Κατανομές Πιθανότητας στους Πραγματικούς

Η προσπάθεια μας έγκειται στην διατύπωση εννοιών που μας επιτρέπουν να περιγράφουμε κατανομές πιθανότητας στο $\mathbb{R}$, αποφεύγοντας τον δύσχρηστο ορισμό. Μια τέτοια επιβοηθητική έννοια είναι αυτή του στηρίγματος.

Η Έννοια του Στηρίγματος

Για τα παρακάτω επισημαίνεται χωρίς μεγάλη ακρίβεια ότι κάποιο υποσύνολο του $\mathbb{R}$ θεωρείται κλειστό αν το συμπλήρωμα του θεωρείται ανοικτό. Παραδείγματα είναι το $\mathbb{R}$, το $\emptyset$, τα μονοσύνολα, τα κλειστά διαστήματα, τα διακριτά υποσύνολα του $\mathbb{R}$, πεπερασμένες ενώσεις και αυθαίρετου πλήθους τομές αυτών κ.ο.κ.

Αναλόγως όταν $A, B \subseteq \mathbb{R}$ τότε το A θα αναφέρεται μικρότερο του B ανν $A \subset B$ (το A γνήσιο υποσύνολο του B).

Ορισμός. Έστω κατανομή πιθανότητας P στο $\mathbb{R}$. Το στήριγμα (supp - από το support) αυτής είναι το μικρότερο κλειστό υποσύνολο του $\mathbb{R}$ στο οποίο η P αποδίδει μοναδιαία πιθανότητα.

Σχόλια: Είναι δυνατόν να αποδειχθεί ότι το supp υπάρχει και είναι μοναδικό για κάθε P, επειδή εξ ορισμού του επιβάλλεται να είναι κλειστό. Προφανώς $\text{supp} \in \Sigma_{\mathbb{R}}$ (γιατί;).

Η χρησιμότητα της έννοιας φαίνεται στο παρακάτω αποτέλεσμα:

Πόρισμα: Αν $A \in \Sigma_{\mathbb{R}}$ τότε $P(A) = P(A \cap supp)$.

Απόδειξη. Αφού $A, supp \in \Sigma_{\mathbb{R}}$ τότε (γιατί;)

$$P(A) = P(A \cap supp) + P(A \cap supp')$$

αλλά (γιατί;) $P(A \cap supp') \leq P(supp') = 1 - P(supp) =$
$= 1 - 1 = 0$.

Επομένως $P(A \cap supp') \leq 0 \Rightarrow P(A \cap supp') = 0$ (γιατί;)

οπότε και προκύπτει το ζητούμενο. $\square$

Παράδειγμα: Στο παράδειγμα της προηγούμενης διάλεξης για την P^* έχουμε (γιατί;) ότι όταν $q \in (0,1)$ $supp = \{0,1\}$ (ποιό είναι το supp όταν $q=0$, $q=1$;). Θυμηθείτε ότι αυτό αποτέλεσμα παράδειγμα κατανομής στο $\mathbb{R}$ που περιγράφεται "εύκολα", και αυτό έχει σχέση με το ότι το στήριγμα είναι διακριτό (δείτε παρακάτω).

Το προηγούμενο πόρισμα μας λέει ότι ο υπολογισμός του $P(A)$ ανάγεται στον υπολογισμό του $P(A \cap supp)$ κάτι το οποίο μπορεί να αποτελεί διευκόλυνση (γιατί;).

Η έννοια του στηρίγματος ανάλογα με την "μορφή" που αυτό μπορεί να πάρει μας παρέχει καταρχάς την παρακάτω κατηγοριοποίηση των κατανομών στο $\mathbb{R}$:

a. Η P ονομάζεται διακριτή αν το στήριγμά της είναι διακριτό υποσύνολο του $\mathbb{R}$ (π.χ. πεπερασμένο, $\mathbb{N}$, $\mathbb{Z}$, κ.ο.κ.).

β. Η P θα ονομάζεται συνεχής αν το στήριγμα της είναι διάστημα.

γ. Η P θα ονομάζεται μεικτή σε κάθε άλλη περίπτωση (προσοχή: ο όρος αυτός χρησιμοποιείται μόνο στα πλαίσια του μαθήματος για λόγους οικονομίας ορολογίας και δεν αποδίδει αναγκαστικά καλά το πόσο περίπλοκη μπορεί να είναι μια κατανομή που εμπεριέχεται σε αυτή την κατηγορία. Π.χ. σε αυτή θα ανήκουν κατανομές με supp $=[0,1]\cup\{2\}$ αλλά και κατανομές με περισσότερο "περίπλοκο" στήριγμα - δείτε αν θέλετε το σχετικό λήμμα της Wikipedia για την κατανομή Cantor (Cantor Distribution)).

Διακριτές Κατανομές

Το supp μιας διακριτής κατανομής είναι εξ' ορισμού διακριτό, το οποίο σημαίνει ότι αποτελείται από "απομονωμένους" πραγματικούς, με την έννοια ότι για κάθε έναν από αυτούς μπορούμε να βρούμε ανοικτό διάστημα με κέντρο αυτόν και περατωμένη ακτίνα, το οποίο δεν θα περιλαμβάνει κανένα άλλο μέρος του supp. Διακριτά είναι αναγκαστικά τα πεπερασμένα υποσύνολα του $\mathbb{R}$. Υπάρχουν και διακριτά απειροσύνολα υποσύνολα του $\mathbb{R}$ (π.χ. $\mathbb{N}$, $\mathbb{Z}$, άρτιοι, περιττοί), ενώ είναι δυνατόν να αποδειχθεί ότι το πλήθος των στοιχείων ένος απειροσύνολου διακριτού είναι αναγκαστικά ίσο με το πλήθος των στοιχείων του $\mathbb{N}$ (επομένως στα παρακάτω είναι άμεσα εφαρμόσιμη η ιδιότητα της αριθμήσιμης προσθετικότητας).

Για μια τέτοια κατανομή λοιπόν θα γράφουμε το supp ως $\{x_1, x_2, \ldots, x_n, \ldots\}$. Εξαιτίας της μορφής του η περιγραφή μιας διακριτής κατανομής δεν χρειάζεται "περαιτέρω έννοιες", αφού:

$$\text{αν } A\in\Sigma_{\mathbb{R}},\quad P(A) \underset{(\text{γιατί;})}{=} P(A\cap \text{supp}) = P(A\cap\{x_1,x_2,\ldots,x_n,\ldots\})$$

$$= P\big((A\cap\{x_1\})\cup(A\cap\{x_2\})\cup\dots\cup(A\cap\{x_n\})\cup\dots\big) \underset{\text{αρ. προσθ.}}{=}$$

$$= P(A\cap\{x_1\}) + P(A\cap\{x_2\}) + \dots + P(A\cap\{x_n\}) + \dots$$

ενώ $P(A\cap\{x_i\}) = \begin{cases} P(\emptyset), & x_i\notin A \\ P(\{x_i\}), & x_i\in A \end{cases} = \begin{cases} 0, & x_i\notin A \\ P(\{x_i\}), & x_i\in A \end{cases}$

$\forall n:\ x_n \in supp$.

Επομένως για τον υπολογισμό της $P(A)$ αρκεί να γνωρίζουμε τις πιθανότητες που αποδίδει η κατανομή σε κάθε στοιχείο του στηρίγματος. Το παρακάτω πόρισμα μας λέει ότι σε αυτά και μόνο σε αυτά αποδίδει μια διακριτή κατανομή θετική πιθανότητα.

Πόρισμα. Αν η P διακριτή τότε $P(\{x\})>0$ ανν $x\in supp$.

Απόδειξη. Έστω ότι $P(\{x\})=0$ για $x\in supp$. Τότε $supp-\{x\}$ είναι επίσης διακριτό, άρα αριθμήσιμο, ενώ $supp-\{x\}\subset supp$. Επίσης $P(supp-\{x\}) \underset{(\text{γιατί;})}{=} P(supp)-P(\{x\}) = 1-0=1$, και άρα υπάρχει και μικρότερο του supp αριθμήσιμο υποσύνολο του $\mathbb{R}$ μοναδιαίας P-πιθανότητας. Αντίστροφα, αν $x\in supp'$ τότε $P(\{x\})\le P(supp') = 0$, επομένως $P(\{x\})=0$. □

Σχόλιο: Σε μη διακριτές κατανομές είναι δυνατόν $P(\{x\})=0$ για κάποιο $x\in supp$. Τα παραπάνω μας λένε ότι είναι δυνατόν να περιγράψουμε μια διακριτή κατανομή δίνοντας το στήριγμα και τις πιθανότητες που αυτή αποδίδει σε κάθε στοιχείο του στηρίγματος.

Σημειώστε ότι η συνάρτηση $p: supp\to\mathbb{R}$ ορισμένη ως $p(x):=P(\{x\})$ ονομάζεται συνάρτηση μάζας πιθανότητας της διακριτής κατανομής P (pmf - probability mass function) ενώ συνήθως επεκτείνεται σε όλο το $\mathbb{R}$, ως $p(x):=\begin{cases} P(\{x\}), & x\in supp \\ 0, & x\notin supp \end{cases}$.

Παραδείγματα Διακριτών Κατανομών στους Πραγματικούς

Στα παρακάτω θα εξετάσουμε παραδείγματα διακριτών κατανομών, ορίζοντας τις όπως αναφέραμε παραπάνω. Προκειμένου να ελέγξουμε το κατά πόσο ορίζουν κάτι καλά, θα πρέπει να διαπιστώνουμε:
I. Είναι το δεδομένο supp διακριτό;
II. Ισχύει ότι $P(\{i\})>0,\ \forall i\in supp$;
III. Ισχύει ότι $P(supp)=1$;

Εφόσον τα I, II, III τηρούνται τότε το παράδειγμα θα ορίζει μια καλά ορισμένη διακριτή κατανομή στο $\mathbb{R}$.

1. Εκφυλισμένη κατανομή στο 0.

$$supp=\{0\}$$
$$P(\{0\})=1$$

Προφανώς οι I, II, III τηρούνται. (Πως θα ορίζονταν η εκφυλισμένη κατανομή στο x όπου x αυθαίρετος πραγματικός; Πόσες τέτοιες υπάρχουν;)

2. Έστω $q\in(0,1)$. Κατανομή Bernoulli με παράμετρο q ($Ber(q)$).

$$supp=\{0,1\}$$
$$P(\{0\})=1-q$$
$$P(\{1\})=q$$

Προφανώς οι I, II, III τηρούνται αφού $q\neq 0,1$ (γιατί;)
Το παράδειγμα ορίζει επί της ουσίας μια οικογένεια από κατανομές Bernoulli, μία για κάθε τιμή του q, επομένως τόσες όσες τα στοιχεία του $(0,1)$.

3. Έστω $n\in\mathbb{N}^*$, $q\in(0,1)$. Διωνυμική κατανομή με παραμέτρους (n,q) [Binomial Distribution - $Bin(n,q)$]

$$supp=\{0,1,\dots,n\}$$
$$P(\{i\})=\binom{n}{i}q^i(1-q)^{n-i},\ \forall i\in supp,$$

όπου $\binom{n}{i} = \frac{n!}{i!(n-i)!}$, και $j! = \begin{cases} 1\cdot 2\cdots j, & j=1,2,\dots \\ 1, & j=0 \end{cases}$.

Οι I-II προφανώς τηρούνται εφόσον $q\in(0,1)$. Για το III δίνεται ότι αν $a,b\in\mathbb{R}$, $n\in\mathbb{N}$ τότε $(a+b)^n = \sum_{i=0}^{n} \binom{n}{i} a^i b^{n-i}$

το οποίο ονομάζεται διωνυμικό ανάπτυγμα – binomial expansion
π.χ. όταν $n=2$ το παραπάνω αποδίδει το τετραγωνικό ανάπτυγμα,
οπότε $P(\text{supp}) = P(\{0,1,\dots,n\}) \underset{(\text{γιατί;})}{=} \sum_{i=0}^{n} P(\{i\}) = \sum_{i=0}^{n} \binom{n}{i} q^i (1-q)^{n-i} \underset{\substack{a=q\\ b=1-q}}{=}$

$= (q+1-q)^n = 1.$

Επομένως τα παραπάνω καλώς ορίζουν την εν λόγω κατανομή. Σημειώνεται ότι όταν $n=1$ αποκτούμε την $Ber(q)$ [δηλ. $Ber(q) = Bin(1,q)$ (γιατί;)]. Και πάλι τα παραπάνω επί της ουσίας ορίζουν μια οικογένεια από διωνυμικές κατανομές, αφού αποκτούμε μια διαφορετική $Bin(n,q)$ για κάθε διαφορετική δυνατή τιμή του (n,q).

4. Έστω $\lambda>0$, η κατανομή Poisson με παράμετρο λ $Pois(\lambda)$ ορίζεται από:

$$\text{supp} = \mathbb{N}$$
$$P(\{i\}) = e^{-\lambda}\frac{\lambda^i}{i!}, \quad \forall i\in\mathbb{N}.$$

Προφανώς οι I-II ισχύουν, ενώ για την III δίνεται ότι $e^x = \sum_{i=0}^{\infty} x^i/i!$, $\forall x\in\mathbb{R}$ - ανάπτυγμα McLaurin της εκθετικής συνάρτησης. Για λόγους εκτός του εύρους του μαθήματος σημειώνουμε ότι μπορούμε να μεταχειριζόμαστε αλγεβρικά τα απειροπληθή αθροίσματα που θα συναντήσουμε - σειρές - όπως και τα αθροίσματα πεπερασμένου πλήθους προσθετέων. Συνεπώς, $P(\text{supp}) = P(\mathbb{N}) = P(\{0,1,2,\dots\}) =$ (γιατί;)

$$\sum_{i=0}^{\infty} P(\{i\}) = \sum_{i=0}^{\infty} e^{-\lambda}\frac{\lambda^i}{i!} = e^{-\lambda}\sum_{i=0}^{\infty}\frac{\lambda^i}{i!} \underset{x=\lambda}{=} e^{-\lambda}e^{\lambda} = e^{0} = 1.$$

Επομένως η παραπάνω είναι καλώς ορισμένη.
Π.χ. για την εν λόγω κατανομή έχουμε:

$$P(\mathbb{Z}) = P(\mathbb{Z}\cap\mathbb{N}) = P(\mathbb{N}) = 1,$$
$$P((0,1)) = P((0,1)\cap\mathbb{N}) = P(\emptyset) = 0,$$
$$P([0,1]) = P([0,1]\cap\mathbb{N}) = P(\{0,1\})$$
$$= P(\{0\}) + P(\{1\}) = e^{-\lambda}\frac{\lambda^0}{0!} + e^{-\lambda}\frac{\lambda^1}{1!} = (1+\lambda)e^{-\lambda}.$$

Και πάλι: σημειώνεται ότι τα παραπάνω ορίζουν την οικογένεια από κατανομές Poisson, αφού αποκτούμε μια διαφορετική κατανομή για κάθε διαφορετική δυνατή τιμή του λ, ενώ το πλήθος των μελών αυτής της οικογένειας ισούται με το πλήθος των στοιχείων του $(0,+\infty)$ (γιατί;).

Οι σημειώσεις βρίσκονται σε στάδιο διαρκούς διόρθωσης και δεν υποκαθιστούν τις διαλέξεις. Παρακαλώ αναφέρτε όποιο λάθος στο stelios@aueb.gr ή στο e-class του μαθήματος.