Лекция 13

Математикалық статистикаға кіріспе
Ықтималдықтық-статистикалық модель және математикалық статистиканың негізгі есептері

Математикалық статистика– бұл ықтималдықтар теориясына тектес (жақын) қолданбалық бағыттағы математикалық пән. Ол ықтималдықтар теориясының ұғымдары мен әдістеріне сүйенеді, бірақ та өзінің арнайы есептерін өз әдістерімен шешеді. Кез келген математикалық теория нақты құбылыстардың белгілі бір аумағын сипаттайтын қандай да бір модельдер аясында дамытылатыны белгілі. Статистикалық модельді анықтап және математикалық статистиканың арнайы айналысатын есептерін жете түсіну үшін алдымен ықтималдықтар теориясының маңызды нәтижелеріне қысқаша тоқтала кетелік.

Ықтималдықтар теориясында зерттелінетін кездейсоқ құбылыстардың математикалық модельдері $(\Omega, \mathcal{A}, P)$- *ықтималдық кеңістігі* ұғымына негізделеді, мұндағы: **а)** $\Omega = \{\omega\}$ - *элементар оқиғалар кеңістігі* (элементар оқиға ω - тәжірибенің жалғыз мүмкіндікті өзара сиыспайтын нәтижелері), $\Omega \neq \varnothing$ (бос жиын); **ә)** $\mathcal{A}$ - Ω -ның қандай да бір ішкі жиындарының σ -алгебрасы (оқиғалар өрісі). $\mathcal{A}$ -ның элемент-тері *кездейсоқ оқиғалар* (көбіне қысқаша *оқиғалар*) деп аталады. **б)** $P : \mathcal{A} \to [0,1]$– *ықтималдық (ықтималдықтық функция)*. Оқырманның есіне сала кетелік, бұл жағдайларда $\mathcal{A}$ σ -алгебрасы **А1 , А2', А3** шарттарын, ал P ықтималдығы **Р1, Р2, Р3 (** немесе **Р3'**, немесе **Р3''**, немесе **Р3''')** аксиомаларын қанағаттандырады. $(\Omega, \mathcal{A}, P)$ жалпы ықтималдық кеңістігі жағдайында P ықтималдығы толық анықталған деп есептелінеді де, ықтималдықтар теориясының негізгі есебі ретінде әлдеқайда қарапайым оқиғалардың белгілі ықтималдықтары арқылы әрқилы күрделі оқиғалардың ықтималдықтарын есептей алу әдістерін табу проблемалары түсініледі.

Бірақ та іс жүзінде жекелеген тәжірибелерді зерттеу барысында P ықтималдығы толық белгілі болу жағдайы өте сирек кездеседі. Көбіне алдын-ала тек P ықтималдығы қандай да бір берілген ықтималдықтар класы $\mathcal{P}$ -ның элементі ($P \in \mathcal{P}$) деп қана тұжырымдай аламыз. Бұл класс не $\mathcal{A}$ -да анықтауға болатын барлық ықтималдықтарды қамтуы мүмкін (толық анықталмағандық жағдайы); басқа жағдайларда одан көрі кедейлеу, белгілі бір айқын түрде берілген ықтималдықтар класына жатуы мүмкін (алдын-ала нендей де бір ақпарат, мәлімет белгілі болатын жағдай). Қалай десек те $\mathcal{P}$ - бұл берілген жағдайда (берілген экспериментті сипаттау үшін) мағынасы болатын P ықтималдықтарының жиынтығы. Егер $\mathcal{P}$ класы берілсе, онда *ықтималдықтық-статистикалық* (не-месе қысқаша *статистикалық*) модель берілген дейді және бұл модель ретінде $(\Omega, \mathcal{A}, \mathcal{P})$ үштігін түсінеді.

Мысал. Бернуллидің тәуелсіз сынақтар тізбегін қарастыралық. Егер i -ші сынақта "табыс" болса $\omega_i = 1$; сәтсіздік болса $\omega_i = 0$ ($i = 1, 2, \ldots, n$) деп белгілесек, онда $\Omega = \{\omega = (\omega_1, \omega_2, ..., \omega_n): \omega_i = 0,\ не\ \omega_i = 1\}$, $\mathcal{A} = \{A : A \subseteq \Omega\}$. Бұл модельде, егер табыс ықтималдығы p белгілі болса, онда ω элементар оқиғасының ықтималдығы $P(\omega) = p^{|\omega|}(1-p)^{n-|\omega|}$, мұндағы $|\omega| = \omega_1 + \omega_2 + ... + \omega_n$, функциясы арқылы анықталады.

Айталық, енді табыс ықтималдығы алдын-ала белгілі болмасын. Егер оны θ арқылы белгілесек (ескерте кетелік, бұдан былай қарай барлық уақытта дерлік белгісіз параметрді (скаляр, вектор) θ арқылы белгілейтін боламыз), онда біз алдын-ала тек $\theta \in \Theta = [0,1]$ болатыны туралы айта алған болар едік. Ендеше бұл жағдайда болжамдай алатын ықтималдықтарымыздың жиынтығы $\mathcal{P} = \{P_\theta, \theta \in \Theta\}$, мұндағы $P_\theta(\omega) = \theta^{|\omega|}(1-\theta)^{n-|\omega|}$.

Сонымен зерттелініп отырған эксперименттің ықтималдықтық моделінде P ықтималдығын беруде нендей де бір анықталмағандықтар болады, ал математикалық статистиканың негізгі міндеті, енді осы қолда бар статистикалық деректер (эксперименттің бақыланатын нәтижелері) арқылы алынған ақпараттарды, мәліметтерді пайдалана отырып P -ға байланысты анықталмағандықты азайту, қандай да бір қорытындыларға, шешімдерге келу. Басқаша айтқанда математикалық статистика- ол статистикалық қорытындылар туралы ғылым десек те болады. Белгілі бір мағынада математикалық статистика ықтималдықтар теориясының есептеріне кері есептерді шығарумен айналысады- ол алынған бақылаулар нәтижелері арқылы статистикалық модельдің құрылымын сипаттайды, дәлдендіреді, ол туралы қорытындылар жасайды.

Басқа да математикалық пәндер сияқты математикалық статистиканың да пайда болуы мен дамуы практикалық сұранымдар мен қажеттіліктерден туындады; қазіргі уақытта оның әдістері мен нәтижелері техниканың әртүрлі салаларында, экономикада, медицинада, геологияда, психологияда, социологияда, актуарлық математикада, қаржылық математикада т.с.с. кеңінен қолданылуда.

Енді бұл тарауда бұдан былай қарай қолданылатын негізгі белгілеулер мен терминдерге қысқаша тоқталайық.

Негізгі белгілеулер мен терминдер

Көптеген жағдайларда *статистикалық деректер* деп зерттелініп отырған эксперименттің нәтижесін сипаттайтын қандай да бір $X=(X_1, X_2 ,...,X_n)$ кездейсоқ шамаларының бақыланатын нәтижелерін айтамыз. Бұны әдетте былайша тұжырымдайды: эксперимент i -ші сынағының нәти-жесі X_i, $i=1,2,\ldots,n$, кездейсоқ шамасымен сипатталатын n сынақ жүргізуден тұрады. Бақыланатын кездейсоқ шамалардың жиынтығы $X=(X_1,X_2,...,X_n)$ *таңдама* деп, X_i шамалары– *таңдаманың элементтері*, ал олардың саны n - *таңдаманың көлемі* деп аталады. X таңдамасының қабылдайтын мәндерін кіші $x=(x_1, x_2,..., x_n)$ арқылы белгілейміз. Кейде бақыланатын кездейсоқ шамаларды белгілеу үшін бас Y, Z, U, V т.с.с. әріп-терін де пайдаланатын, ал олардың қабылдайтын мәндерін сәйкес y, z, u, ϑ т.с.с. кіші әріптермен белгілейтін боламыз.

X таңдамасының барлық мүмкін мәндер жиынын, яғни X-тің үлестірімі берілетін жиынды $\aleph=\{x\}$ арқылы белгілейік ($P_X(B)=\sum\limits_{x\in B} P\{X=x\},\ B\in\aleph$). Математикалық статистикада $\aleph$ кеңістігі *таңдамалық кеңістік* деп аталады.

Егер X көп өлшемді (абсолютті) үзіліссіз кездейсоқ шама болса, онда не $\aleph=R^n$ не $\aleph\subset R^n$; егер де X (көп өлшемді) дискретті кездейсоқ шама болса, онда $\aleph$ ақырлы не саналымды нүктелерден тұрады. Берілген жағдайда эксперименттің статистикалық моделі ретінде $(\aleph, \mathcal{P})$ жұбын түсінеміз, мұндағы $\mathcal{P}$ X-тің $\aleph$-те беріл-ген үлестірімдерінің класы. Бірақ, кез келген кездейсоқ шаманың үлестірімі оның үлестірім функциясы арқылы бірмәнді анықталатындықтан, әдетте статистикалық модель үлестірім функциялары терминінде беріледі. Сонымен статистикалық модель $\aleph$ таңдамалық кеңістігімен және $F_X(x)=P\{X_1\le x_1,..., X_n\le x_n\}$ белгісіз үлестірім функциясын қамтитын $\mathcal{F}$ үлестірім функциялар класы (жиынтығы) арқылы беріледі: $(\aleph, \mathcal{F})$.

Көбіне $X=(X_1, X_2 ,...,X_n)$ таңдамасының $X_1, X_2 ,...,X_n$ компо-ненталары тәуелсіз және олардың бәрі қандай да бір ξ кездейсоқ шамасымен бірдей үлестірілген жағдай қарастырылады. Бұл ξ кездейсоқ шамасы бақыланатын тәуелсіз сынақтарды қайталаудан тұратын экспериментке сәйкес келеді ($X_i - \xi$ -дің i -ші “данасы”, “көшірмесі”) және де бұл жағдайда

$$F_{X_i}(x_i) = F_{\xi}(x_i) = P\{\xi \le x_i\},\ i = 1,2,\ldots,n;$$

$$F_X(x) = F_{X_1}(x_1)\, F_{X_2}(x_2) \ldots F_{X_n}(x_n) .$$

Мұндай модельді $F_{\xi}(x)$ функциясы терминінде беруге болады және бұл модель үшін *X=(X₁, X₂ ,...,Xₙ)* таңдамасы ξ кездейсоқ шамасының үлестірімінен алынған таңдама деп аталады. Кейде үлестірім функциясы $F_{\xi}(x)$ болатын ξ кездейсоқ шамасының мүмкін мәндер жиыны үлестірім функциясы $F_{\xi}(x)$ болатын *бас жиынтық* немесе жай *жиынтық* деп, ал X шамасы көлемі n -ге тең *таңдама* деп аталады. Бұдан былай қарай ξ -дің үлес-тірімін $G(\xi)$ арқылы белгілейтін боламыз (мәселен, $G(\xi) \in N\left(a, \sigma^2\right)$- ξ параметрлері $\left(a, \sigma^2\right)$ болатын нормаль кездейсоқ шама т.с.с.). Сонымен “$X=(X_1, X_2, \ldots, X_n)$ $G(\xi)$ үлестірімінен алынған таңдама” дегеніміз

$$F_{X_i}(x) = F_{\xi}(x) = P\{\xi \le x\}, \quad i = 1,2,\ldots, n ,$$

болатынын білдіреді.

Қайталама тәуелсіз бақылаулар үшін статистикалық модельді бұдан былай қарай қысқаша $F = \{F_{\xi}\}$ түрінде жазатын, ал материалдарды баяндау барысында (ешқандай түсініспеушілік болмайтын болса) үлестірім функциясындағы ξ индексін жазбайтын боламыз.

Егер F класының үлестірім функциялары өзгеру облысы Θ болатын қандай да бір θ параметріне (ол скаляр да, вектор да болуы мүмкін) дейінгі дәлдікпен берілсе, онда бұл модель $F = \{F(x,\theta) : \theta \in \Theta\}$ арқылы белгіленетін болады. Бұл бақыла-натын кездейсоқ шаманың үлестірім функциясының типі бел-гілі, бірақ үлестірім функциясы тәуелді болатын параметр θ белгісіз деген сөз. Θ жиыны *параметрлік* жиын деп аталады.

Мысалы, $G(\xi) \in N\left(\theta, \sigma^2\right)$ (σ^2 -белгілі, θ -белгісіз) үшін статистикалық модель $F = \{F(x,\theta), \theta \in \Theta = (-\infty, +\infty)\}$, мұндағы $F(x,\theta)$ тығыздығы

$$f(x,\theta)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\theta)^2}{2\sigma^2}},\ -\infty<x<\infty,$$

болатын үлестірім функциясы. Егер математикалық күтім де, дисперсия да белгісіз болса ($G(\xi)\in N\left(\theta_1,\theta_2^2\right)$), онда $F=\{F(x,\theta):\theta=(\theta_1,\theta_2)\in\Theta\}$, мұндағы

$$\Theta=\{(\theta_1,\theta_2):-\infty<\theta_1<\infty,\ 0<\theta_2<\infty\},$$

ал $F(x,\theta)$-ның үлестірім тығыздығы

$$f(x,\theta)=\frac{1}{\sqrt{2\pi}\theta_2}e^{-\frac{(x-\theta_1)^2}{2\theta_2{}^2}},\ -\infty<x<+\infty.$$

Егер $G(\xi)\in\Pi(\theta)$ болса, онда $F=\{F(x,\theta):\theta\in\Theta\},\Theta=(0,\infty)$,

$$F(x;\theta)=\sum_{y:y\le x}e^{-\theta}\frac{\theta^y}{y!}.$$

Егер F класының үлестірім функциялары абсолютті үзіліссіз немесе дискретті болса, онда $F=\{F_\xi\}$ статистикалық моделі сәйкес абсолютті үзіліссіз статистикалық модель не дискретті статистикалық модель деп аталады.

Бұдан былай қарай үнемі $f_\xi(x)$ арқылы (параметрлік модельдер үшін $f(x,\theta)$ арқылы) абсолютті үзіліссіз модельдер үшін - тығыздықты, дискретті модельдер үшін $P\{\xi=x\}$ ықти-малдықтарын белгілейтін боламыз. Сонымен бірге параметрлік модель үшін $\aleph$ таңдамалық кеңістігінде параметр θ -ға сәйкес келетін ықтималдықтық үлестірімді P_θ арқылы, ал P_θ үлестірімі бойынша алынатын математикалық күтім, дисперсия т.с.с. сәйкес M_θ, D_θ т.с.с. арқылы белгілейтін боламыз. Мәселен, берілген $T(x)$ функциясының X таңдамасындағы мәнінің математикалық күтімі

$$M_\theta T(X)=\int_{\aleph}T(x)dF_X(x;\theta)$$

Стильтес интегралы: егер $X=(X_1, X_2 ,...,X_n)$ абсолютті үзіліссіз болса, онда

$$M_\theta T(X)=\int_{\aleph}...\int_{\aleph}T(x_1,...,x_n)f(x_1;\theta)...f(x_n;\theta)dx_1...dx_n,$$

ал дискретті X үшін

$$M_{\theta}T(X)=\sum_{\aleph}T(x_1,x_2,\ldots,x_n)f(x_1;\theta)\ldots f(x_n;\theta).$$

Енді кейбір көп кездесетін аса маңызды статистикалық модельдерге қысқаша тоқталайық. Егер бақыланатын кездейсоқ шама ξ қандай да бір стандартты типті болса, мәселен $G(\xi)\in N\left(a,\sigma^2\right)$, $G(\xi)\in\Pi(\lambda)$ т.с.с., онда сәйкес статистикалық модельді де сол атпен атайтын боламыз. Мәселен, егер $G(\xi)\in N\left(a,\sigma^2\right)$ болса, онда нормаль модель туралы айтамыз (тағы да бір рет еске сала кетелік: бізде *X=(X₁, X₂ ,...,Xₙ)* таңдамасының компоненталары тәуелсіз кездейсоқ шамалар және олардың әрқайсысы ξ секілді үлестірілген

$$\left(F_{X_i}(x_i)=F_{\xi}(x_i)=P\{\xi\le x_i\}\right)).$$

Төменде кестеде жиі кездесетін статистикалық модельдер келтірілген (кестеде көрсетілмеген x-тің мәндері үшін $f(x,\theta)$ функциясы нөлге тең). Кестенің екінші бағанындағы жазудың мағынасы мынада: жалпы нормаль модель үшін

$$G(\xi)\in N\left(\theta_1,\theta_2^2\right)$$

жазуы модель $F=\{F(x,\theta),\theta\in\Theta\}$ үлестірім функциялар класы-мен берілетінін білдіреді, ал тығыздық $f(x,\theta)$ мен параметрлік жиын кестенің сәйкес 2-ші және 3-ші бағандарында келтірілген.

Нормаль модель статистикалық эксперименттің нәтижесіне әрқайсысының эксперименттің түпкілікті нәтижесіне тигізетін әсері өте аз болатын көптеген өзара тәуелсіз кездейсоқ факторлар әсер ететін жағдайларда пайда болады. Онда орталық шектік теореманың негізінде бақыланатын ξ кездейсоқ шамасы жуықтап алғанда параметрлері $\left(a,\sigma^2\right)$ болатын (олардың екеуі де белгісіз; не біреуі белгілі, екіншісі белгісіз болуы мүмкін) нормаль кездейсоқ шама болады деп тұжырымдай аламыз. Мәселен өлшемдердің қателері теория-сында біріктірілген (бір-біріне қосылған) қате әрқайсысының шамасы өте аз болатын көптеген кішкентай қателердің қосындысынан тұрады деп есептелінеді.

Биномдық үлестірім табысының ықтималдығы θ-ға тең Бернуллидің n тәуелсіз сынақтарындағы табыс санының үлестірімін сипаттайды. Бұл модельдің жеке жағдайы $-Bi(1,\theta)$ *бернуллилік моделі* ықтималдықтар теория-сы мен математикалық статистиканың қолданымдарында жиі кездеседі, өйткені тек қана екі нәтижесі бар эксперимент– ол ең қарапайым статистикалық эксперимент. Оның үстіне параметрі (n,θ) болатын биномдық кездейсоқ шаманы параметрлері $(1,\theta)$ болатын n тәуелсіз бернуллилік кездейсоқ шаманың қосындысы ретінде жазуға болатыны белгілі.

Теріс биномдық үлестірім $\overline{B}i(n,\theta)$- Бернуллидің тәуелсіз сынақтарындағы n-ші табыс $n+r$-ші сынақта пайда болуын білдіретін кездейсоқ шаманың үлестірімі, ал геометриялық үлестірім - ол бірінші сәтсіздікке дейінгі табыс-тар санының үлестірімі. $\overline{B}i(n,\theta)$ заңымен үлестірілген кездейсоқ шаманы $\overline{B}i(1,\theta)$ заңымен үлестірілген n тәуелсіз кездейсоқ шаманың қосындысы ретінде жазуға болатыны да белгілі.

$\Pi(\theta)$ *пуассондық моделі* әдетте сирек оқиғалардың схе-масын сипаттайды (телефон станциясына белгілі бір t уақыт аралығында келіп түскен қоңыраулар саны, есептеуішке белгілі бір t уақыт аралығында тіркелген радиоактивті бөлшектер саны т.с.с.)

$R(a,b)$ *бірқалыпты үлестірімі* (a,b) аралығынан кездейсоқ нүкте алуды сипаттайды. Мәселен, егер (a,b) интервалы аялдамадан автобустың екі көрші уақыт сәтінде кететін интервалы болса, онда аялдамаға келген жолаушының (автобустардың қозғалыс кестесі белгілі болмаған жағдайдағы) автобусты күту уақыты $R(a,b)$ үлестіріміне бағынады.

Реттік статистикалар және таңдаманың вариациялық қатары

Айталық, $X=(X_1,X_2,...,X_n)-G(\xi)$ үлестірімінен алынған көлемі n-ге тең таңдама (еске салалық, айтылғанды былай түсіну керек: $X_1,X_2,...,X_n$ әрқайсысы ξ кездейсоқ шама-сымен бірдей үлестірілген

$$(F_{X_i}(x)=P\{X_i\le x\}=P\{\xi\le x\}=F_{\xi}(x),\ x\in R)$$

тәуелсіз кездейсоқ шамалар, n-олардың саны (көлемі)), ал $x=(x_1,x_2,...,x_n)-$ X-тің бақыланған мәні болсын. X-тің әрбір жүзеге асырылуы (бақыланған мәні) x-ке мынадай реттелген тізбекті сәйкес қоюға болады:

$$x_{(1)}\le x_{(2)}\le ...\le x_{(n)}, \qquad (1)$$

мұндағы $x_{(1)}$-бақыланған мәндердің ең кішісі, $x_{(2)}$ $-$ $x_1,x_2,\ ...,x_n$-дердің ішіндегі мәнінің кішілігі бойынша екіншісі, т.с.с., $x_{(n)}$-бақыланған мәндердің ең үлкені:

$$x_{(1)} = \min\,(x_1, x_2, ..., x_n), \quad x_{(2)} = \min_i\,(x_i : x_i \ge x_{(1)}),$$
$$x_{(3)} = \min_i\,(x_i : x_i \ge x_{(2)}), ...,$$
$$x_{(n)} = \min_i\,(x_i : x_i \ge x_{(n-1)}) = \max\,(x_1, x_2, ..., x_n)\,.$$

$X_1, X_2, ..., X_n$ таңдамасының әрбір жүзеге асырылуы $x = (x_1, x_2, ..., x_n)$ үшін $X_{(k)}$ арқылы $x_{(k)}$ мәнін $(k = 1,2,..., n)$ қабылдайтын кездейсоқ шаманы белгілейік. Осылайша X таң-дамасы арқылы жаңа, *реттік статистикалар* деп аталатын $X_{(1)}, X_{(2)},..., X_{(n)}$ кездейсоқ шамалар тізбегін анықтаймыз. Бұл жағдайда $X_{(k)}$ – k -ші *реттік статистика* деп аталады. Құ-растыруымыз бойынша

$$X_{(1)} \le X_{(2)} \le ... \le X_{(n)}\,. \qquad (2)$$

(2)-қатар таңдаманың *вариациялық қатары* деп аталады. (2)-тізбектің шеттеріне байланысты симметриялы $X_{(m)}$ және $X_{(n-m+1)}$ $(m = 1,2,...)$ реттік статистикалар таңдаманың m *-ші ең кіші* және m *-ші ең үлкен* мәндері деп аталады. Сонымен вариациялық қатар дегеніміз – ол таңдаманың өсу ретімен орна-ластырылған элементтері. Атай кетелік, $X = (X_1, X_2, ..., X_n)$ таң-дамасының $x = (x_1, x_2, ..., x_n)$ жүзеге асырылуы (бақыланған мәндері) үшін (2) тізбектің жүзеге асырылуы – ол (1) тізбек. Әрине, $X_{(1)}, X_{(2)},..., X_{(n)}$ тәуелсіз кездейсоқ шамалар болуы міндетті емес (бізде $X_1, X_2, ..., X_n$ - тәуелсіз бірдей үлестірілген кездейсоқ шамалар!), себебі олар (2)-теңсіздіктерге бағынады. Реттік статистикалар туралы біз толығырақ үшінші параграфта айтатын боламыз.

Эмпирикалық үлестірім функциясы

Әрбір $x \in R$ үшін $X = (X_1, X_2, ..., X_n)$ таңдамасының x - тен аспайтын элементтерінің санын $\mu_n(x)$ арқылы белгілелік. Онда бұл кездейсоқ шаманы былайша

$$\mu_n(x) = \sum_{i=1}^{n} \mathrm{I}_{\{X_i \le x\}} \qquad (3)$$

жаза аламыз, мұндағы $\mathrm{I}_A(\omega)$ – A оқиғасының индикаторы: $\mathrm{I}_A(\omega) = 1$, егер $\omega \in A$ болса; $\mathrm{I}_A(\omega) = 0$ егер $\omega \notin A$ болса. Әрине, $\mu_n(x)$ үшін

$M\mu_n(x) = \sum\limits_{i=1}^{n} M\mathrm{I}_{\{X_i \le x\}} = \sum\limits_{i=1}^{n} P\{X_i \le x\} = F(x)$, мұндағы $F(x) = F_\xi(x)$. Егер таңдаманың i -ші элементі $(-\infty, x]$ интервалына түссе “табыс”, түспесе – “сәтсіздік” деп есептесек, онда $\mu_n(x)$-ті табысының ықтималдығы $F(x)$ болатын Бер-нуллидің n сынағы нәтижесіндегі табыс саны ретінде қарастыра аламыз. Онда табыстың салыстырмалы жиілігі ретінде анық-талған

$$\hat{F}_n(x) = \frac{\mu_n(x)}{n} = \frac{1}{n}\sum_{i=1}^{n} \mathrm{I}_{\{X_i \le x\}} \qquad (4)$$

функциясы *эмпирикалық үлестірім функциясы* деп аталады да, бақыланатын ξ кездейсоқ шамасының үлестірім функциясы $F(x)$ *теориялық* (немесе *гипотетикалық*) үлестірім функциясы деп аталады. Анықтамасы бойынша $\hat{F}_n(x)$- кездейсоқ шама және $\mu_n(x) \sim Bi(n, F(x))$ болғандықтан

$$P\left\{\hat{F}_n(x) = \frac{k}{n}\right\} = P\{\, \mu_n(x) = k\} = C_n^k (F(x))^k (1 - F(x))^{n-k}, \qquad (5)$$

$$k = 0,1,\ldots, n.$$

Сонымен әрбір $x \in R$ үшін $\hat{F}_n(x)$ эмпирикалық үлестірім функциясы 0, $1/n$, $2/n$,…,(n-1)/n, $n/n = 1$ мәндерін (5) фор-мулаларға сәйкес ықтималдықтармен қабылдайды және X таңдамасының әрбір жүзеге асырылуы үшін бірмәнді анық-талған. $\hat{F}_n(x)$ эмпирикалық үлестірім функциясының кездейсоқ шаманың үлестірім функциясының біз білетін барлық қасиеттерін қанағаттандыратынын байқау қиын емес: *нөл мен бірдің арасында өзгереді; кемімейді; оң жағынан үзіліссіз*. Сонымен бірге ол бөлікті-тұрақты және тек (1)-тізбек нүктелерінде ғана өседі. Егер (1)-қатардың теңсіздіктерінің бәрі қатаң теңсіздіктер ($x_{(1)} < x_{(2)} < \ldots < x_{(n)}$) болса, онда, әрине

$$\hat{F}_n(x) = \begin{cases} 0, \; \textit{егер} \;\; x < x_{(1)} \;\; \textit{болса}, \\ k/n, \; \textit{егер} \;\; x_{(k)} \le x < x_{(k+1)}, \; k = 1,2,\ldots, n-1 \;\; \textit{болса}, \\ 1, \; \textit{егер} \;\; x \ge x_{(n)} \;\; \textit{болса}, \end{cases}$$

яғни бұл жағдайда функцияның барлық секірістерінің шамасы $1/n$ ($\hat{F}_n\left(x_{(k)}\right) - F_n\left(x_{(k)} - 0\right) = 1/n$) және оның графигі $x_{(k)}$ нүкте-леріндегі секірісі $1/n$ болатын монотонды кемімейтін бөлікті-тұрақты функция.

Эмпирикалық үлестірім функциясының математикалық статистикада атқаратын рөлі өте зор, олардың ішіндегі ең маңыздыларының бірі-таңдаманың көлемі өскен сайын оның теориялық функцияға “мейлінше жақындай” беретіндігі.

Гистограмма және салыстырмалы жиіліктер полигоны

Жоғарыда айтқандарымыздан бақыланатын ξ кездейсоқ шамасының үлестірімі белгісіз болған жағдайларда статис-тикалық деректер ($X = \left(X_1, X_2, ..., X_n\right)$ таңдамасы) арқылы оның үлестірімі туралы қорытындылар жасау үшін $\hat{F}_n(x)$ эмпи-рикалық үлестірім функциясын пайдалану ыңғайлы болатынын көреміз.

Статистикалық деректерді көрнекі түрде сипаттайтын басқа да әдістер көп, солардың бірі– *гистограмма* салу әдісі. Бұл жағдайда бақыланатын ξ кездейсоқ шамасының қабылдайтын мәндер облысы ұзындықтары h бірдей болатын интервалдарға бөлінеді де, $G(\xi)$-ден алынған X таңдамасының бақыланатын $x = \left(x_1, x_2, ..., x_n\right)$ мәндері үшін сәйкес интервалға түсетін x_i–координаталарын санайды (деректер топтастырылады). Мәселен, егер $\Delta_j - j$–ші нөмірлі интервал болса, онда $\nu_j = \sum\limits_i I_{\{x_i \in \Delta_j\}}$–осы интервалға түскен x_i–лердің саны. Сосын осындай әрбір Δ_j интервалын табаны етіп алып, биіктігі $\nu_j/(nh)$ болатын тіктөртбұрыштар салады. Осылайша алынған фигура *гистограмма* деп аталады. Әрине, әрбір тіктөрт-бұрыштың ауданы $h\dfrac{\nu_j}{nh} = \dfrac{\nu_j}{n} - \Delta_j$–ші интервалға түскен таңдамалық мәндердің салыстырмалы жиілігі, ал Бернулли схемасы үшін үлкен сандар заңы бойынша $\dfrac{\nu_j}{n} \xrightarrow{P} P\left\{\xi \in \Delta_j\right\}$. Егер ξ–дің үзіліссіз тығыздығы $f(x)$ бар болса, ал интервалдың ұзындығы h аз болса, онда соңғы ықтималдық шамамен $f\left(z_j\right)h$–қа тең, мұндағы $z_j - \Delta_j$–дің орта нүктесі. Сонымен таңдаманың көлемі n жеткілікті үлкен, интервалдардың ұзындығы жеткілікті аз болған жағдайда Δ_j интервалына тұрғызылған тіктөртбұрыштың ауданы $\frac{\nu_j}{nh} \approx f\left(z_j\right)$–тығыздықтың Δ_j интервалының орта нүктесіндегі мәні. Бұдан гистограмманың жоғарғы шекарасын бақыланатын ξ кездейсоқ шамасының жуықталған үлестірім тығыздығы ретінде алуға болатынын көреміз.

Бұл статистикалық деректерді беру әдісін тек үзіліссіз кездейсоқ шамалар үшін қолдану керек және де бұл әдістің көпе-көрінеу көзге ұрып тұрған кемшіліктері де жоқ емес (интервалдарды қалай құру керектігінің анықталмағандығы; деректерді топтастыру кезінде қандай да бір ақпараттарды жоғалтып алу қаупі (гистограмма құру барысында таңда-малардың мәндері емес, олардың интервалдарға түсу жиілігі пайдаланылады) т.с.с.). Сондықтан да гистограмманы статис-тикалық деректерді алдын-ала сараптау кезінде қолданған жөн.

Гистограммалар әдісінде белгісіз үлестірім тығыздығы бөлікті-тұрақты қисықпен жуықталады. Алайда, анализден белгілі екендігіндей, жеткілікті біртегіс функция бөлікті-сызықтық функциялармен әлдеқайда жақсы жуықталады. Бұдан жеткілікті біртегіс тығыздықты жуықтаудың *жиіліктер полигонына* негізделген әлдеқайда дәл әдісін ұсынуға болатынын байқаймыз. *Салыстырмалы жиіліктер полигоны* дегеніміз– ол былайша салынатын қисық: егер таңдама дискретті үлестірімнен алынса, онда $\left(x_i, \nu_i / n\right)$ нүктелерін қосатын сынық қисық, мұндағы $\nu_i - x_i$–дің бақылану саны; егер таңдама үзіліссіз үлестірімнен алынса, онда ол гистограмма құрастырған кездегі интервалдардың орта нүктесіне сәйкес келетін гистограмманың нүктелерін қосатын кесінділерден тұратын сынық қисық.