Γνωστές Κατανομές Διακριτών Τ.Μ. 5η Διάλεξη 26/3/2024 1

↳ Σκοπός Στατιστικής: Χρησιμοποιώντας δειγματικά δεδομένα $Y_1, \dots, Y_n$ να εξαχθούν συμπεράσματα για έναν ολόκληρο πληθυσμό.

Παραμετρική Στατιστική | Μη Παραμετρική Στατιστική

→ 1ο Βήμα: Υπόθεση για την κατανομή (τυχαίος μηχανισμός) που "γέννησε" τα δεδομένα που παρατηρούμε στο δείγμα που εξετάζουμε.

2ο Βήμα: Για την κατανομή του 1ου βήματος εκτίμηση των παραμέτρων της έτσι ώστε η συγκεκριμένη κατανομή (με τις συγκεκριμένες) τιμές παραμέτρων να ταιριάζει όσο το δυνατόν περισσότερο με τα δεδομένα που μελετούμε.

Για την εκτέλεση του 1ου βήματος απαιτείται η γνώση κάποιων κοινά χρησιμοποιούμενων κατανομών. Παρακάτω παρατίθεται μια λίστα με τις πιο κοινές κατανομές (συνεχείς κ' διακριτές)

Διακριτές Κατανομές

1) **Διακριτή Ομοιόμορφη κατανομή**: Χρησιμοποιείται στην περίπτωση πεπερασμένου δειγματικού χώρου και όταν όλα τα πιθανά ενδεχόμενα συμβαίνουν με την ίδια πιθανότητα.

↳ **Ορισμός**: Για κάθε $x \in \{1, .., N\}$

$$f(x) = P(X=x) = \frac{1}{N}$$

$$E(x) = \frac{N+1}{2}, \quad Var X = \frac{(N+1)(N-1)}{12}$$

Άσκηση: Ν.δ.ο. η $f(x)$ είναι συνάρτηση πιθανότητας κ' να αποδείξετε τους τύπους για $E(x)$ κ' $Var X$.

Λύση

Έχουμε ότι $\sum_x f(x) = \sum_{x=1}^{N} \frac{1}{N} = \underbrace{\frac{1}{N} + \frac{1}{N} + \dots + \frac{1}{N}}_{N \text{ φορές}} =$

$= N \cdot \frac{1}{N} = 1$ (i) και (ii) $f(x) \geq 0 \ \forall x$

άρα η $f(x)$ είναι συνάρτηση πιθανότητας

$$E(X) = \sum_x x f(x) = \sum_{x=1}^{N} x \frac{1}{N} = 1 \cdot \frac{1}{N} + 2 \cdot \frac{1}{N} + \dots + N \frac{1}{N} =$$

$$= \frac{1+2+\dots+N}{N} = \frac{\frac{\cancel{N}(N+1)}{2}}{\cancel{N}} =$$

$$= \frac{N+1}{2}$$

$$\operatorname{Var} X = EX^2 - (EX)^2 = EX^2 - \frac{(N+1)^2}{4}$$

$$EX^2 = \sum_x x^2 f(x) = \sum_{x=1}^{N} x^2 \frac{1}{N} = \frac{1}{N}\sum_{x=1}^{N} x^2 = \frac{1}{\cancel{N}} \frac{\cancel{N}(N+1)(2N+1)}{6}$$

Οπότε

$$EX^2 - (EX)^2 = \frac{(N+1)(2N+1)}{6} - \frac{(N+1)^2}{4} =$$

$$= \frac{(2N+2)(2N+1) - 3(N+1)^2}{12} =$$

$$= \frac{\cancel{4N^2} + \cancel{2N} + \cancel{4N} + 2 - \cancel{3N^2} - \cancel{6N} - 3}{12}$$

$$= \frac{N^2 - 1}{12} = \frac{(N-1)(N+1)}{12}$$

2) Bernoulli Κατανομή: Χρησιμοποιείται όταν το υπό μελέτη πείραμα έχει μόνο 2 πιθανά αποτελέσματα (Επιτυχία/αποτυχία) και η πιθανότητα Επιτυχίας είναι ίση με p. Συμβολικά γράφουμε

$$X \sim \text{Bernoulli}(p)$$

Για $x \in \{0, 1\}$

παράμετρος της κατανομής

$$f(x) = P(X = x) = p^x (1-p)^{1-x}$$

$$E(X) = p \qquad \operatorname{Var} X = p(1-p)$$

Παραδείγματα: i) Ρίψη νομίσματος (μοναδικής)
ii) Το αποτέλεσμα της προσπάθειας ενός πωλητή να πουλήσει ένα προϊόν

3) Διωνυμική Κατανομή : Χρησιμοποιείται όταν έχουμε n ανεξάρτητες επαναλήψεις ενός πειράματος με 2 πιθανά αποτ. και σε κάθε επανάληψη η πιθ. επιτυχίας είναι p

$$X \sim Bin(\underbrace{n, p}_{\text{παράμετροι της κατανομής}})$$

• Για $x \in \{0, 1, .., n\}$

$$f(x) = P(X=x) = \binom{n}{x} p^x (1-p)^{n-x} =$$

$$= \frac{n!}{(n-x)!\,x!} p^x (1-p)^{n-x}$$

$$EX = np \quad , \quad Var X = np(1-p)$$

Ιδιότητες

• Για $n=1$ η Bernoulli(p) είναι ειδική περίπτωση της Binomial(n, p).

• Εάν $X \sim Bin(n, p)$ κ' $Y \sim Bin(m, p)$

τότε $X+Y \sim Bin(n+m, p)$

Παράδειγμα: Ένα προϊόν βιομηχανικής παραγωγής μπορεί να είναι ελαττωματικό με πιθανότητα 0.005 ⑤

α) Ποια η πιθ. να έχουμε ένα τουλάχιστον ελατ. σε δείγμα μεγέθους 15?

β) Εάν κάθε μέρα παράγονται 1000 προϊόντα πόσα αναμένεται να είναι ελαττωματικά

Λύση

α) Έστω X η τ.μ. που "μετράει" τον αριθμό των ελαττωματικών σε 15 προσπάθειες

Τότε $X \sim Bin(15, 0.005)$

$$P(X \geq 1) = 1 - P(X=0) =$$

$$1 - \binom{15}{0} p^0 (1-p)^{15-0} =$$

$$1 - \frac{15!}{(15-0)!\,0!} 0.005^0 \cdot 0.995^{15} = 1 - 0.995^{15} =$$

$$= 0.072 = 7,2\%$$

β) $X \sim Bin(1000, 0.005)$

$$EX = 1000 \cdot 0.005 = 5$$

4) Poisson Κατανομή: Χρησιμοποιείται για να "μετρήσουμε" τον αριθμό γεγονότων σε μια συγκεκριμένη χρονική περίοδο.

$$X \sim Pois(\lambda)$$

Για $x \in \{0, 1, 2, \dots\}$, $f(x) = \dfrac{e^{-\lambda}\lambda^x}{x!}$

$EX = VarX = \lambda$

Ιδιότητες:

- Αν $X_i \sim Pois(\lambda_i)$, $i = 1, \dots, n$ και X_i ανεξ. τ.μ.

  τότε $\sum_{i=1}^{n} X_i \sim Pois\left(\sum_{i=1}^{n} \lambda_i\right)$

Παράδειγμα: Ο αριθμός των μη θανατηφόρων τροχαίων ανά εβδομάδα ακολουθεί την Poisson κατανομή με μέσο 1.

α) Ποια η πιθ. να έχουμε τουλάχιστον ένα ατύχημα κάποια εβδομάδα

β) Ποια η πιθ. μέσα σε ένα μήνα να μη συμβούν ατυχήματα

Λύση

α) $X \sim Pois(1)$, $P(X \geq 1) = 1 - P(X = 0)$

$$= 1 - e^{-1}\frac{1^0}{0!} = 1 - \frac{1}{e}$$

$$= 0.632 = 63,2\%$$

β) Αφού την εβδομάδα έχουμε κατά μέσο όρο 1 ατύχημα, το μήνα θα έχουμε κατά μέσο όρο 4 ατυχήματα.

Άρα αν X ο αριθμός μη θανατηφόρων ατυχημάτων σε ένα μήνα τότε

$$X \sim Pois(4)$$

άρα $P(X=0) = e^{-4}\frac{4^0}{0!} = \frac{1}{e^4} = 0,018 = 1,8\%$

* Άλλες κοινές διακριτές κατανομές:

↳ Αρνητική Διωνυμική
↳ Γεωμετρική
↳ Υπεργεωμετρική
...

→ Συνεχείς Κατανομές:

1) ~~Ομοι~~ Ομοιόμορφη κατανομή στο διάστημα (a,b)

Για $x \in (a,b)$ έχουμε $f(x) = \frac{1}{b-a}$

$EX = \frac{b+a}{2}$, $Var\,X = \frac{(b-a)^2}{12}$

Παράδειγμα: Έστω ότι επισκεπτόμαστε μια πόλη στο εξωτερικό και φτάνουμε στον σταθμό τρένων. Ξέρουμε μόνο ότι τρένο για τον προορισμό μας περνάει κάθε 45 λεπτά αλλά όχι την ακριβή ώρα α) Ποια η πιθ. να περιμένουμε το πολύ 10 λεπτά; (8)

β) Ποιος ο αναμενόμενος χρόνος μέχρι την επόμενη αναχώρηση

Λύση

Έστω X ο χρόνος σε λεπτά ανάμεσα σε 2 τρένα

Τότε $X \sim U(0, 45)$

α) $P(X \leq 10) = \int_0^{10} \frac{1}{45} dx = \frac{X}{45}\Big|_0^{10} = \frac{10}{45} = 0,22 = 22\%$

β) $EX = \frac{0+45}{2} = 22,5$ λεπτά

2) Κανονική Κατανομή (Normal or Gaussian distribution)

$X \sim N(\mu, \sigma^2)$ → παράμετροι

→ Η πιο συχνά χρησιμοποιούμενη κατανομή στην Στατιστική και την Οικονομετρία

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

$EX = \mu$, $Var X = \sigma^2$

Παράδειγμα: Μια εταιρεία παράγει μεταλλικούς άξονες των οποίων το μήκος ακολουθεί την κανονική κατανομή με μέση τιμή 0,2508cm και τυπική απόκλιση 0,0015cm. Για να είναι αποδεκτός στην αγορά ο άξονας πρέπει να έχει μήκος
0,2500± 0,0015 cm

α) Ποιο το ποσοστό των παραγόμενων αξόνων πληροί τις προδιαγραφές;

β) Αν μπορούσαμε να διορθώσουμε τη γραμμή παραγωγής σε ποιο μέσο θα πρέπει να στοχεύσαμε για να ελαχιστοποιηθούν οι παραγόμενοι μη αποδεκτοί άξονες.

Λύση

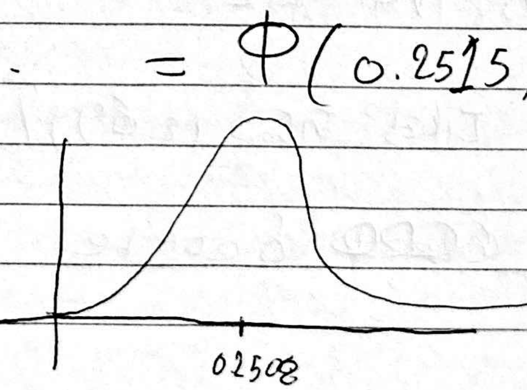
α) $X \sim N(\mu = 0{,}2508,\ \sigma^2 = 0.0015^2)$

$$P(0.2485 < X < 0.2515) =$$

$$= P(X < 0.2515) - P(X < 0.2485) =$$

$$= \int_{-\infty}^{0.2515} \frac{1}{0.0015\sqrt{2\pi}} e^{-\frac{1}{2\cdot 0.0015^2}(x - 0{,}2508)^2} dx -$$

$$- \int_{-\infty}^{0.2485} \dots = \Phi(0.2515) - \Phi(0.2485)$$



Ιδιότητες:

• Αν $\mu = 0$ και $\sigma = 1$ τότε η κατανομή λέγεται τυπική κανονική κατανομή και συμβολίζεται ως $Z \sim N(0,1)$

Ισχύει ότι: i) Αν $X \sim N(\mu, \sigma^2)$

τότε $\frac{X-\mu}{\sigma} \sim N(0,1)$

ii) Αν $Z \sim N(0,1)$ τότε

$X = \sigma Z + \mu \sim N(\mu, \sigma^2)$

• Ο κανόνας των τριών τυπικών αποκλίσεων

i) $P(|X-\mu| \leq \sigma) = 0,6826 = 68,26\%$

ii) $P(|X-\mu| \leq 2\sigma) = 0,9544 = 95,44\%$

iii) $P(|X-\mu| \leq 3\sigma) = 0,9973 = 99,73\%$

π.χ. αν $X \sim N(\mu = 15, \sigma^2 = 4)$

τότε $\sigma = \sqrt{4} = 2$

και οι τιμές της με 99% πιθ. ανήκουν

στο διάστημα $(9, 21)$

• Αν X_1, X_2 ανεξάρτητες κανονικές τ.μ. με μέσες τιμές μ_1, μ_2 και διασπορές σ_1^2, σ_2^2 τότε

$$X_1 \pm X_2 \sim N(\mu_1 \pm \mu_2, \sigma_1^2 + \sigma_2^2)$$

• Αν $X_i, i=1,\dots,n$ ανεξάρτητες τυπικές κανονικές (δηλ. $X_i \sim N(0,1)$)

$$\text{τότε } Y = \sum_{i=1}^{n} X_i^2 = \chi_n^2$$

• Αν $X_i \sim N(\mu, \sigma^2)$, $i=1,\dots,n$, ανεξάρτητες τότε

$$t = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1} \qquad \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

• Αν $X_i, i=1,\dots,n \sim N(0,1)$ και $Y_j, j=1,\dots,m \sim N(0,1)$ ανεξάρτητες τότε

$$F = \frac{S_X^2/n}{S_Y^2/m} \sim F_{n,m}$$

Αν $X \sim t_v$ τότε $X \in (-\infty, +\infty)$

$$\text{και } f(x) = \frac{\Gamma\left(\frac{v+1}{2}\right)}{\Gamma\left(\frac{v}{2}\right)} \frac{1}{\sqrt{v\pi}} \left(1 + \frac{x^2}{v}\right)^{-\frac{v+1}{2}}$$

$$EX = 0, \ (v>1) \qquad Var X = \frac{v}{v-2}, \ v>2$$

* Άλλες συνεχείς κατανομές: LogNormal
Gamma
Exponential
⋮

~ · ~ · ~ · ~ · ~ · ~ · ~ · ~

Πολυμεταβλητές Κατανομές:

↳ Διακριτό τυχαίο διάνυσμα είναι το ζευγάρι τ.μ. (X,Y) το οποίο παίρνει τιμές σε αριθμήσιμο σύνολο. Η συνάρτηση

$$f: \mathbb{R}^2 \to \mathbb{R} \text{ με}$$

$f(x,y) = P(X=x, Y=y)$ λέγεται από κοινού συνάρτηση πιθανότητας του (X,Y) εάν

i) $f(x,y) \geq 0$

ii) $\sum_x \sum_y f(x,y) = 1$

- Για $A \subseteq \mathbb{R}^2$ έχουμε ότι

$$P((X,Y)\in A) = \sum_{(x,y)\in A} f(x,y)$$

- Από κοινού μέση τιμή:

$$E(g(x,y)) = \sum_x \sum_y g(x,y) f(x,y)$$

• Περιθώριες Κατανομές

$$f_X(x) = P(X=x) = \sum_y f(x,y)$$
$$f_Y(y) = P(Y=y) = \sum_x f(x,y)$$

Παράδειγμα:

f(x, y) Y \ X	0	1	2	$f_Y(y)$
0	0.05	0.1	0.03	0.18
1	0.21	0.11	0.19	0.51
2	0.08	0.15	0.08	0.31
$f_X(x)$	0.34	0.36	0.30	1.00

• Από κοινού Κατανομή για συνεχείς τ.μ.

Αν (X, Y) συνεχές (δισδιάστατο) τυχαίο διάνυσμα τότε

$$P((X,Y) \in A) = \iint_A f(x,y)\,dx\,dy$$

και $f(x,y)$ ονομάζεται από κοινού συνάρτηση πυκνότητας πιθανότητας αν

i) $f(x,y) \geq 0$ και ii) $\iint_{\mathbb{R}^2} f(x,y) = 1$

• $$E(g(x,y)) = \iint_{\mathbb{R}^2} g(x,y) f(x,y)\,dx\,dy$$

• Περιθώριες Κατανομές $f_X(x) = \int_y f(x,y)\,dy$

$$f_Y(y) = \int_x f(x,y)\,dx$$