Εβδομάδα 0 - Ασκήσεις

Οι ασκήσεις της Εβδομάδας 0 εμβαθύνουν στην κατασκευή Διαστημάτων Εμπιστοσύνης για παράμετρο που εκφράζει μία απλή πιθανότητα σε δοκιμές Bernoulli. Η σύνδεση με το μάθημα βρίσκεται στο ότι η εμπειρική συνάρτηση κατανομής σε κάθε σημείο της μπορεί να εκφραστεί ως ο δειγματικός μέσος μίας ακολουθίας ανεξάρτητων δοκιμών Bernoulli. Έτσι η κατασκευή ενός Δ.Ε. για το $F(x)$ στο σημείο $x \in \mathbb{R}$ ανάγεται στην κατασκευή Δ.Ε. για το p ενός τυχαίου δείγματος από Bernoulli. Επιπλέον, αναφέρονται κάποιες ιδιότητες που συνδέονται με τα διατεταγμένα δείγματα.

0.1 (διασποράς-σταθεροποιητικοί μετασχηματισμοί): Η μέθοδος Δέλτα είναι πολύ χρήσιμη για να βρίσκουμε ασυμπτωτικές κατανομές εκτιμητριών παραμέτρων που μπορούν να εκφραστούν ως κατάλληλες συναρτήσεις εκτιμητριών με γνωστή ασυμπτωτική κατανομή.

1. Αναφέρετε τη μέθοδο και τις συνθήκες εφαρμογής της για διάσταση 1.
2. Έστω Θ ένα ανοικτό υποσύνολο του $\mathbb{R}$ που αποτελεί τον παραμετρικό χώρο ενός παραμετρικού μοντέλου με μονοδιάστατη παράμετρο θ και $\widehat{\theta}_n$ μία εκτιμήτρια του θ που είναι ασυμπτωτικά κανονική:

   $$\sqrt{n}\left(\widehat{\theta}_n - \theta\right) \xrightarrow[n\to\infty]{d} \mathcal{N}(0, \sigma^2(\theta)), \quad \forall \theta \in \Theta.$$

   Ένας μετασχηματισμός φ καλείται *διασποράς-σταθεροποιητικός* (variance stabilising), αν η $\widehat{\phi}_n := \varphi(\widehat{\theta}_n)$ έχει ασυμπτωτική κατανομή με διασπορά ανεξάρτητη του θ, δηλ. υπάρχει $c > 0$:

   $$\sqrt{n}\left(\varphi(\widehat{\theta}_n) - \varphi(\theta)\right) \xrightarrow[n\to\infty]{d} \mathcal{N}(0, c), \quad \forall \theta \in \Theta.$$

   Βρείτε κατάλληλο διασποράς-σταθεροποιητικό μετασχηματισμό για την παράμετρο p ενός μοντέλου $\mathcal{B}e(p)$ και κατασκευάστε ένα ασυμπτωτικό $(1-\alpha)-$Δ.Ε. βασισμένοι σε αυτόν τον μετασχηματισμό.
3. Συμπληρώστε το Παράδειγμα 3.1. με τον κατάλληλο κώδικα ώστε να εμφανιστούν τα αντίστοιχα $95\%-$Δ.Ε. με τη μέθοδο αυτή για $n = 10$ και $n = 100$.

0.2 (Wald-Δ.Ε. με προσθήκη ψευδοπαρατηρήσεων): Η Μπεϋζιανή μεθοδολογία μπορεί να χρησιμοποιηθεί για να "εμπνεύσει" τροποποιήσεις κατάλληλων εκτιμητριών έτσι ώστε οι κλασικές εκτιμήτριες να συμπεριφέρονται καλύτερα για μικρότερα μεγέθη δείγματος.

1. Υποθέτουμε ότι στο μοντέλο $\mathcal{B}e(p)$ διαθέτουμε 4 ψευδοδεδομένα στα οποία έχουμε 2 επιτυχίες και 2 αποτυχίες πριν δούμε τις πραγματικές παρατηρήσεις. Τροποποιήστε κατάλληλα την εκτιμήτρια του p ώστε να λάβει υπόψη τις ψευδοπαρατηρήσεις και κατασκευάστε με τη μέθοδο Wald τροποποιημένα ασυμπτωτικά Δ.Ε.
2. (προαιρετικό) Ποιά prior κατανομή για το p θα μπορούσαμε να είχαμε χρησιμοποιήσει έτσι ώστε η μπεϋζιανή εκτιμήτρια του p (η σημειακή κάτω από την L^2 συνάρτηση απώλειας) να συμπίπτει με αυτήν την εκτιμήτρια ?
3. Συμπληρώστε το Παράδειγμα 3.1. με τον κατάλληλο κώδικα ώστε να εμφανιστούν τα αντίστοιχα $95\%-$Δ.Ε. με τη μέθοδο αυτή για $n = 10$ και $n = 100$. Συμβουλευτείτε το άρθρο των Agresti and Coull και το πακέτο `binom` της R που αναφέρεται στις Σημειώσεις. Ποιές είναι οι παρατηρήσεις σας ?
4. (προαιρετικό) Κατασκευάστε 95% διάστημα αξιοπιστίας για το p χρησιμοποιώντας τη μπεϋζιανή εκτιμήτρια του Ερωτ. 2 και αντιπαραβάλλεται τα διαστήματα αυτά σε σχέση με αυτά που προκύπτουν από τα 95% Δ.Ε..

0.3 (εκτίμηση πιθανοτήτων κάλυψης): Οι 3 μέθοδοι κατασκευής Δ.Ε. για το p, CP, Wald and Wilson συμπληρώθηκαν με άλλες δύο, την VS (variance-stabilising) και την τροποποιημένη του Wald. Σε αυτήν την άσκηση θα συγκρίνουμε όλες τις μεθόδους ως προς την ικανότητά τους να αναπαράγουν την ονομαστική αξία 95% στα Δ.Ε. που κατασκευάζουν. Η πραγματική πιθανότητα κάλυψης μπορεί να εκτιμηθεί με Monte-Carlo για διάφορες τιμές του p.

1. Να παραχθούν $M = 10000$ φορές μεγέθη δείγματος $n \in \{10, 20, 100\}$ για $p \in \{0.1, 0.3, 0.5\}$ και να εκτιμηθούν οι πιθανότητες κάλυψης των 95%−Δ.Ε. που κατασκευάζονται με τις 5 διαφορετικές μεθόδους κατασκευής. Καταγράψτε τις παρατηρήσεις σας. Χρησιμοποιήστε τις ίδιες παρατηρήσεις (για κοινές τιμές του p) όταν συγκρίνετε τις πιθανότητες κάλυψης.
2. Όταν παράγεται ένα δείγμα $\mathcal{B}e(p)$ για διαφορετικά p συνήθως η προσομοίωση των δεδομένων γίνεται από την αρχή. Πώς θα μπορούσατε να κάνετε την προσομοίωση δεδομένων για όλα τα p ταυτόχρονα ? Αφού το καταφέρετε, φτιάξτε γραφήματα της πιθανότητας κάλυψης για $n \in \{10, 20, 100\}$. Η μέθοδος αυτή είναι προφανώς πολύ πιο αποδοτική. Εξηγείστε τους λόγους και κάντε τις παρατηρήσεις σας σχετικά με τις πιθανότητες κάλυψης των 5 διαφορετικών μεθόδων.
3. (προαιρετικό) Αντιπαραβάλλετε τα συμπεράσματά σας για την απόδοση των μεθόδων με τα συμπεράσματα των Agresti and Coull.

0.4 (Clopper-Pearson Δ.Ε. με ποσοστημόρια Fisher): Έστω X, Y δύο ανεξάρτητες τ.μ. με $X \sim \mathcal{G}(\alpha, \theta)$ και $Y \sim \mathcal{G}(\beta, \theta)$, όπου $\alpha, \beta, \theta > 0$ και θ παράμετρος κλίμακας.

1. Να δείξετε ότι αν $Z = X/(X + Y)$, τότε $Z \sim \mathcal{B}(\alpha, \beta)$.
2. Να δείξετε ότι αν $(\alpha, \beta) = (n_1, n_2)$ και $n_1, n_2 \in \mathbb{N}^*$, τότε η τ.μ.

$$W = \frac{n_2}{n_1}\,\frac{X}{Y} \sim \mathcal{F}_{2n_1,2n_2},$$

όπου $\mathcal{F}_{n_1,n_2}$ είναι η κατανομή Fisher με n_1 και n_2 βαθμούς αντίστοιχα.

3. Να δείξετε ότι τα $100(1-\alpha)\%$ Δ.Ε. για το p με τη μέθοδο Clopper-Pearson για $i = 1, 2, \ldots, n-1$ μπορούν να αναπαρασταθούν στη μορφή

$$I^{CP}_{1-\alpha}(s) = \left[\left(1 + \frac{n-i+1}{i\, f_{2i,2(n-i+1);1-\alpha/2}}\right)^{-1}, \left(1 + \frac{n-i}{(i+1)\, f_{2(i+1),2(n-i);\alpha/2}}\right)^{-1}\right],$$

όπου i είναι το παρατηρούμενο πλήθος των μονάδων στο δείγμα μεγέθους n (το πλήθος των επιτυχιών) και $f_{n_1,n_2;\alpha}$ το α-άνω ποσοστημόριο της κατανομής Fisher $\mathcal{F}_{n_1,n_2}$.

0.5 (διατεταγμένα δείγματα):
Έστω X τυχαία μεταβλητή με απόλυτα συνεχή κατανομή και F η συνάρτηση κατανομής της. Θέτουμε $U = F(X)$ και

$$I = \left\{ x \in \mathbb{R} : \text{υπάρχει η } F'(x) \text{ και } F'(x) > 0 \right\}.$$

Μπορούμε να υποθέσουμε χ.β.γ. ότι η συνάρτηση πυκνότητας πιθανότητας $f(x)$ είναι γνήσια θετική ακριβώς στο I.

1. Να αιτιολογήσετε ότι για την εύρεση της σ.κ. F_U της U μπορούμε να περιοριστούμε στο σύνολο $F(I)$.
2. Να δείξετε ότι $U \sim \mathcal{U}(0,1)$.
3. Αν $X_{(1)}, \ldots, X_{(n)}$ είναι ένα διατεταγμένο δείγμα από την F, τότε να δείξετε ότι το τυχαίο διάνυσμα $\Big(F(X_{(1)}), \ldots, F(X_{(n)})\Big)$ είναι ισόνομο με ένα τυχαίο διάνυσμα $\Big(U_{(1)}, \ldots, U_{(n)})\Big)$ που αντιστοιχεί σε ένα διατεταγμένο δείγμα από ομοιόμορφη κατανομή $\mathcal{U}(0,1)$.

Εβδομάδα 1 - Ασκήσεις

Οι ασκήσεις της Εβδομάδας 1 αφορούν ιδιότητες και εφαρμογές της πολυωνυμικής και της πολυδιάστατης κανονικής κατανομής.

1.1: Έστω $X = (X_1, \dots, X_m) \sim \mathcal{M}_m(n, p)$, $n, m \geq 1$, $p = (p_1, \dots, p_m) \in \Delta_0^m$. Θα λέμε ότι μία δ.τ.μ. $S = (S_1, \dots, S_m)$ ακολουθεί την *αθροιστική πολυωνυμική* και θα γράφουμε $S \sim s\mathcal{M}_m(n, p)$, αν $S_i = X_1 + X_2 + \dots + X_i$, $1 \leq i \leq m$.

1. Θέτουμε $\Sigma_X = \mathbb{V}(X)$. Να διατυπωθεί ικανή και αναγκαία συνθήκη για να είναι ο πίνακας Σ_X αντιστρέψιμος και να βρεθεί ο αντίστροφος στην περίπτωση που υπάρχει.
2. Να βρεθεί το στήριγμα και η συνάρτηση πιθανότητας του S.
3. Προσδιορίστε τα $\mathbb{E}(S)$ και $\mathbb{V}(S)$. Θέτουμε $\Sigma_S = \mathbb{V}(S)$. Να διατυπωθεί ικανή και αναγκαία συνθήκη για να είναι ο πίνακας Σ_S αντιστρέψιμος και να βρεθεί ο αντίστροφος στην περίπτωση που υπάρχει.
4. (προαιρετικό) Να υπολογιστούν τα μέτρα (οι πίνακες) πληροφορίας του Fisher I_X και I_S που αντιστοιχούν σε μία παρατήρηση των X και S αντίστοιχα.

1.2: Έστω (X_n) ακολουθία α.ι.δ.τ.μ. με $X_n \sim \boldsymbol{\mathcal{C}at}_m(p)$, $m \geq 2$ και $p \in \Delta^{m-1}$. Στην άσκηση αυτή υποθέτουμε ότι το p είναι άγνωστο και θέλουμε να το εκτιμήσουμε.

1. Να βρεθεί η εκτιμήτρια ροπών $\overline{p}_n$ του p, επεκτείνοντας κατάλληλα τη μέθοδο ροπών για δ.τ.μ.. Τί κατανομή ακολουθεί ?
2. Να δειχθεί ότι η $\overline{p}_n$ είναι ισχυρά συνεπής εκτιμήτρια του p.
3. Να βρεθεί η ασυμπτωτική κατανομή της $\overline{p}_n$.
4. Συμπεράνετε από τα παραπάνω ότι αν $\pi_i(n)$ είναι το ποσοστό των ενδείξεων που είναι $i \in \{1, \dots, 6\}$, σε n-ανεξάρτητες ρίψεις ενός τίμιου ζαριού, τότε
$$\pi(n) = \big(\pi_1(n), \dots, \pi_6(n)\big) \xrightarrow{a.s.} \left(\frac{1}{6}, \dots, \frac{1}{6}\right),$$
και προσεγγίστε ασυμπτωτικά την κατανομή του $\big(\pi_5(n), \pi_6(n)\big)$ και του $\pi_5(n)+\pi_6(n)$.

1.3: Στο ΛΟΤΤΟ είναι γνωστό ότι σε κάθε κλήρωση επιλέγονται τυχαία 6 αριθμοί από τους $\{1, 2, \dots, 49\}$. Έστω $N_n = \big(N_n^1, N_n^2, \dots, N_n^{49}\big)$, $n \geq 1$, το διάνυσμα που περιγράφει το πλήθος εμφανίσεων των αριθμών μετά από n-κληρώσεις.

1. Να υπολογίσετε την $\mathbb{E}(N_n)$ και το $\mathbb{V}(N_n)$.
2. Να προσδιοριστεί το σταθερό διάνυσμα c για το οποίο $N_n/n \xrightarrow{a.s.} c$.
3. Να βρεθεί η ασυμπτωτική κατανομή του N_n/n.
4. Αν υποθέσουμε ότι σε κάθε κλήρωση επιλέγεται ένα ζεύγος αριθμών (i, j), $1 \leq i < j \leq 49$, με άγνωστη πιθανότητα $0 < p_{i,j} < 1$, τότε να επαναλάβετε τα ερωτήματα (2) και (3).

1.4: Έστω X, Y δύο δ.τ.μ. έτσι ώστε (X, Y) να ακολουθεί πολυδιάστατη κανονική.

1. Να αποδειχθεί ότι X, Y ανεξάρτητες, αν και μόνο αν, X, Y ασυσχέτιστες. Μπορεί να παραλειφθεί η υπόθεση ότι η από κοινού κατανομή τους είναι κανονική ? Να δώσετε παραδείγματα ορθογώνιων μετασχηματισμών και να δείξετε ότι οι ορθογώνιοι μετασχηματισμοί διατηρούν την ανεξαρτησία.
2. Αν $(X, Y) \sim \mathcal{N}_2(0_2, I_2)$, τότε να βρεθεί η κατανομή της δ.τ.μ. (R, Θ) που προκύπτει με "πέρασμα" στις πολικές συντεταγμένες έτσι ώστε η R να καταγράφει την απόσταση ενός σημείου από την αρχή των αξόνων και Θ τη γωνία που σχηματίζει το διάνυσμα OM του σημείου $M = (x, y)$ με τον θετικό ημιάξονα Ox κατά τη θετική φορά διαγραφής.
3. Στην προσομοίωση ένας τρόπος παραγωγής κανονικά κατανεμημένων τ.μ. είναι ο μετασχηματισμός Box-Muller. Με τη βοήθεια του παραπάνω ερωτήματος να δειχθεί ότι χρησιμοποιώντας μόνο δύο ανεξάρτητες ομοιόμορφες τ.μ. $U_1, U_2 \sim \mathcal{U}(0, 1)$ και κατάλληλο μετασχηματισμό μπορούν να παραχθούν δύο ανεξάρτητες τ.μ. $X, Y \sim \mathcal{N}(0, 1)$.

1.5: Έστω X, Y δύο δ.τ.μ. διαστάσεων m, n αντίστοιχα έτσι ώστε $(X, Y) \sim \mathcal{N}_{m+n}(\mu, \Sigma)$. Θέτουμε $\mu = (\mu_X, \mu_Y)$, $\mathbb{V}(X) = \Sigma_X$, $\mathbb{V}(Y) = \Sigma_Y$ και $\mathbb{C}(X, Y) = \Sigma_{X,Y}$. Υποθέτουμε ότι ο πίνακας Σ_Y είναι θετικά ορισμένος.

1. Να δειχθεί ότι οι δ.τ.μ. Y και $X - \Sigma_{X,Y}\Sigma_Y^{-1}Y$ είναι ανεξάρτητες.
2. Να δειχθεί ότι η δεσμευμένη κατανομή της $[X|Y = y] \sim \mathcal{N}_m(\mu_{X|Y}, \Sigma_{X|Y})$, όπου

$$\begin{aligned}
\mu_{X|Y} &= \mu_X + \Sigma_{XY}\Sigma_Y^{-1}(y - \mu_Y) \\
\Sigma_{X|Y} &= \Sigma_X - \Sigma_{XY}\Sigma_Y^{-1}\Sigma_{XY}^\top.
\end{aligned}$$

   Επιπλέον, να εξειδικευθεί το παραπάνω αποτέλεσμα για $m = n = 1$ χρησιμοποιώντας τυπικές αποκλίσεις και συντελεστή γραμμικής συσχέτισης, όπως επίσης και για την περίπτωση που $m = 1$ (με αυθαίρετο n), όταν οι συνιστώσες της Y δοθείσης της τ.μ. X είναι ανεξάρτητες (καλούνται τότε δεσμευμένα ανεξάρτητες).
3. Υποθέτουμε τώρα ότι σε μία μπεϋζιανή προσέγγιση εκτίμησης έχουμε n παρατηρήσεις $\{X_i\}_{i=1}^n$ για τις οποίες υποθέτουμε ότι είναι δεσμευμένα ανεξάρτητες με $[X_i|\theta] \sim \mathcal{N}(\theta, \sigma_x^2)$ και θέλουμε να βρούμε την εκ των υστέρων κατανομή του θ, όταν το σ_x^2 θεωρείται γνωστό και a priori $\theta \sim \mathcal{N}(\mu, \sigma^2)$, όπου μ, σ^2 γνωστά. Να βρεθούν η περιθώρια κατανομή της $X = (X_1, \ldots, X_n)$ και η εκ των υστέρων κατανομή $\big[\theta|X = x\big]$.

Εβδομάδες 5-6 - Ασκήσεις

Οι ασκήσεις των Εβδομάδων 5 και 6 αφορούν αποτελέσματα που σχετίζονται με διάφορες μορφές του ελέγχου χ^2.

2.1: Στο πρόβλημα με τον ισχυρισμό του ερευνητή, συμπεραίνουμε εύκολα ότι οι εκτιμήτριες

$$\begin{aligned}\widehat{p}_n &= \frac{N_{1,n}+N_{2,n}+N_{3,n}+N_{4,n}}{6n}\\ \overline{p}_n &= \frac{N_{1,n}+N_{2,n}+0.5N_{3,n}+0.5N_{4,n}}{4n}\end{aligned}$$

είναι αμερόληπτες και συνεπείς εκτιμήτριες του p.

1. Να βρεθούν οι εκτιμήτριες $\widehat{p}_n^{\mathrm{LX}}$ ελαχίστου χ^2 (ελαχιστοποιώντας την $Q_n(p)$) και ελαχίστων τετραγώνων $\widehat{p}_n^{\mathrm{LS}}$.
2. Να βρεθεί η ασυμπτωτική κατανομή κάποιων εκτιμητριών από τις παραπάνω και να συγκριθεί η ασυμπτωτική τους διασπορά [Ζητείται το όριο κατά κατανομή: $\sqrt{n}(\widetilde{p}_n - p)$]
3. Θέτουμε $Q_n(\widetilde{p}_n)$ το στατιστικό χ^2 για τις παραπάνω εκτιμήτριες στα 6 κελιά. Να βρεθεί η ασυμπτωτική κατανομή κάποιων εκτιμητριών από τις παραπάνω.

2.2: Στο πρόβλημα με την κατανομή Poisson υπενθυμίζουμε τα δεδομένα και τον πίνακα πιθανοτήτων και αναμενόμενων τιμών για απλή μηδενική $\mathcal{P}(1)$. Θα διερευνήσουμε λίγο βαθύτερα την ισχύ του ελέγχου χ^2. Η σύσταση διεξαγωγής του ελέγχου αφορά την ομαδοποίηση σε κλάσεις έτσι ώστε $E_i \geq 5$. Στο συγκεκριμένο παράδειγμα $n = 20$, σχετικά μικρό.

Πινακας 2.1: Δεδομένα Παραδείγματος

i	0	1	2	3	4	5+	Άθροισμα
O_i	3	3	6	5	3	0	20
$p_{i,0}$	0.368	0.368	0.184	0.061	0.015	0.004	1
E_i	7.358	7.358	3.679	1.227	0.307	0.073	20

1. Κάτω από την H_0 πόσο άραγε "απέχει" η κατανομή της ελεγχοσυνάρτησης από την θεωρητική ? Για $\lambda = 1$ προσομοιώστε $M = 10^5$ φορές δεδομένα μεγέθους $n = 20$. Σκεφτείτε ότι κάθε φορά η ομαδοποίηση θα ήταν σταθερή, αφού εξαρτάται από τις αναμενόμενες τιμές και θα οδηγούσε σε 3 κελιά. Υπολογίστε τις τιμές του στατιστικού στα M δείγματα και φτιάξτε ένα ιστόγραμμα συχνοτήτων. Αντιπαραβάλλετε το ιστόγραμμα με το γράφημα της χ_2^2 και τις τιμές των δειγματικών άνω ποσοστημορίων για $a \in \{0.01, 0.05, 0.1\}$ με τα αντίστοιχα θεωρητικά. Κάντε και έναν έλεγχο Kolmogorov-Smirnov με θεωρητική κατανομή την χ_2^2. Διατυπώστε τα συμπεράσματά σας.
2. Επαναλάβετε το παραπάνω πείραμα κάνοντας έλεγχο της οικογένειας Poisson. Σε αυτήν την περίπτωση προσομοιώστε $M = 10^5$ φορές, δεδομένα μεγέθους $n = 20$,

με $\lambda = 1$. Εκτιμήστε το λ κάθε φορά με την ε.μ.π. και αλλάξτε την ομαδοποίηση, αν αυτό χρειαστεί, τηρώντας πάντα τη σχέση $\widehat{E}_i \geq 5$ για να υπολογίσετε την τιμή του στατιστικού. Αντιπαραβάλλετε το ιστόγραμμα με το γράφημα της χ_1^2 και τις τιμές των δειγματικών άνω ποσοστημορίων για $a \in \{0.01, 0.05, 0.1\}$ με τα αντίστοιχα θεωρητικά. Κάντε και έναν έλεγχο Kolmogorov-Smirnov με θεωρητική κατανομή την χ_1^2. Διατυπώστε τα συμπεράσματά σας.

3. Επαναλάβετε το παραπάνω πείραμα προσομοιώνοντας από $\lambda = 5$. Επιπλέον, κάντε έλεγχο ομοιογένειας Kolmogorov-Smirnov 2 δειγμάτων χρησιμοποιώντας κατάλληλα την εντολή ks.test στην R.
4. Επαναλάβετε τα παραπάνω 2 πειράματα με σταθερές κλάσεις, εκείνες που προκύπτουν από την ε.μ.π. των αρχικών δεδομένων.