ΣΤΑΤΙΣΤΙΚΗ ΙΙΙ

Θεωρία 07

Δήμητρα Κυριακοπούλου

Τμήμα Οικονομικών Επιστημών

Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών

- Στο κομμάτι αυτής της Θεωρίας δίνονται οι **βασικές έννοιες των ελέγχων υποθέσεων**.

- Για την κατανόηση όσων παρουσιάζονται, είναι απαραίτητη η γνώση των εννοιών:

    - της πιθανότητας,
    - της δεσμευμένης πιθανότητας,
    - των μετασχηματισμών των τυχαίων μεταβλητών,
    - του τυχαίου δείγματος,
    - της στατιστικής συνάρτησης και
    - της συνάρτησης πιθανοφάνειας.

- Ένας έλεγχος ονομάζεται **παραμετρικός** αν είναι γνωστή η συναρτησιακή μορφή της συνάρτησης κατανομής αλλά είναι άγνωστη η τιμή της παραμέτρου θ από την οποία εξαρτάται η κατανομή και οι υποθέσεις αφορούν σε αυτήν την παράμετρο (αν οι υποθέσεις αφορούν στη μορφή της κατανομής, τότε ο έλεγχος λέγεται μη παραμετρικός).

Έλεγχος Υπόθεσης (Hypothesis testing)

- Στους παραμετρικούς ελέγχους, η κλάση κατανομών στην οποία ανήκει η κατανομή που μελετάμε περιορίζεται στον παραμετρικό χώρο Θ, για τον οποίο προκύπτει μια διαμέριση σε δύο σύνολα Θ_0και Θ_1, τέτοια ώστε:

$$\Theta_0 \cup \Theta_1 = \Theta.$$

- Ένας γενικός έλεγχος υπόθεσης που αφορά σε μία παράμετρο θ μιας κατανομής έχει την εξής μορφή:

$$H_0 : \theta \in \Theta_0$$
$$H_1 : \theta \in \Theta_1$$

όπου H_0 καλείται ως μηδενική (ή αλλιώς αρχική) υπόθεση και H_1 ως εναλλακτική υπόθεση.

- Μια υπόθεση καλείται απλή (simple hypothesis) αν κάθε ένα από τα σύνολα $\Theta_i, i = 0, 1$ αποτελείται από ένα μόνο στοιχείο και σύνθετη (composite hypothesis) σε άλλη περίπτωση.

- Μια <u>απλή υπόθεση</u> είναι της μορφής:

$$H_0 : \theta = \theta_0$$
$$H_1 : \theta = \theta_1$$

- Αν η υπόθεση περιείχε το σύμβολο $<$ ($\leq$) ή $>$ ($\geq$) ή $\neq$ τότε θα λέγαμε ότι πρόκειται για μια <u>σύνθετη υπόθεση</u>.

- Οι σύνθετες υποθέσεις που περιέχουν $<$ ($\leq$) ή $>$ ($\geq$) λέγονται **μονόπλευρες** (one-sided), ενώ οι υποθέσεις που περιλαμβάνουν $\neq$ λέγονται **δίπλευρες ή αλλιώς αμφίπλευρες** (two-sided).

Σφάλματα Τύπου I και II

- Το <u>σφάλμα τύπου I</u> είναι όταν:
  - απορρίπτουμε την H_0 δεδομένου ότι η H_0 είναι αληθής.
    - Η πιθανότητα $\alpha = P(\text{σφάλμα τύπου I})$, που είναι το μέγεθος του ελέγχου, καλείται **επίπεδο σημαντικότητας**.
- Το <u>σφάλμα τύπου II</u> είναι όταν:
  - δεχόμαστε την H_0 δεδομένου ότι η H_1 είναι αληθής.
    - Συμβολίζουμε την πιθανότητα να έχουμε σφάλμα τύπου II με β, δηλαδή $\beta = P(\text{σφάλμα τύπου II})$.
- Συνήθως, η μηδενική υπόθεση είναι εκείνη για την οποία η λανθασμένη απόρριψή της προκαλεί μεγαλύτερους κινδύνους σε σχέση με τους κινδύνους που προκαλεί η λανθασμένη αποδοχή της.

- Η **κρίσιμη περιοχή** (critical area / region), η οποία καλείται και απορριπτική περιοχή της μηδενικής υπόθεσης, είναι:

$$X = \{x : x \in \mathbb{R}^n\}$$

δηλαδή

$$X \in C,$$

τέτοια ώστε να απορρίπτουμε την H_0, όπου C είναι ένα υποσύνολο της κλάσης κατανομών που ανήκει η κατανομή που μελετάμε και περιέχει την άγνωστη παράμετρο.

- Ο κανόνας απόρριψης ή μη γίνεται μέσω μιας **ελεγχοσυνάρτησης** (test statistic), η οποία αντιστοιχεί στα σημεία $x \in \mathbb{R}^n$ είτε τις τιμές 0 ή 1, είτε κάποιο μέτρο πιθανότητας.

- Τα α και β είναι ανταγωνιστικά μεταξύ τους, δηλαδή όταν μικραίνει το α μεγαλώνει το β.

- Επίσης ισχύει ότι:

$$\alpha = P\left(X \in C \mid H_0\right) = \int\limits_{x \in C} L_0\left(x\right) dx$$

και

$$\beta = P\left(X \in \bar{C} \mid H_1\right) = \int\limits_{x \in \bar{C}} L_1\left(x\right) dx,$$

όπου $L_i\left(x\right)$ είναι η πιθανοφάνεια της παρατήρησης x αν $\theta = \theta_i$ και $\bar{C}$ ένα άλλο υποσύνολο της κλάσης κατανομών.

- Να σημειωθεί ότι βάσει όλων των παραπάνω, αναφερόμαστε σε **μη τυχαιοποιημένο έλεγχο** (nonrandomized test), κατα τον οποίο η ελεγχοσυνάρτηση αντιστοιχεί στο σημείο $x \in \mathbb{R}^n$ μόνο τις τιμές 0 ή 1.

- **Το πρόβλημα το οποίο αντιμετωπίζουμε είναι** να αποφασίσουμε ποια από τις υποθέσεις είναι σωστή, με άλλα λόγια σε ποιο από τα σύνολα $C, \bar{C}$ ανήκει η κατανομή που μελετάμε, που έχει παράμετρο θ.

- Η **ισχύς του ελέγχου** ορίζεται ως η πιθανότητα π που δίνεται από:

$$1-\beta = P\left(X \in C \mid H_1\right) = \int\limits_{x \in C} L_1\left(x\right) dx.$$

- <u>Παράδειγμα:</u>

  Έστω συνάρτηση κατανομής:

$$f\left(x\right) = \begin{cases} \left(\theta+1\right) x^{\theta}, 0 \leq x \leq 1, \quad \theta > 0 \\ 0, \alpha\lambda\lambda o\acute{\upsilon} \end{cases}$$

  και έστω ότι εξετάζουμε την εξής υπόθεση:

$$H_0 : \theta = 0$$

$$H_1 : \theta = 1$$

  Τότε ισχύει ότι:

$$f_0\left(x\right) = 1,$$

$$f_1\left(x\right) = 2x.$$

Για $x \in [0,1]$, η κρίσιμη περιοχή (στην οποία απορρίπτουμε την H_0) είναι $x > \frac{1}{2}$.

Επίσης, έχουμε ότι:

$$\begin{aligned} P_0\left(x > \frac{1}{2}\right) &= 1 - P_0\left(x \leq \frac{1}{2}\right) \\ &= 1 - \int_0^{1/2} dx \\ &= 1 - [x]_0^{1/2} \\ &= 1 - \frac{1}{2} \\ &= \frac{1}{2} \end{aligned}$$

και

$$
\begin{aligned}
P_1\left(x > \frac{1}{2}\right) &= 1 - P_1\left(x \leq \frac{1}{2}\right) \\
&= 1 - \int_0^{1/2} 2x dx \\
&= 1 - \left[x^2\right]_0^{1/2} \\
&= 1 - \frac{1}{4} \\
&= \frac{3}{4}.
\end{aligned}
$$

Οπότε:

$$\alpha = P\left(x > \frac{1}{2} \mid \theta = 0\right) = \frac{1}{2},$$
$$\beta = P\left(x \leq \frac{1}{2} \mid \theta = 1\right) = \frac{1}{4},$$
$$1 - \beta = P\left(x > \frac{1}{2} \mid \theta = 1\right) = \frac{3}{4}.$$

Αν η κρίσιμη περιοχή ήταν $x > \frac{3}{4}$τότε:

$$\alpha = P\left(x > \frac{3}{4} \mid \theta = 0\right) = \frac{1}{4},$$
$$\beta = P\left(x \leq \frac{3}{4} \mid \theta = 1\right) = 0.5625,$$
$$1 - \beta = P\left(x > \frac{3}{4} \mid \theta = 1\right) = 0.4375.$$

Και γενικά μπορούμε να θέσουμε το ερώτημα:

$$\kappa = ?, \quad \alpha\nu \;\; \pi\chi \;\; \alpha = 0.15$$

δηλαδή αν η κρίσιμη περιοχή ορίζεται ως $x > \kappa$, να βρεθεί η τιμή του κ αν ισχύει ότι η πιθανότητα να κάνουμε σφάλμα τύπου I, ή με άλλα λόγια το επίπεδο σημαντικότητας, ισοδυναμεί με 0.15.

Τότε:

$$\alpha = P\left(x > \kappa \mid \theta = 0\right) = 0.15 \Rightarrow \kappa = 0.85$$
$$\beta = P\left(x \leq 0.85 \mid \theta = 1\right) = 0.81,$$
$$1 - \beta = P\left(x > 0.85 \mid \theta = 1\right) = 0.19.$$

Γενικά, στο Παράδειγμά μας αυτό, η κρίσιμη περιοχή είναι:

$$C = \{x \in [0,1] \mid x > \kappa\}.$$

Αν $\alpha = 0.10$,

$$P(x > \kappa \mid \theta = 0) = 0.10 \Rightarrow \kappa = \frac{9}{10}.$$

Αν $\alpha = \alpha_0$,

$$\begin{aligned} \alpha_0 &= P(x > \kappa_0 \mid \theta = 0) \\ &= 1 - \kappa_0 \Rightarrow \\ \kappa_0 &= 1 - \alpha_0. \end{aligned}$$

Έλεγχος Μεγίστης Ισχύος και το Θεμελιώδες Λήμμα των **Neyman-Pearson**

- Ο έλεγχος που μεγιστοποιεί το $1-\beta$ για δεδομένο α καλείται έλεγχος μεγίστης ισχύος μεγέθους α
  - επειδή $\min \beta \iff \max(1-\beta)$:=ισχύς
- Στη συνέχεια διατυπώνεται το θεμελιώδες λήμμα των Neyman-Pearson, που αποτελεί την βάση της θεωρίας των ελέγχων υποθέσεων.
- Έστω $(X_1, X_2, ..., X_n)$ τυχαίο δείγμα από πληθυσμό που ακολουθεί κατανομή $f(x;\theta)$ και ας συμβολίσουμε με $\prod_{i=1}^{n} f(x_i;\theta)$ την από κοινού συνάρτηση πυκνότητας ή πιθανότητας του δείγματος.
- Έστω ότι θέλουμε να εξετάσουμε την απλή υπόθεση:

$$H_0 : \theta = \theta_0$$

$$H_1 : \theta = \theta_1$$

- Αν υπάρχει σταθερά κ και αντίστοιχη κρίσιμη περιοχή C μεγέθους α για την οποία ισχύει ότι ο λόγος των πιθανοφανειών του δείγματος κάτω από την εναλλακτική και μηδενική υπόθεση:

$$\frac{L_1(\theta_1;x)}{L_0(\theta_0;x)} > \kappa, \quad \textit{εντός της } C$$
$$\frac{L_1(\theta_1;x)}{L_0(\theta_0;x)} \leq \kappa, \quad \textit{εκτός της } C$$

  τότε η κρίσιμη περιοχή C είναι κρίσιμη περιοχή ελέγχου μεγίστης ισχύος μεγέθους α για το εν λόγω πρόβλημα ελέγχου και η σταθερά κ λέγεται **κρίσιμο σημείο ή σημείο αποκοπής** (critical value).

- Ο έλεγχος που καθορίζεται από τις παραπάνω σχέσεις καλείται **ισχυρότατος έλεγχος (IE)** και ονομάζεται συνήθως και έλεγχος λόγου πιθανοφανειών (likelihood ratio test).

- Σημείωση: η σταθερά κ πρέπει να προσδιοριστεί έτσι ώστε ο έλεγχος να έχει πραγματικά μέγεθος α.

• Έστω δύο έλεγχοι με μέγεθος α:

$$P\left(x \in C \mid \theta_0\right) = P\left(x \in D \mid \theta_0\right) = \alpha.$$

• Τότε:

$$\begin{aligned}
\pi_C - \pi_D &= P\left(x \in C \mid \theta_1\right) - P\left(x \in D \mid \theta_1\right) \\
&= P\left[x \in C \cap \bar{D} \mid \theta_1\right] + P\left[x \in C \cap D \mid \theta_1\right] \\
&\quad - \left\{P\left[x \in D \cap \bar{C} \mid \theta_1\right] + P\left[x \in C \cap D \mid \theta_1\right]\right\} \\
&= P\left[x \in C \cap \bar{D} \mid \theta_1\right] - P\left[x \in D \cap \bar{C} \mid \theta_1\right] \\
&= \int_{C \cap \bar{D}} L_1(x)\, dx - \int_{D \cap \bar{C}} L_1(x)\, dx
\end{aligned}$$

- Ισχύει ότι:

$$C \cap \bar{D} \subseteq C$$

και

$$L_1(x) > \kappa L_0(x)$$

- Καθώς και ότι:

$$D \cap \bar{C} \subseteq \bar{C}$$

και

$$L_1(x) \leq \kappa L_0(x)$$

• Άρα:

$$\begin{aligned}
\pi_C - \pi_D &\geq \kappa \int_{C\cap\bar{D}} L_0(x)\,dx - \kappa \int_{D\cap\bar{C}} L_0(x)\,dx \\
&= \kappa \left[\int_{C\cap\bar{D}} L_0(x)\,dx - \int_{D\cap\bar{C}} L_0(x)\,dx \right] \\
&= \kappa \left[\begin{array}{l} \displaystyle\int_{C\cap\bar{D}} L_0(x)\,dx + \int_{C\cap D} L_0(x)\,dx \\ \displaystyle - \int_{C\cap D} L_0(x)\,dx - \int_{D\cap\bar{C}} L_0(x)\,dx \end{array} \right] \\
&= \kappa \left[\int_{C} L_0(x)\,dx - \int_{D} L_0(x)\,dx \right] \\
&= \kappa \left[P(x \in C \mid \theta_0) - P(x \in D \mid \theta_0) \right] \\
&= \kappa (\alpha - \alpha) \\
&= 0.
\end{aligned}$$

- Παράδειγμα 1:

  Έστω $(X_1, X_2, ..., X_{10}) = (1, 0, 1, 1, 0, 0, 1, 1, 0, 0)$ τυχαίο δείγμα από κατανομή $B(n, p)$:

$$K\Gamma KK\Gamma\Gamma KK\Gamma\Gamma$$

με συνάρτηση πιθανοφάνειας:

$$\begin{aligned} L(x) &= p^{\sum x_i}(1-p)^{n-\sum x_i} \\ &= p^5 (1-p)^5 \end{aligned}$$

Έστω ότι έχουμε το πρόβλημα του ελέγχου:

$$H_0 : p = \frac{1}{2}$$
$$H_1 : p = \frac{3}{4}$$

οπότε ισχύει:

$$L_0(x) = \left(\frac{1}{2}\right)^5 \left(\frac{1}{2}\right)^5 = \left(\frac{1}{2}\right)^{10}$$
$$L_1(x) = \left(\frac{1}{4}\right)^5 \left(\frac{3}{4}\right)^5$$

Επίσης προκύπτει ότι:

$$\begin{aligned}\frac{L_1(x)}{L_0(x)} &= \frac{\left(\frac{3}{4}\right)^{\sum x_i}\left(\frac{1}{4}\right)^{n-\sum x_i}}{\left(\frac{1}{2}\right)^n} \\ &= \frac{\left(\frac{3}{4}\right)^{\sum x_i}}{\left(\frac{1}{4}\right)^{\sum x_i}}\frac{\left(\frac{1}{4}\right)^n}{\left(\frac{1}{2}\right)^n} \\ &= 3^{\sum x_i}\left(\frac{1}{2}\right)^n .\end{aligned}$$

Και:

$$\frac{L_1(x)}{L_0(x)} > \kappa \Rightarrow$$

$$\sum x_i \log 3 + n \log\frac{1}{2} > \log\kappa \Rightarrow$$

$$\sum x_i > \frac{\log\kappa - n\log\frac{1}{2}}{\log 3} = \kappa^*$$

Σε συνέχεια των προηγούμενων, έστω ότι $Y = \sum X_i \sim B(n,p)$, $n = 10$ και έστω ότι $\kappa = 5$.

Τότε:

$$\begin{aligned}
\alpha &= P(Y > \kappa \mid H_0) = P\left[B\left(10, \frac{1}{2}\right) > 5\right] \\
&= 1 - 0.6230 \\
&= 0.377
\end{aligned}$$

και

$$\begin{aligned}
\beta &= P(Y \leq \kappa \mid H_1) = P\left[B\left(10, \frac{3}{4}\right) \leq 5\right] \\
&= P\left[B\left(10, \frac{1}{4}\right) \geq 5\right] \\
&= 1 - P\left[B\left(10, \frac{1}{4}\right) \leq 4\right] \\
&= 1 - 0.9219 \\
&= 0.0781.
\end{aligned}$$

Επίσης, μπορούμε να θέσουμε το ερώτημα πόσο πρέπει να είναι η τιμή του κ αν δίνεται ότι $\alpha = 7.81\%$;

Η απάντηση είναι $\kappa = 4$.

Ή, πόσο πρέπει να είναι το κ αν $\alpha = 1.97\%$;

Η απάντηση είναι $\kappa = 5$.

- <u>Παράδειγμα 2:</u>

  Έστω ότι $(X_1, X_2, ..., X_n) \sim N\left(\mu, \sigma^2\right)$, όπου σ^2 είναι γνωστή και ας θεωρήσουμε τον έλεγχο:

$$H_0 : \mu = \mu_0$$
$$H_1 : \mu = \mu_1$$

όπου $\mu_1 > \mu_0$.

Έχουμε ότι:

$$\begin{aligned}
\frac{L_1(\mu_0; x)}{L_0(\mu_1; x)} &= \frac{\prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2\sigma^2}(x_i - \mu_1)^2\right\}}{\prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2\sigma^2}(x_i - \mu_0)^2\right\}} \\
&= \frac{\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^n \left(x_i^2 - 2n\bar{x}\mu_1 + n\mu_1^2\right)\right\}}{\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^n \left(x_i^2 - 2n\bar{x}\mu_0 + n\mu_0^2\right)\right\}} \\
&= \exp\left\{\frac{1}{2\sigma^2}\left(2n\bar{x}(\mu_1 - \mu_0) + n(\mu_1 - \mu_0)^2\right)\right\}.
\end{aligned}$$

Συνήθως κατά την αναζήτηση ισχυρότατου ελέγχου με το λήμμα Neyman-Pearson, διευκολύνει στις πράξεις να θεωρούμε αντί του λόγου πιθανοφανειών τον λογάριθμό του.

Έτσι, για μια σταθερά κ:

$$
\begin{aligned}
\log \frac{L_1(\mu_0; x)}{L_0(\mu_1; x)} = \frac{1}{2\sigma^2}\left(2n\bar{x}(\mu_1 - \mu_0) + n(\mu_1 - \mu_0)^2\right) > \kappa &\iff \\
2(\mu_1 - \mu_0)\bar{x} > \kappa^* &\iff \\
\bar{x} > c.&
\end{aligned}
$$

Οπότε προκύπτει ότι:

$$C = \{(X_1, X_2, ..., X_n) : \bar{x} > c\}.$$

Η επιλογή της σταθεράς c γίνεται χρησιμοποιώντας ότι:

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \sim N(0,1),$$

κάτω από τη μηδενική υπόθεση H_0.

Οπότε, θα απορρίψουμε την H_0 αν $z > z_\alpha$.

- <u>Παράδειγμα 3:</u>

  Έστω $(X_1, X_2, ..., X_n)$ τυχαίο δείγμα από πληθυσμό που ακολουθεί την Εκθετική κατανομή:

$$f(x;\theta) = \theta e^{-\theta x}, \quad x > 0, \quad \theta > 0$$

  και έστω ότι θέλουμε να βρούμε τον έλεγχο με <u>κρίσιμη περιοχή ελέγχου μεγίστης ισχύος</u> μεγέθους α για να εξετάσουμε την απλή υπόθεση:

$$H_0 : \theta = \theta_0$$
$$H_1 : \theta = \theta_1$$

  με $\theta_1 > \theta_0$.

  Θα εφαρμόσουμε το λήμμα Neyman-Pearson, το οποίο μας λέει ότι ο λόγος των πιθανοφανειών κάτω από την μηδενική και εναλλακτική υπόθεση πρέπει να είναι μικρότερος από μια σταθερά κ:

$$\begin{aligned}\frac{L_0\left(x\right)}{L_1\left(x\right)} &= \frac{\theta_0^n e^{-\theta_0 \sum x_i}}{\theta_1^n e^{-\theta_1 \sum x_i}} \\ &= \left(\frac{\theta_0}{\theta_1}\right)^n e^{-(\theta_0-\theta_1)\sum x_i} \leq \kappa\end{aligned}$$

$$\Longleftrightarrow e^{-(\theta_0-\theta_1)\sum x_i} \leq \kappa \left(\frac{\theta_1}{\theta_0}\right)^n$$

$$\Longleftrightarrow -(\theta_0 - \theta_1)\sum_{i=1}^{n} x_i \leq \log\left[\kappa \left(\frac{\theta_1}{\theta_0}\right)^n\right]$$

$$\Longleftrightarrow (\theta_1 - \theta_0)\sum_{i=1}^{n} x_i \leq \log\left[\kappa \left(\frac{\theta_1}{\theta_0}\right)^n\right]$$

$$\Longleftrightarrow \sum_{i=1}^{n} x_i \leq \frac{1}{\theta_1-\theta_0}\log\left[\kappa \left(\frac{\theta_1}{\theta_0}\right)^n\right] = \kappa^*,$$

όπου κ^* είναι μια νέα σταθερά.

Οπότε, η κρίσιμη περιοχή ελέγχου μεγίστης ισχύος (στην οποία απορρίπτουμε την H_0) είναι $\sum_{i=1}^{n} x_i \leq \kappa^*$.

Στη συνέχεια μπορούμε να υπολογίσουμε την πιθανότητα Σφάλματος Τύπου I, α:

$$\begin{aligned}
\alpha &= P\left(x < \kappa \mid \theta = \theta_0\right) \\
&= \int_0^{\kappa} \theta_0 e^{-\theta_0 x} dx \\
&= -\int_0^{\kappa} \left(e^{-\theta_0 x}\right)' dx \\
&= -\left[e^{-\theta_0 x}\right]_0^{\kappa} \\
&= -\left[e^{-\theta_0 \kappa} - 1\right] \\
&= 1 - e^{-\theta_0 \kappa}.
\end{aligned}$$

Από την τελευταία σχέση, έχουμε ότι:

$$\begin{aligned} e^{-\theta_0\kappa} &= 1-\alpha \Rightarrow \\ -\theta_0\kappa &= \log(1-\alpha) \Rightarrow \\ \kappa &= -\frac{1}{\theta_0}\log(1-\alpha) \\ &= \log(1-\alpha)^{-1/\theta_0}. \end{aligned}$$

Αν $\theta_0 = 1$:

$$\kappa = -\log(1-\alpha) = \log(1-\alpha)^{-1},$$

όπου αν $\alpha = 0.05$:

$$\kappa = -\log 0.95 = 0.051,$$

αν $\alpha = 0.10$:

$$\kappa = -\log 0.90 = 0.105,$$

αν $\alpha = 0.01$:

$$\kappa = -\log 0.99 = 0.01.$$