Σπουδαιότητα και ιδιότητες του συντελεστή παραλλακτικότητας

Μενεξές Γ., Κουτσός Θ.

Τμήμα Γεωπονίας Α.Π.Θ., Εργαστήριο Γεωργίας

Περίληψη

Ο συντελεστής παραλλακτικότητας ή μεταβλητότητας (coefficient of variation - *CV*) είναι δείκτης σχετικής διακύμανσης (ή διασποράς), ο οποίος εκφράζει την ομοιογένεια των τιμών μιας ποσοτικής τυχαίας μεταβλητής. Είναι "καθαρός" αριθμός που παρέχει τη δυνατότητα σύγκρισης ακόμη και της διακύμανσης των τιμών δύο ή περισσότερων ποσοτικών τυχαίων μεταβλητών που έχουν διαφορετικές μονάδες μέτρησης. Η σπουδαιότητα του δείκτη αναδεικνύεται από την ευρεία χρήση του σε ποικίλα επιστημονικά πεδία των Βιολογικών, Οικονομικών, Κοινωνικών και άλλων Επιστημών. Στην παρούσα εργασία παρουσιάζονται ορισμένες ιδιότητες του δείκτη και αναδεικνύεται η σημασία του.

1 Συντελεστής Μεταβλητότητας - Παραλλακτικότητας (Coefficient of Variation - *CV*)

A **Εκφράζει την ομοιογένεια ενός συνόλου μετρήσεων:**

- **Υπολογισμός για δεδομένα δείγματος:**
  όπου S η τυπική απόκλιση και
  $\bar{Y}$ ο αριθμητικός μέσος όρος των μετρήσεων του δείγματος
  $$CV = \frac{S}{\bar{Y}} \text{ x } 100 \quad \text{(A)}$$
- **Υπολογισμός για δεδομένα πληθυσμού:**
  όπου σ η τυπική απόκλιση και
  μ ο αριθμητικός μέσος όρος των μετρήσεων του πληθυσμού
  $$\gamma = \frac{\sigma}{\mu} \text{ x } 100$$

B **Εκφράζει την ακρίβεια ενός Πειραματικού Σχεδίου:**
$$CV = \frac{\sqrt{MSE}}{\bar{Y}_{..}} \text{ x } 100 \quad \text{(B)}$$
όπου *MSE* το μέσο τετράγωνο των σφαλμάτων (από την κατάλληλη ANOVA) και $\bar{Y}_{..}$ ο γενικός αριθμητικός μέσος όρος των μετρήσεων του πειράματος)

Παρατηρήσεις:
α) Αν οι μετρήσεις είναι αρνητικοί αριθμοί τότε στις σχέσεις (A) και (B) χρησιμοποιείται η απόλυτη τιμή του μέσου όρου.
β) Σε πολλές εφαρμογές χρησιμοποιείται το τετράγωνο του δείκτη *CV* (CV^2)

Σε πολλά επιστημονικά πεδία η τυπική απόκλιση εκφράζει Σταθερότητα (Stability) ή Αβεβαιότητα (Uncertainty). Υπάρχουν οικογένειες κατανομών με μεγάλη τυπική απόκλιση σε σχέση με το μέσο όρο και επομένως σε αυτές τις περιπτώσεις αναμένονται υψηλές τιμές του δείκτη *CV*.

2 *CV*

Είναι σημαντικός και χρήσιμος δείκτης!

- Στις Βιολογικές Επιστήμες
- Στις Οικονομικές Επιστήμες
- Στις Κοινωνικές Επιστήμες
- Στις Ανθρωπιστικές Επιστήμες
- Στις Θετικές Επιστήμες

3 Ιδιότητες του δείκτη *CV*

- Είναι "καθαρός" αριθμός (δεν έχει μονάδες μέτρησης)
- Εκφράζει την τυπική απόκλιση ως ποσοστό % του μέσου όρου
- Είναι δείκτης σχετικής μεταβλητότητας
- Επιτρέπει τη σύγκριση της μεταβλητότητας δύο ή περισσοτέρων συνόλων μετρήσεων που έχουν διαφορετικές μονάδες μέτρησης
- Επιτρέπει τη σύγκριση της μεταβλητότητας δύο ή περισσοτέρων συνόλων μετρήσεων που έχουν την ίδια μονάδα μέτρησης αλλά αρκετά διαφορετικούς μέσους όρους ή/και τυπικές αποκλίσεις
- Η μέγιστη τιμή του δείκτη μπορεί να φθάσει την τιμή $\sqrt{1\text{-}N}$, όπου N είναι το μέγεθος δείγματος ή την τιμή $\sqrt{N}$, όπου N είναι το μέγεθος πληθυσμού (όταν όλες οι τιμές πλην μίας είναι ίσες με μηδέν)
- Χρησιμοποιείται στον καθορισμό του ελάχιστου μεγέθους δείγματος (επαναλήψεις ανά επέμβαση) σε πειράματα
- Δεν παραμένει αμετάβλητος σε μετασχηματισμούς των αρχικών μετρήσεων ($\log(Y)$, $1/Y$, $Y^{1/2}$)
- Δεν πρέπει να χρησιμοποιείται όταν οι μετρήσεις περιλαμβάνουν αρνητικούς και θετικούς αριθμούς
- Δεν πρέπει να χρησιμοποιείται όταν ο μέσος όρος είναι κοντά στο μηδέν
- Δεν μπορεί να υπολογιστεί όταν ο μέσος όρος είναι ίσος με μηδέν
- Δεν έχει νόημα (ερμηνεία) όταν οι τιμές του δείγματος δεν είναι μετρημένες σε κλίμακα αναλογίας (ratio scale)

4 Επιθυμητές τιμές του δείκτη *CV*

Εν γένει, $CV \leq 10\%$

Οι **αποδεκτές τιμές** του δείκτη *CV* καθορίζονται ανάλογα:

Α) με τη "φύση" των μετρήσεων
(π.χ. απόδοση, ύψος, βάρος, συγκέντρωση)

Β) με το είδος των πειραματικών ή δειγματοληπτικών μονάδων
(π.χ. ρύζι, σιτάρι, αγελάδα, άνθρωπος)

Γ) με τις συνθήκες πειραματισμού
(π.χ. αγρός, θερμοκήπιο, εργαστήριο)

Σε πειραματικούς σχεδιασμούς:
Τιμές *CV*: 0-10% δηλώνουν υψηλή ακρίβεια
Τιμές *CV*: 10-20% δηλώνουν μέτρια ακρίβεια
Τιμές *CV*: 20-30% δηλώνουν μικρή ακρίβεια
Τιμές *CV*: >30% δηλώνουν πολύ μικρή ακρίβεια

Σε πειράματα αγρού αποδεκτές τιμές του δείκτη *CV* έως και 33%

Η τιμή του δείκτη μπορεί να μειωθεί με κατάλληλο μετασχηματισμό των αρχικών δεδομένων (π.χ. log μετασχηματισμός). Όμως, τότε, η φυσική του ερμηνεία δεν είναι ξεκάθαρη.

Η **"βελτιστοποίηση"** της τιμής του δείκτη *CV* εξαρτάται από το σκοπό της έρευνας. Διότι:
$$CV = f(s, \bar{Y})$$
Αλλά:
$$s = g(\bar{Y}, N)$$
Συνεπώς:
$$CV = \varphi(s, \bar{Y}, N)$$

Δηλαδή: Μπορεί τρία δείγματα να έχουν ίδια τιμή του δείκτη *CV* αλλά με τελείως διαφορετική "σύνθεση" σε σχέση με την τυπική απόκλιση και το μέσο όρο (και το μέγεθος δείγματος).

Παράδειγμα:

CV=40%, με τυπική απόκλιση 40 και μέσο όρο 100 (N=50)
CV=40%, με τυπική απόκλιση 20 και μέσο όρο 50 (N=500)
CV=40%, με τυπική απόκλιση 80 και μέσο όρο 200 (N=20)

Συνεπώς: η βελτιστοποίηση της τιμής του δείκτη *CV* εξαρτάται από το εάν, για παράδειγμα, επιδίωξη της έρευνας είναι η επίτευξη υψηλών μέσων όρων (π.χ. απόδοσης) με μικρή μεταβλητότητα (π.χ. σταθερότητα, ομοιογένεια) ή η επίτευξη μεγάλης παραλλακτικότητας με υψηλούς μέσους όρους κ.λπ.

5 Όρια Εμπιστοσύνης

Αν οι μετρήσεις ακολουθούν Κανονική Κατανομή τότε ένα (1-α)% διάστημα εμπιστοσύνης για το δείκτη *CV* δίνεται από τη σχέση (Abdi, 2010):

$$\hat{C}_v \pm t_{\alpha/2,\nu} Sc_v$$

Όπου:
$$\hat{S}c_v = \frac{\hat{C}_v}{\sqrt{2N}} \qquad \hat{C}_v = \left(1 + \frac{1}{4N}\right) CV$$
$\hat{C}_v$: αμερόληπτος εκτιμητής του δείκτη *CV*

***N*=μέγεθος του δείγματος**
***ν*=*N*-1 βαθμοί ελευθερίας της *t*-Κατανομής**

$t_{\alpha/2,\nu}$: Η κρίσιμη τιμή της *t*-Κατανομής για ν βαθμούς ελευθερίας σε επίπεδο σημαντικότητας $\alpha/2$.

Παρατήρηση: Υπάρχουν και άλλες μεθοδολογικές προσεγγίσεις στην κατασκευή διαστημάτων εμπιστοσύνης για το δείκτη (π.χ. μέθοδοι που βασίζονται στην μη κεντρική *t*-Κατανομή).

6 Στατιστική σύγκριση δεικτών *CV*

Για τη σύγκριση δύο δεικτών *CV* που προέρχονται από ανεξάρτητα δείγματα έχει προταθεί ο παρακάτω στατιστικός έλεγχος:

H_0: $\gamma_1 = \gamma_2$,
H_1: $\gamma_1 \neq \gamma_2$,

Σε επίπεδο σημαντικότητας α (π.χ. α=0,05)

Το στατιστικό του ελέγχου υπολογίζεται με βάση την παρακάτω σχέση (Forkman, 2009):

$$F = \frac{c_1^2 / (1 + c_1^2 (n_1 - 1)/n_1)}{c_2^2 / (1 + c_2^2 (n_2 - 1)/n_2)}$$

Όπου:
$c_1 = CV_1$ του πρώτου δείγματος
$c_2 = CV_2$ του δευτέρου δείγματος

n_1: μέγεθος πρώτου δείγματος
n_2: Μέγεθος δευτέρου δείγματος

Η απόφαση για την απόρριψη ή όχι της μηδενικής υπόθεσης H_0 βασίζεται στην κρίσιμη τιμή της *F*-Κατανομής.

7 Παραδείγματα άλλων δεικτών σχετικής μεταβλητότητας

A) Non Parametric *CV*

$$CV_{np} = \frac{Q_{50}}{\frac{Q_{75} - Q_{25}}{2}} \text{ x } 100 = \frac{2Q_{50}}{Q_{75} - Q_{25}} \text{ x } 100 \quad \text{ή} \quad CV_{np} = \frac{Q_{50}}{Q_{75} - Q_{25}} \text{ x } 100$$

Όπου: Q_{50} : *διάμεση τιμή*
Q_{25} : *πρώτο τεταρτημόριο,*
Q_{75} : *τρίτο τεταρτημόριο*

B) Proportional Variability

$$PV = \frac{1}{c} \sum_{i \neq j} \left(1 - \frac{min(y_i, y_j)}{max(y_i, y_j)}\right)$$, όπου $i, j = 1, ..., N$ και $c = \binom{N}{2}$

και y_i οι μετρήσεις του δείγματος (Heath and Borowski, 2013)

8 Ένας "έξυπνος" μετασχηματισμός

Με βάση την παρακάτω σχέση [1] μπορούμε να προσθέσουμε στις αρχικές τιμές ενός συνόλου μετρήσεων (με τυπική απόκλιση s και συντελεστή CV=V ως ποσοστό) έναν κατάλληλο αριθμό λ, ώστε οι νέες μετασχηματισμένες μετρήσεις να έχουν συγκεκριμένη επιθυμητή τιμή συντελεστή μεταβλητότητας δ (ως ποσοστό).

$$\lambda \geq s\left(\frac{1}{\delta} - \frac{1}{V}\right) \quad [1]$$

Παράδειγμα:

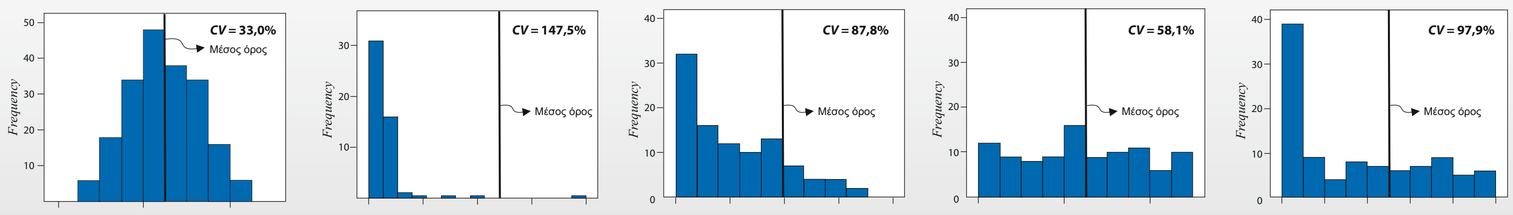
Έστω $\bar{Y}$=40 και s=30, V=0,75 και δ=0,20
Συνεπώς:

$\lambda \geq 30(5\text{-}1{,}333\ldots) \Rightarrow \lambda \geq 110$

Επομένως, αν στις αρχικές τιμές προστεθεί η ποσότητα 110 τότε ο δείκτης *CV* για τις μετασχηματισμένες τιμές θα είναι 0,20 ή 20%.

9 Συνέπειες μεγάλων τιμών του δείκτη *CV*

Γενικά: Υψηλές τιμές του δείκτη *CV* (>50%) δηλώνουν μη κανονικότητα των δεδομένων ή πρόβλημα στη διαδικασία παραγωγής των δεδομένων.
Στα παρακάτω διαγράμματα παρουσιάζονται κατανομές τιμών δειγμάτων και οι αντίστοιχοι δείκτες *CV*.



Διαπιστώνεται ότι οι κατανομές είτε είναι ασύμμετρες είτε προσεγγίζουν την ομοιόμορφη κατανομή (αλλά με μεγάλο εύρος τιμών). Δεν παρουσιάζουν κανονικότητα και η ύπαρξη ακραίων τιμών (outliers) προκαλεί ασυμμετρία και μεγάλη παραλλακτικότητα.
Συνεπώς:

- Ο αριθμητικός μέσος όρος δεν είναι κατάλληλος δείκτης κεντρικής τάσης και δεν μπορεί να αξιοποιηθεί περαιτέρω σε άλλες στατιστικές αναλύσεις.
  Ο μέσος όρος δεν "αντιπροσωπεύει" τα δεδομένα.
- Η περαιτέρω στατιστική επεξεργασία των δεδομένων θα πρέπει να γίνει με βάση Μη Παραμετρικές μεθόδους ή με χρήση Γενικευμένων Γραμμικών Υποδειγμάτων.

Παρατήρηση: Αν η τιμή του δείκτη *CV* είναι 100%, π.χ. 120%, τότε είναι προτιμότερο ο δείκτης να γράφεται ως 1,2.

10 Συμπεράσματα

- Ο δείκτης *CV* είναι χρήσιμος και σημαντικός δείκτης σχετικής μεταβλητότητας ή/και ακρίβειας σε πολλά επιστημονικά πεδία.
- Πολύ υψηλές τιμές του δείκτη *CV* δηλώνουν μη κανονικότητα των δεδομένων ή πρόβλημα στη διαδικασία παραγωγής των δεδομένων.
- Σε δεδομένα που προέρχονται από μη κανονικές κατανομές, με μεγάλη ασυμμετρία και μεγάλη παραλλακτικότητα ο αριθμητικός μέσος όρος δεν είναι "καλός" δείκτης κεντρικής τάσης.
- Η τιμή του δείκτη *CV* μπορεί να μεταβληθεί σε επιθυμητά ή αποδεκτά επίπεδα μέσω κατάλληλων μετασχηματισμών.
- Ο δείκτης *CV* δεν θα πρέπει να χρησιμοποιείται όταν η σχέση των δύο "συστατικών" του, δηλαδή του μέσου όρου και της τυπικής απόκλισης, είναι "ανταγωνιστική" (π.χ. απόδοση *vs* σταθερότητα).

Βιβλιογραφία

Abdi H. (2010). Coefficient of Variation. In Neil Salkind (Ed.), *Encyclopedia of Research Design*. Thousand Oaks, CA: Sage. 2010.
Heath, J.P. and Borowski, P. (2013). Quantifying Proportional Variability. *PLoS ONE*, 8(12): e84074. doi:10.1371/journal.pone.0084074.
Forkman, J. (2009). Estimator and Tests for Common Coefficients of Variation in Normal Distributions. *Communications in Statistics - Theory and Methods*, Volume: 38 Number: 2, pp 233-251.