2}} = 554.3717.$$

Λύση (συνέχεια)

ii) Για $a = 0.01$ είναι $t_{n-1,a/2} = 2.60076$, οπότε

$$B = t_{n-1,a/2} S/\sqrt{n} = \frac{2.60076 \cdot 4000}{\sqrt{200}} = \frac{2.60076 \cdot 400}{\sqrt{2}} = 735.606$$

και το ζητούμενο 99% δ.ε. είναι το

$$[14500 \pm 735.606]$$

Δ.Ε. για τον πληθυσμιακό μέσο όταν σ^2 άγνωστη

```
import numpy as np
import scipy.stats as st
a, n, xbar, s = 0.05, 200, 14500, 4000
df = n-1
q = st.t.ppf(1-a/2,df)
qz = st.norm.ppf(1-a/2)
print("confidence level = %s, sample mean = %s"%(1-a, xbar))
print("t_{n-1,a/2}=%s, P(T>=t_{n-1,a/2}) = %s"%(q, 1-st.t.
    cdf(q,df)))
print("z_{a/2}=%s, P(Z>=z_{a/2}) = %s"%(qz, 1-st.norm.cdf(qz
    )))
print("(1-a)100% CI (using t):", st.t.interval(alpha = 1-a,
    df = n-1, loc = xbar, scale = s/np.sqrt(n)))
print("(1-a)100% CI (using normal):", st.norm.interval(alpha
     = 1-a, loc = xbar, scale = s/np.sqrt(n)))
```

Δ.Ε. για τον πληθυσμιακό μέσο όταν σ^2 άγνωστη

Output:

```
confidence level = 0.95, sample mean = 14500
t_{n-1,a/2}=1.971956544249395, P(T>=t_{n-1,a/2}) =
    0.02500000000013569
z_{a/2}=1.959963984540054, P(Z>=z_{a/2}) =
    0.02500000000000022
(1-a)100% CI (using t): (13942.246462142424,
    15057.753537857576)
(1-a)100% CI (using normal): (13945.638470260128,
    15054.361529739872)
```

Δ.Ε. για τον πληθυσμιακό μέσο όταν σ^2 άγνωστη

Παράδειγμα

Ένα κτηματολογικό γραφείο θέλει να εκτιμήσει την μέση τιμή των σπιτιών μιας περιοχής. Ένα δείγμα 25 σπιτιών έδωσε δειγματικό μέσο $\overline{X} = 50$ και διακύμανση $S^2 = 64$. Να βρεθεί ένα 90% διάστημα εμπιστοσύνης για το μ.

Λύση

Για $a = 0.1$ και $n = 25$, είναι $t_{n-1,a/2} = 1.711$, οπότε $B = t_{n-1,a/2} S/\sqrt{n} = 1.711 \cdot 8/5 = 2.7376$, και το ζητούμενο δ.ε. είναι το $[\overline{X} \pm B] = [50 \pm 2.7376]$.

Δ.Ε. για πληθυσμιακό ποσοστό

Πρόταση (Δ.Ε. ποσοστού όταν το δείγμα είναι μεγάλο)

Έστω τυχαίο δείγμα $(X_1, X_2, \ldots, X_n)$ από τον πληθυσμό $\text{Bernoulli}(p)$. Τότε,

$$Z = \frac{\overline{P} - p}{\sqrt{\overline{P}(1-\overline{P})/n}} \to N(0,1)$$

και ένα προσεγγιστικό $(1-a)\cdot 100\%$ διάστημα εμπιστοσύνης για το ποσοστό p είναι το

$$[\overline{P} \pm B], \qquad \text{όπου } B = z_{a/2}\sqrt{\overline{P}(1-\overline{P})/n}.$$

Απόδειξη.

Ως γνωστό, $X = X_1 + \cdots + X_n \sim \mathrm{Binom}(\mathrm{np}, \mathrm{np}(1-\mathrm{p}))$, οπότε η $\overline{P} = X/n$ έχει μέση τιμή p και διακύμανση $\sigma^2 = p(1-p)/n$, οι οποίες είναι άγνωστες.
Θέτοντας $\overline{S}^2 = \overline{P}(1-\overline{P})/n$, ως εκτιμήτρια της σ^2, έχουμε ότι

$$E(\overline{S}^2) = \frac{1}{n}E(\overline{P}(1-\overline{P})) = \frac{1}{n}(E(\overline{P}) - E(\overline{P}^2)) = \frac{1}{n}(p - \frac{p(1-p)}{n} - p^2)$$
$$= \frac{1}{n^2}(np - p(1-p) - np^2) = \frac{n-1}{n^2}p(1-p) = \frac{n-1}{n}\sigma^2 \approx \sigma^2$$

Επειδή, βάσει του ΚΟΘ, είναι $\overline{P} \to N(p, \sigma^2)$ και
$Z' = (\overline{P} - p)/\sigma \to N(0,1)$, η ζητούμενη ακτίνα του δ.ε. θα είναι
$B \approx z_{a/2}\sigma \approx z_{a/2}\sqrt{\overline{P}(1-\overline{P})/n}$. □

Δ.Ε. για πληθυσμιακό ποσοστό

Παράδειγμα

Από τα προϊόντα μιας μηχανής λαμβάνεται τυχαίο δείγμα μεγέθους $n = 125$. Σε αυτά υπάρχουν 7 ελαττωματικά. Να βρεθεί ένα 98% διάστημα εμπιστοσύνης του ποσοστού των ελαττωματικών προϊόντων που παράγονται από την μηχανή.

Λύση

Για $a = 0.02$ και $\overline{P} = 7/125 = 0.056$, είναι $z_{a/2} = 2.326$, οπότε $B = z_{a/2}\sqrt{\overline{P}(1-\overline{P})/n} = 0.047833$.

Δ.Ε. για πληθυσμιακό ποσοστό

```
import numpy as np
import scipy.stats as st
a, p, n = 0.02, 7/125, 125
s = np.sqrt(p*(1-p)/n)
print("%d"%((1-a)*100)+"% CI (using normal):", st.norm.
    interval(alpha = 1-a, loc = p, scale = s))
```

Output:

```
98% CI (using normal): (0.00815906462425664,
    0.10384093537574336)
```

Πρόταση (Δ.Ε. για τη διακύμανση σ^2 όταν μ γνωστό)

Έστω τυχαίο δείγμα $X_1, X_2, \ldots, X_n$ από κανονικό πληθυσμό $N(\mu, \sigma^2)$ με γνωστή μέση τιμή μ.
Η στατιστική συνάρτηση

$$Y = \frac{1}{\sigma^2} \sum_{i=1}^{n} (X_i - \mu)^2$$

ακολουθεί την χ^2_n και ένα $(1-a)\cdot 100\%$ διάστημα εμπιστοσύνης για την διασπορά σ^2 είναι το

$$\frac{\sum_{i=1}^{n} (X_i - \mu)^2}{\chi^2_{n,a/2}} \leq \sigma^2 \leq \frac{\sum_{i=1}^{n} (X_i - \mu)^2}{\chi^2_{n,(1-a/2)}}$$

Απόδειξη.

Οι ΤΜ $Z_i = (X_i - \mu)/\sigma$ είναι ανεξάρτητες και ακολουθούν την $N(0,1)$, άρα $Y = \sum_{i=1}^{n} Z_i^2 \sim \chi^2_n$.
Για το δ.ε., αναζητάμε $y_1 < y_2$, με $P(Y > y_2) = a/2 \Leftrightarrow y_2 = \chi^2_{n,a/2}$ και

$$P(Y < y_1) = a/2 \Leftrightarrow P(Y \geq y_1) = 1 - a/2 \Leftrightarrow y_1 = \chi^2_{n,1-a/2}$$

Επομένως,

$$1 - a = P(y_1 \leq Y \leq y_2) = P(\frac{1}{y_2} \leq \frac{1}{Y} \leq \frac{1}{y_1})$$

$$= P\left(\frac{\sum_{i=1}^{n}(X_i - \mu)^2}{\chi^2_{n,a/2}} \leq \sigma^2 \leq \frac{\sum_{i=1}^{n}(X_i - \mu)^2}{\chi^2_{n,(1-a/2)}} \right).$$

□

Πρόταση (Δ.Ε. για τη διακύμανση σ^2 όταν μ άγνωστο)

Έστω τυχαίο δείγμα $X_1, X_2, \ldots, X_n$ από κανονικό πληθυσμό $N(\mu, \sigma^2)$ με γνωστή μέση τιμή μ. Η στατιστική συνάρτηση

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

ακολουθεί την χ^2_{n-1} και ένα $(1-a)\cdot 100\%$ διάστημα εμπιστοσύνης για την σ^2 είναι το

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

Απόδειξη.

Ομοίως με την προηγούμενη (άσκηση). □

Δ.Ε. για την πληθυσμιακή διακύμανση

Παράδειγμα

Ο υπεύθυνος για τον ποιοτικό έλεγχο των προϊόντων μια εταιρείας ενδιαφέρεται να εκτιμήσει την διακύμανση των μηκών των μεταλλικών ράβδων που παράγει μια νέα μηχανή προκειμένουν να διαπιστώσει ότι ικανοποιούν τις προδιαγραφές. Για το σκοπό αυτό, λαμβάνει τυχαίο δείγμα 25 ράβδων για το οποίο υπολογίζει ότι $s = 1.1$ cm. Να βρεθεί ένα 99% διάστημα εμπιστοσύνης για την διακύμανση του μήκους των ράβδων που παράγει η μηχανή.

Λύση

Ο μέσος είναι άγνωστος, οπότε θεωρούμε την ΤΜ $Y = (n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$. Για $a = 0.01$, είναι $\chi^2_{n-1,a/2} = 45.5585$ και $\chi^2_{n-1,1-a/2} = 9.886$, οπότε το ζητούμενο δ.ε. είναι το

$$[(n-1)S/\chi^2_{n-1,a/2}, (n-1)S/\chi^2_{n-1,1-a/2}] = [0.58, 2.67]$$

Δ.Ε. για την πληθυσμιακή διακύμανση

```
import numpy as np
import scipy.stats as st
a, n, s  = 0.01, 25, 1.1
df = n-1
u = st.chi2.ppf(1-a/2, df)
l = st.chi2.ppf(a/2, df)
#(l,u) = st.chi2.interval(1-a, df)
print(l,u)
print((n-1)*s*s/u,(n-1)*s*s/l)
```

Output:

```
9.886233502241467 45.558511936530586
0.6374220483859704 2.937417975553165
```

Παράδειγμα

Ένα δείγμα μεγέθους $n = 15$ έδωσε $S^2 = 17.2$. Να βρεθεί ένα 90% διάστημα εμπιστοσύνης για την σ^2.

Λύση

Η μέση τιμή είναι άγνωστη, οπότε θα χρησιμοποιηθεί η στατιστική: $Y = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$. Το $(1-a)100\%$ δ.ε. είναι το $[(n-1)s^2/\chi^2_{a/2,n-1}, (n-1)s^2/\chi^2_{1-a/2,n-1}]$. Για $n = 15$, $s^2 = 17.2$, $a = 0.1$, είναι $\chi^2_{n-1,a/2} = 23.685$ και $\chi^2_{n-1,1-a/2} = 6.571$ και τελικά το ζητούμενο δ.ε. είναι το

$$[\frac{14 \cdot 17.2}{23.685}, \frac{14 \cdot 17.2}{6.571}] = [10.166, 36.645]$$

Δ.Ε. για την πληθυσμιακή διακύμανση

```
import numpy as np
import scipy.stats as st
a, n, s2  = 0.1, 15, 17.2
df=n-1
(l,u) = st.chi2.interval(1-a, df)
print(l,u)
print((n-1)*s2/u,(n-1)*s2/l)
```

Output:

```
6.57063138378934 23.684791304840576
10.16686180176671 36.64792406314057
```

Παράδειγμα

Μια μηχανή έχει ρυθμιστεί να παράγει προϊόντα μέσου βάρους $18kg$. Δείγμα $n = 9$ προϊόντων έδωσε

$$18.2,\ 17.9,\ 18.3,\ 18,\ 18.3,\ 17.9,\ 18.1,\ 17.9,\ 18.2$$

Να βρεθεί ένα 95% διάστημα εμπιστοσύνης για την διασπορά του βάρους των προϊόντων.

Λύση

Επειδή δίνεται $\mu = 18$, χρησιμοποιούμε τη στατιστική συνάρτηση $Y = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \mu)^2 \sim \chi^2_n$.

Δ.Ε. για την πληθυσμιακή διακύμανση

```
import numpy as np
import scipy.stats as st
data = np.array([18.2,  17.9,  18.3,  18,  18.3,  17.9,
    18.1,  17.9,  18.2])
a, n, mu = 0.05, len(data), 18
d2 = data - 18
sse = np.dot(d2,d2)
(l1,u1) = st.chi2.interval(1-a, n)
print("chi2 points:",(l1,u1))
print("%d%% CI for sigma^2:"%(100*(1-a)), (np.sqrt(sse/u1),
    np.sqrt(sse/l1)))
```

Output:

```
chi2 points: (2.7003894999803584, 19.022767798641 63)
95% CI for sigma^2: (0.1255809528608859, 0.333309292725 6988)
```

Δ.Ε. για την πληθυσμιακή διακύμανση

Παράδειγμα

Οι χωρητικότητες 10 μπαταριών μετρήθηκαν ως εξής:

$$140, 136, 150, 144, 148, 152, 138, 141, 143, 151.$$

i) Να υπολογισθεί ένα 99% δ.ε. για την πληθυσμιακή διακύμανση σ^2.

ii) Να υπολογισθεί μια τιμή v ώστε να είναι $\sigma^2 < v$ με 90% εμπιστοσύνη.

Λύση

Θεωρούμε τη στατιστική συνάρτηση $Y = (n-1)S^2/\sigma^2$.

ii) Είναι

$$P(\sigma^2 < v) \geq 0.9 \Leftrightarrow P(Y > (n-1)S^2/v) \geq 0.9 \Leftrightarrow (n-1)S^2/v \leq \chi^2_{n-1,0.9}$$
$$\Leftrightarrow v \geq (n-1)S^2/\chi^2_{n-1,0.9},$$

$S^2 = 32.23$ και $\chi^2_{n-1,0.9} = 4.168$, οπότε $v = 69.6$.

Δ.Ε. για την πληθυσμιακή διακύμανση

```
import numpy as np
import scipy.stats as st
data = np.array([140, 136, 150, 144, 148, 152, 138, 141,
    143, 151])
a, n = 0.01, len(data)
S2 = st.tvar(data)
(l,u) = st.chi2.interval(1-a, n-1)
print("sample variance S^2 =", S2)
print("chi2 points:",(l,u))
print("%d%% CI for sigma^2:"%(100*(1-a)), ((n-1)*S2/u, (n-1)
    *S2/l))
p = 0.9
q = st.chi2.ppf(1-p, n-1) #q = 0.1-lower percent point
print("P(chi^2 < %s) = %s" %(q, st.chi2.cdf(q, n-1)))
print("P(sigma^2 < %s) = %s" %((n-1)*S2/q,p))
```

Output:

```
sample variance S^2 = 32.23333333333333
chi2 points: (1.7349329049966606, 23.589350781257387)
99% CI for sigma^2: (12.297922172173351, 167.2110772494446)
P(chi^2 < 4.168159008146107) = 0.09999999999999999
P(sigma^2 < 69.59907226020852) = 0.9
```

Πρόταση (Δ.Ε. διαφοράς δύο μέσων τιμών)

Έστω δύο ανεξάρτητα τυχαία δείγματα $(X_1, X_2, \ldots, X_{n_1})$ και $(Y_1, Y_2, \ldots, Y_{n_2})$ από κανονικούς πληθυσμούς $N(\mu_1, \sigma_1^2)$ και $N(\mu_2, \sigma_2^2)$ αντίστοιχα, με σ_1^2 και σ_2^2 γνωστές. Αν $\overline{X}$ και $\overline{Y}$ οι δειγματικοί μέσοι, τότε

$$\overline{X} - \overline{Y} \sim N(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}) \quad \text{και} \quad Z = \frac{(\overline{X} - \overline{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1)$$

και ένα $(1-a)\cdot 100\%$ διάστημα εμπιστοσύνης για την διαφορά $\mu_1 - \mu_2$ είναι το

$$[(\overline{X} - \overline{Y}) \pm B], \qquad \text{όπου } B = z_{a/2} \cdot \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}.$$

Δ.Ε. για την διαφορά μέσων δύο πληθυσμών

Πρόταση (Δ.Ε. διαφοράς δύο μέσων τιμών - άγνωστη διασπορά - μεγάλα δείγματα)

Έστω δύο ανεξάρτητα τυχαία δείγματα $(X_1, X_2, \ldots, X_{n_1})$ και $(Y_1, Y_2, \ldots, Y_{n_2})$ από κανονικούς πληθυσμούς με άγνωστες μέσες τιμές μ_1, μ_2 και άγνωστες διακυμάνσεις σ_1^2, σ_2^2 αντίστοιχα. Τότε,

$$Z = \frac{(\overline{X} - \overline{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \to N(0,1)$$

και ένα προσεγγιστικό $(1-a)\cdot 100\%$ διάστημα εμπιστοσύνης για την διαφορά $\mu_1 - \mu_2$ είναι το

$$[(\overline{X} - \overline{Y}) \pm B], \qquad \text{όπου } B = z_{a/2} \cdot \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}.$$

Δ.Ε. για την διαφορά μέσων δύο πληθυσμών

Πρόταση (Δ.Ε. διαφοράς δύο μέσων τιμών - άγνωστες διασπορές αλλά ίσες - μικρά δείγματα)

Έστω δύο ανεξάρτητα τυχαία δείγματα $(X_1, X_2, \ldots, X_{n_1})$ και $(Y_1, Y_2, \ldots, Y_{n_2})$ από κανονικούς πληθυσμούς με άγνωστες μέσες τιμές μ_1, μ_2 και άγνωστες διακυμάνσεις $\sigma_1^2 = \sigma_2^2 = \sigma^2$ αντίστοιχα. τότε,

$$T = \frac{(\overline{X} - \overline{Y}) - (\mu_1 - \mu_2)}{\sqrt{S_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim t_{n_1+n_2-2}, \text{ όπου } S_p^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}$$

και ένα $(1-a)\cdot 100\%$ διάστημα εμπιστοσύνης για την διαφορά $\mu_1 - \mu_2$ είναι το

$$[(\overline{X} - \overline{Y}) \pm B], \qquad \text{όπου } B = t_{n_1+n_2-2,a/2} \cdot \sqrt{S_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}.$$

Δ.Ε. για την διαφορά μέσων δύο πληθυσμών

Παράδειγμα

Μετρήσεις των τελευταίων 25 ετών έδειξαν ότι η μέση βροχόπτωση στην περιοχή A για έναν ορισμένο μήνα είναι 12.2 cm. Σε μια άλλη περιοχή B τα τελευταία 20 έτη για τον ίδιο μήνα η μέση βροχόπτωση ήταν 10.5 cm. Αν θεωρήσουμε ότι οι κατανομές των δύο βροχοπτώσεων είναι $N(\mu_1, \sigma_1^2)$ και $N(\mu_2, \sigma_2^2)$ να βρεθεί ένα διάστημα εμπιστοσύνης με επίπεδο σημαντικότητας 95% για την διαφορά των μέσων βροχοπτώσεων στις δύο περιοχές όταν

i) $\sigma_1 = 1.5$ cm, $\sigma_2 = 0.5$ cm.

ii) $\sigma_1^2 = \sigma_2^2 = \sigma^2$ άγνωστη, αλλά $S_1 = 1.2$ cm, $S_2 = 0.3$ cm.

Λύση

Είναι $\overline{X} = 12.2$, $n_1 = 25$, $\overline{Y} = 10.5$, $n_2 = 20$ και $a = 0.05$, οπότε $z_{a/2} = 1.96$ και $t_{n_1+n_2-2,a/2} = 2.0167$.

Δ.Ε. για την διαφορά μέσων δύο πληθυσμών

```
import numpy as np
import scipy.stats as st
xbar, ybar, n1, n2, a = 12.2, 10.5, 25, 20, 0.05
#known variance
sigma1, sigma2 = 1.5, 0.5
se = np.sqrt(sigma1**2/n1 + sigma2**2/n2)
z = st.norm.ppf(1-a/2)
B = z*se
print("i) %s%% CI for mu1 - mu2:"%(1-a),(xbar-ybar - B, xbar
    -ybar + B))
#unknown variance
s1, s2 = 1.2, 0.3
t = st.t.ppf(1-a/2, n1+n2-2)
sp2 = ((n1-1)*s1**2+(n2-1)*s2**2)/(n1+n2-2)
B2 = t*np.sqrt(sp2*(1.0/n1 + 1.0/n2))
print("ii) %s%% CI for mu1 - mu2:"%(1-a),(xbar-ybar - B2,
    xbar-ybar + B2))
#approximate
se2 = np.sqrt(s1**2/n1 + s2**2/n2)
B3 = z*se2
print("ii) approximated %s%% CI:"%(1-a),(xbar-ybar - B3,
    xbar-ybar + B3))
```

Δ.Ε. για την διαφορά μέσων δύο πληθυσμών

Output:

```
i) 0.95% CI for mu1 - mu2: (1.0725053553048052,
    2.3274946446951934)
ii) 0.95% CI for mu1 - mu2: (1.1443511472091799,
    2.255648852790818 7)
ii) approximated 0.95% CI: (1.211579491866787,
    2.1884205081332113)
```

Δ.Ε. για την διαφορά ποσοστών δύο πληθυσμών

Πρόταση (Δ.Ε. για την διαφορά ποσοστών δύο πληθυσμών)

Αν $(X_1, X_2, \ldots, X_n)$, $(Y_1, Y_2, \ldots, Y_n)$ είναι δύο ανεξάρτητα τυχαία δείγματα από πληθυσμούς $\text{Bernoulli}(p_1)$ και $\text{Bernoulli}(p_2)$, μεγέθους n_1 και n_2 αντίστοιχα, τότε

$$Z = \frac{\overline{P}_1 - \overline{P}_2 - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} \sim N(0,1)$$

όπου $\overline{P}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} X_i$ και $\overline{P}_2 = \frac{1}{n_2}\sum_{i=1}^{n_2} Y_i$ και ένα προσεγγιστικό $(1-a)\cdot 100\%$ δ.ε. για την διαφορά $p_1 - p_2$ είναι το

$$[\overline{P}_1 - \overline{P}_2 \pm z_{a/2} B], \qquad \text{όπου } B = \sqrt{\frac{\overline{P}_1(1-\overline{P}_1)}{n_1} + \frac{\overline{P}_2(1-\overline{P}_2)}{n_2}}.$$

Δ.Ε. για την διαφορά ποσοστών δύο πληθυσμών

Παράδειγμα

Σε μια ομάδα 200 ανδρών βρέθηκαν 82 καπνιστές, ενώ σε μια ομάδα 300 γυναικών βρέθηκαν 87 καπνίστριες. Να βρεθεί ένα 95% δ.ε. για την διαφορά των ποσοστών $p_1 - p_2$ όπου p_1 το ποσοστό των καπνιστών και p_2 το ποσοστό των καπνιστριών.

Λύση

```
import numpy as np
import scipy.stats as st
n1, n2, a = 200, 300, 0.05
p1, p2 = (1.0)*82/200, (1.0)*87/300
z = st.norm.ppf(1-a/2)
B = z*np.sqrt(p1*(1-p1)/n1 + p2*(1-p2)/n2)
print("Approximated %s%% CI for p1 - p2:"%(1-a),(p1-p2-B, p1
    -p2+B))
```

Output:

```
Approximated 0.95% CI for p1 - p2: (0.03466087836812895,
    0.2053391216318710 4)
```

Δ.Ε. για την διαφορά ποσοστών δύο πληθυσμών

Παράδειγμα

Σε μια σφυγμομέτρηση παρατηρήθηκε ότι στην περιοχή A σε τυχαίο δείγμα 500 ψηφοφόρων οι 420 ψηφίζουν το X κόμμα, ενώ στην περιοχή B σε τυχαίο δείγμα 300 ψηφοφόρων οι 219 ψηφίζουν το X κόμμα. Να βρεθεί ένα 99% δ.ε. για την αληθινή διαφορά στο ποσοστό των ψηφοφόρων που θα ψηφίσουν το X κόμμα στις δύο περιοχές.

```
import numpy as np
import scipy.stats as st
n1, n2, a = 500, 300, 0.01
p1, p2 = (1.0)*420/500, (1.0)*219/300
z = st.norm.ppf(1-a/2)
B = z*np.sqrt(p1*(1-p1)/n1 + p2*(1-p2)/n2)
print("Approximated %s%% CI for p1 - p2:"%(1-a),(p1-p2-B, p1
    -p2+B))
```

Output:

```
Approximated 0.99% CI for p1 - p2: (0.03162534036293719 6,
    0.1883746596370628)
```

Δ.Ε. για τον λόγο των διακυμάνσεων δύο πληθυσμών

Πρόταση (Δ.Ε. λόγου διασπορών)

Αν $(X_1, X_2, \ldots, X_{n_1})$ και $(Y_1, Y_2, \ldots, Y_{n_2})$ είναι δύο ανεξάρτητα τυχαία δείγματα από δύο κανονικούς πληθυσμούς, με δειγματικούς μέσους $\overline{X}$, $\overline{Y}$ και δειγματικές διακυμάνσεις S_x^2, S_y^2 αντίστοιχα, τότε

$$F = \frac{S_x^2/\sigma_1^2}{S_y^2/\sigma_2^2} \sim F_{n_1-1,n_2-1}$$

και ένα $(1-a)\cdot 100\%$ διάστημα εμπιστοσύνης για τον λόγο σ_1^2/σ_2^2 είναι το

$$F_{n_2-1,n_1-1,1-a/2}\frac{S_x^2}{S_y^2} < \frac{\sigma_1^2}{\sigma_2^2} < F_{n_2-1,n_1-1,a/2}\frac{S_x^2}{S_y^2}$$

Υπενθύμιση: Ισχύει η σχέση $F_{n_1,n_2,a}F_{n_2,n_1,1-a} = 1$.

Δ.Ε. για τον λόγο των διακυμάνσεων δύο πληθυσμών

Παράδειγμα

Ένα δείγμα μεγέθους $n_A = 8$ από τον πληθυσμό A έδωσε $\overline{X}_A = 43$ και $S_A^2 = 17$. Άλλο δείγμα μεγέθους $n_B = 13$ από τον πληθυσμό B έδωσε $\overline{X}_B = 31$ και $S_B^2 = 22$.
Βρείτε ένα 98% δ.ε. για τον λόγο των διασπορών των δύο πληθυσμών.

```
import numpy as np
import scipy.stats as st
n1, n2, svar1, svar2, a = 8, 13, 17, 22, 0.02
(l,u) = st.f.interval(1-a, n2-1, n1-1)
#u1 = st.f.ppf(1-a/2, n2-1, n1-1)
#l1 = st.f.ppf(a/2, n2-1, n1-1)
ci = (l*svar1/svar2, u*svar1/svar2)
print(l,u)
print("%s%% CI for var1/var2:"%(1-a), ci)
```

Output:

```
0.21554035406113947 6.469091278841487
0.98% CI for var1/var2: (0.1665539095633504,
    4.998843260922968)
```