ΣΤΑΤΙΣΤΙΚΗ ΙΙΙ

Εισαγωγή

Δήμητρα Κυριακοπούλου

Τμήμα Οικονομικών Επιστημών

Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών

Οργάνωση Μαθήματος

- Η ύλη του μαθήματος Στατιστική ΙΙΙ καλύπτει το φάσμα του πεδίου της **Μαθηματικής Στατιστικής** (Mathematical Statistics).

- Το μάθημα επικεντρώνεται στην Επαγωγική Στατιστική (Statistical Inference) και αποτελείται από τις εξής βασικές ενότητες:
  - Εκτιμητική
  - Έλεγχοι υποθέσεων

- Σκοπός του μαθήματος είναι:
  - η εμβάθυνση σε ζητήματα θεμελίωσης διαδικασιών στατιστικής επαγωγής
  - η απόκτηση ενός γενικού θεωρητικού και μαθηματικού υπόβαθρου για την κατανόηση και την περαιτέρω μελέτη θεμάτων Στατιστικής και Οικονομετρικής μεθοδολογίας, καθώς και ειδικότερες μεθόδους, κλασικές και μη κλασικές, που είναι χρήσιμες στην ανάλυση στοχαστικών υποδειγμάτων.

- Στο πρώτο μέρος, το οποίο είναι εισαγωγικό, αναφερόμαστε σε έννοιες που προκύπτουν στο πλαίσιο της θεωρίας πιθανοτήτων.

- Τέτοιες έννοιες συναντώνται όχι μόνο σε οτιδήποτε αφορά στη Στατιστική, αλλά και σε όποιο μέρος της οικονομικής θεωρίας αναφέρεται σε συνθήκες **αβεβαιότητας**. Συνεπώς, η χρησιμότητά τους είναι ευρύτερη στο εύρος των οικονομικών σπουδών.

- Εξετάζουμε έννοιες όπως η κατανομή πιθανότητας και οι αναπαραστάσεις της, η τυχαία μεταβλητή και το τυχαίο διάνυσμα.

- Η θεωρία πιθανοτήτων χρησιμοποιεί έννοιες από τη θεωρία συνόλων (συνολοσυναρτήσεις, πράξεις κ.ο.κ.) και είναι το μαθηματικό θεμέλιο της Στατιστικής.

- Στο δεύτερο και μεγαλύτερο μέρος, το οποίο άπτεται της Στατιστικής, εξετάζεται η έννοια του στατιστικού υποδείγματος καθώς και ζητημάτων εκτιμητικής και ελέγχων στο πλαίσιο της θεωρίας πιθανοφάνειας.

- Με την όσο το δυνατόν μεγαλύτερη μαθηματική αυστηρότητα, η μετάβαση από την κατανόηση αυτών των εννοιών στην κατανόηση περαιτέρω εννοιών εκτιμητικής και ελέγχων στο πλαίσιο μαθημάτων της οικονομετρίας, θα γίνεται πιο ομαλά.

- **Ενδεικτική βιβλιογραφία**:

  - Κοκολάκης, Γ., και Φουσκάκης Δ., «Στατιστική: Θεωρία και Εφαρμογές», Εκδόσεις Συμεών, 2009.

  - Παπαιωάννου, Τ., και Φερεντίνος Κ. «Μαθηματική Στατιστική», Εκδόσεις Σταμούλη.

  - Κοντογιάννης, Ι., και Τουμπής Σ., «Στοιχεία Πιθανοτήτων. Με Εφαρμογές στην Στατιστική και την Πληροφορική», Εκδόσεις Κάλλιπος, 2015 (μπορείτε να το κατεβάσετε ως αρχείο pdf εδώ: https://repository.kallipos.gr/handle/11419/2810?locale=el)

Βασικές Έννοιες

- Η Στατιστική χωρίζεται σε **περιγραφική και επαγωγική**.
- Το πρώτο αναφέρεται στη συλλογή από διαδικασίες που συνοψίζουν την εμπειρική πληροφορία του δείγματος:
    - θεωρούμε ότι το φαινόμενο που σχετίζεται με το δείγμα εξηγείται πιθανοθεωρητικά από μία κατανομή πιθανότητας.
- Το δεύτερο αναφέρεται στη συλλογή από διαδικασίες που χρησιμοποιούν την εμπειρική πληροφορία του δείγματος για την εύρεση της άγνωστης κατανομής (εκτιμητική, έλεγχοι υποθέσεων, διαστήματα εμπιστοσύνης, κ.ο.κ.)
- Προκειμένου να εξεταστούν τα παραπάνω χρειάζεται καλή κατανόηση της έννοιας της **κατανομής πιθανότητας**
    - η βασική ιδέα της θεωρίας πιθανοτήτων είναι η μελέτη τυχαίων μεταβλητών, όπου κάθε μία συνδέεται με μια συνάρτηση κατανομής.

Ορισμός τυχαίας μεταβλητής

- Τυχαία μεταβλητή καλείται η συνάρτηση η οποία απεικονίζει τα δυνατά αποτελέσματα πειράματος τύχης στο σύνολο των πραγματικών αριθμών. Με άλλα λόγια, είναι μια συνάρτηση από ένα δειγματικό χώρο σε έναν αριθμό.

Διακριτή:

- Η τυχαία μεταβλητή X με τιμές $x \in \mathbb{R}$ καλείται διακριτή, εάν το πεδίο τιμών αυτής περιέχει πεπερασμένο ή αριθμήσιμο πλήθος τιμών και εάν υπάρχει συνάρτηση $f(x)$ μέσω της οποίας απεικονίζονται οι τιμές x στις πιθανότητες αυτών, ήτοι:

$$f(x) = P(X = x)$$

  - η $f(x)$ καλείται συνάρτηση πιθανότητας της τυχαίας μεταβλητής X και η αναπαράστασή της γίνεται μέσω ενός ιστογράμματος πιθανότητας.
  - $f(x) \geq 0, \quad \sum_x f(x) = 1$
  - λόγω της παραπάνω αντιστοιχίας, η $f(x)$ καλείται και κατανομή.

Συνεχής:

- Η τυχαία μεταβλητή X με τιμές $x \in \mathbb{R}$ καλείται συνεχής, εάν το πεδίο τιμών αυτής είναι οι πραγματικοί αριθμοί ή διαστήματα αυτών, εάν υπάρχει συνεχής συνάρτηση $f(x)$ ώστε $f(x) \geq 0$ για κάθε $-\infty \leq x \leq \infty$ και εάν ισχύει ότι

$$P(\alpha < X < \beta) = \int_{\alpha}^{\beta} f(x)\, dx, \quad \forall \alpha \leq \beta$$

  - η $f(x)$ καλείται συνάρτηση πυκνότητας πιθανότητας (probability density function) της τυχαίας μεταβλητής X και είναι μια συνεχής συνάρτηση
  - $f(x) \geq 0, \quad \int_{-\infty}^{\infty} f(x)\, dx = 1$
  - λόγω της παραπάνω αντιστοιχίας, η $f(x)$ καλείται και κατανομή.

- Η <u>αθροιστική συνάρτηση κατανομής</u> (cumulative distribution function -CDF) δίνεται από:

$$F(x) = P(X \leq x)$$
$$= \begin{cases} \sum_{z \leq x} f(z) \\ \int_{-\infty}^{x} f(z)\, dz \end{cases}$$

- **Ιδιότητες:**

1. Είναι μη φθίνουσα συνάρτηση (non-decreasing)
2. Είναι συνεχής συνάρτηση από τα δεξιά (right-continuous)
3. τα όρια της είναι στο άπειρο:

$$\lim_{x \to -\infty} F(x) = 0, \quad \lim_{x \to \infty} F(x) = 1.$$

- Η **μέση τιμή** μιας διακριτής τυχαίας μεταβλητής X δίνεται από:

$$E\left(X\right) = \sum_{x} x f\left(x\right),$$

δεδομένου ότι $\sum_x |x| f\left(x\right) < \infty$, δηλαδή πρέπει να συγκλίνει. Διαφοερετικά, η μέση τιμή δεν ορίζεται.

- Η μέση τιμή μιας συνεχούς τυχαίας μεταβλητής X δίνεται από:

$$E\left(X\right) = \int_{-\infty}^{\infty} x f\left(x\right) dx,$$

δεδομένου ότι $\int_{-\infty}^{\infty} |x| f\left(x\right) dx < \infty$ ώστε να ορίζεται η μέση τιμή.

Ροπογεννήτριες Συναρτήσεις (Moment Generating Function)

- Οι ροπές μιας τυχαίας μεταβλητής X μπορούν να υπολογισθούν με την βοήθεια κατάλληλων συναρτήσεων, οι οποίες ονομάζονται ροπογεννήτριες συναρτήσεις και τις συμβολίζουμε ως $M_X(t)$:

$$M_X(t) = E\left(e^{tX}\right),\ \forall t \in (-\delta, \delta), \delta > 0$$

- Αν η τυχαία μεταβλητή X είναι διακριτή με συνάρτηση πιθανότητας $f_X(x)$:

$$M_X(t) = \sum_x e^{tx} f_X(x),\ |t| < \delta$$

- Αν η τυχαία μεταβλητή X είναι συνεχής με συνάρτηση πυκνότητας πιθανότητας $f_X(x)$:

$$M_X(t) = \int_{-\infty}^{\infty} e^{tx} f(x)\, x dx \ \ |t| < \delta$$

Ισχύει ότι:

- $M_X(0) = 1$ πάντα. Η ροπογεννήτρια συνάρτηση μπορεί να ορίζεται ή μπορεί όχι για άλλες τιμές του t.
- Εάν $M_X(t)$ ορίζεται για όλες τις τιμές του t σε ένα ανοιχτό διάστημα που περιλαμβάνει το 0, τότε:
  - οι ροπές όλων των τάξεων υπάρχουν
  - $E(X^r) = M_X^{(r)}(0)$, όπου r η τάξη παραγώγισης
  - $M_X(t)$ ορίζει μοναδικώς την κατανομή της τυχαίας μεταβλητής X:

$$
\begin{aligned}
M'_X(0) &= E(X) \\
M''_X(0) &= E\left(X^2\right) \\
&= \ldots
\end{aligned}
$$

Κανονική Κατανομή με Παραμέτρους (Normal or Gaussian Distribution)

- Μια τυχαία μεταβλητή X ακολουθεί την κανονική κατανομή, αν η συνάρτηση κατανομής της δίνεται από:

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

και συμβολίζεται ως $X \sim N\left(\mu, \sigma^2\right)$.

- Μέση τιμή:

$$E(X) = \mu \in \mathbb{R}$$

- Διακύμανση:

$$Var(X) = \sigma^2 > 0$$

- Ιδιαίτερο παράδειγμα αυτής της οικογένειας κατανομών είναι η τυπική κανονική κατανομή (Standard Normal), για την οποία ισχύει ότι:

$$\mu = 0,\ \sigma^2 = 1$$

και συμβολίζεται ως $N\left(0,1\right)$.

- Μια τυχαία μεταβλητή X ονομάζεται τυποποιημένη αν η συνάρτηση κατανομής της δίνεται από:

$$\phi\left(x\right) = P\left(X \leq x\right)$$

όπου

$$\phi\left(x\right) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2},\ x \in \mathbb{R}$$

Βασικό Πλαίσιο του Μαθήματος

- Το **πλαίσιο της παραμετρικής εκτίμησης** περιλαμβάνει τον ορισμό ενός τυχαίου δείγματος:

$$(x_1, ..., x_n),$$

το οποίο είναι η πραγμάτωση (realization) τυχαίων μεταβλητών $(X_1, ...X_n)$, για τις οποίες υποθέτουμε ότι:

  - έχουν την ίδια κατανομή
  - είναι ανεξάρτητες ανά δύο.

- Αυτό καλείται i.i.d. (independent and identically distributed).

- Στη συνέχεια, θεωρούμε μια οικογένεια κατανομών σε ένα Χώρο πιθανότητας - ο οποίος ορίζεται ως το σύνολο όλων των πιθανών τιμών που οι τυχαίες μεταβλητές μπορεί να πάρουν.

- Υποθέτουμε ότι οι τυχαίες μεταβλητές $X = (X_1, ...X_n)$ είναι ένα υποσύνολο των πραγματικών αριθμών $\mathbb{R}$, δηλαδή $X \in \mathbb{R}$, για παράδειγμα:

$$\{0,1\}, [0,1], [0,\infty), \mathbb{R}, \kappa\lambda\pi.$$

- Συμβολίζουμε την πιθανότητα να παρατηρήσουμε τιμές από ένα υποσύνολο των τυχαίων μεταβλητών, για παράδειγμα $A \subseteq X$, ως $P_\theta(A)$:

$$\{P_\theta(A), \theta \in \Theta\}$$

  όπου θ είναι <u>άγνωστη παράμετρος</u> και Θ το σύνολο τιμών όλων των πιθανών άγνωστων παραμέτρων.

- Οι τυχαίες μεταβλητές έχουν εξαχθεί σύμφωνα με κάποια κατανομή πιθανότητας P_{θ_0} από την παραπάνω οικογένεια κατανομών, για κάποιο άγνωστο $\theta_0 \in \Theta$.

- Στο μάθημα αυτό, εξετάζουμε έναν βασικό σκοπό της Στατιστικής που είναι η **επαγωγική διαδικασία,** δηλαδή η εξαγωγή συμπερασμάτων σχετικά με μια άγνωστη παράμετρο θ που σχετίζεται με έναν πληθυσμό:
  - στη Στατιστική, ο πληθυσμός είναι ένα σύνολο που αποτελείται από παρόμοια στοιχεία που μας ενδιαφέρουν
  - ένας πληθυσμός είναι πολύ μεγάλος για να αντιμετωπιστεί άμεσα
  - στη θέση αυτού, συχνά παίρνουμε δείγματα από τον πληθυσμό και στη συνέχεια χρησιμοποιούμε τα δείγματα για να εκτιμήσουμε μια παράμετρο θ του πληθυσμού, όπως ο μέσος όρος, η διακύμανση, η διάμεσος ή ακόμη και η πραγματική κατανομή του.

Σημαντικά Στατιστικά Μεγέθη

- Δειγματικός μέσος / Μέση τιμή δείγματος:

$$\bar{x} = \frac{x_1 + x_2 + ... + x_n}{n}$$
$$= \frac{1}{n}\sum_{i=1}^{n} x_i$$

- Διακύμανση δείγματος:

$$s^2 = \frac{(x_1 - \bar{x})^2 + (x_1 - \bar{x})^2 + ... + (x_n - \bar{x})^2}{n-1}$$
$$= \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \bar{x})^2$$

- Τυπική απόκλιση:

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

- Επιπλέον, υποθέτουμε ότι η κατανομή μιας τυχαίας μεταβλητής X έχει γνωστή συναρτησιακή μορφή, που περιγράφεται είτε από τη συνάρτηση πυκνότητας πιθανότητας, είτε από τη συνάρτηση πιθανότητας ή από την αθροιστική συνάρτηση κατανομής, αλλά περιλαμβάνει μια άγνωστη, **προς εκτίμηση**, παράμετρο, θ.

- Όταν γνωρίζουμε την τιμή της παραμέτρου, τότε η συνάρτηση κατανομής της X είναι πλήρως ορισμένη και η στοχαστική συμπεριφοράς της στον πληθυσμό είναι πλήρως γνωστή.

- Παράδειγμα 1: Η κανονική κατανομή με μέσο μ και διακύμανση σ^2

$$N\left(\mu, \sigma^2\right)$$

Η άγνωστες παράμετροι για αυτή την οικογένεια δίνονται από το διάνυσμα:

$$\theta = \left(\mu, \sigma^2\right)$$

και έχουμε επίσης ότι:

$$\Theta = \mathbb{R} \times [0, \infty) .$$

- Παράδειγμα 2: Η κατανομή Bernoulli με πιθανότητα επιτυχίας θ:
  Για αυτή την οικογένεια κατανομών έχουμε ότι:
  $$\Theta = \{\theta \in [0,1] \subset \mathbb{R}\}.$$
- Αυτό, λοιπόν, που μας απασχολεί στην Στατιστική Επαγωγή είναι να χρησιμοποιήσουμε το δείγμα $(x_1, ..., x_n)$ ώστε να εξάγουμε συμεράσματα ως προς τις άγνωστες προς εκτίμηση παραμέτρους.
- Αυτό είναι το αντίστροφο πρόβλημα με αυτό με το οποίο μερικώς ασχολείται η θεωρία πιθανοτήτων (σε προηγούμενα μαθήματα Στατιστικής):
  - η θεωρία πιθανοτήτων προσπαθεί να προβλέψει τη συμπεριφορά της τυχαίας μεταβλητής X υποθέτοντας ότι γνωρίζουμε πλήρως την κατανομή της.

Μέσα σε αυτό το πλαίσιο που μόλις δημιουργήσαμε, μελετάμε τα εξής θέματα:

1. [**Εκτιμητική**]: Δεδομένων των τυχαίων μεταβλητών, δηλαδή των παρατηρήσεων που έχουμε διαθέσιμες, $X_1, X_2, ..., X_n$, θέλουμε να εκτιμήσουμε μια άγνωστη παράμετρο θ, ήτοι να βρεθεί συνάρτηση $\hat{\theta} = \hat{\theta}\left(X_1, X_2, ..., X_n\right)$ μέσω της οποίας το $\hat{\theta}$ προσεγγίζει (και μελετάμε πόσο καλά) την παράμετρο θ.

2. [**Έλεγχος Υποθέσεων**]: Θέλουμε να ελέγξουμε συγκεκριμένες υποθέσεις σχετικά με την τιμή της άγνωστης παραμέτρου θ:

$$\theta_0 = \theta_1?$$
$$\theta_0 \lesseqgtr \theta_1$$
$$\theta_0 \neq \theta_1$$