СӘТБАЕВ УНИВЕРСИТЕТІ

SATBAYEV UNIVERSITY

Статистика және деректерді талдау

Ержан Асел Ануаркызы,Ph.D., «Ақпараттық жүйелер» кафедрасының қауымдастырылған профессоры

Assel.yerzhan@satbayev.university



Таңдама және жиынтық

Статистикалық зерттеудің негізгі тұжырымдамалары: бас жиынтықтан таңдама алу, репрезентативтілік шарты, іріктеу әдістері және нәтижені дұрыс интерпретациялау. Бұл презентация статистикалық ойлаудың іргелі логикасын кезең-кезеңімен ашып береді.

Неге бұл тақырып маңызды?

Шынайы өмірде кез келген зерттеуші, сарапшы немесе шешім қабылдаушы бір маңызды мәселеге тіреледі: **барлығын тексеру мүмкін емес.** Миллиондаған тұтынушының пікірін білгіңіз келсе, барлығымен сөйлесуге ресурс жетпейді. Зауыттан шыққан өнімнің сапасын тексергіңіз келсе, бүкіл партияны бұзып-жарып тексере алмайсыз. Халықтың денсаулық жағдайын бағалағыңыз келсе, жан-жақты медициналық тексерудің шығыны асып кетеді.

Дәл осы жерде **таңдама әдісі** кіреді. Бәрін емес, **бір бөлігін ғана** зерттеп, сол арқылы бүкіл жиынтық туралы негізделген қорытынды жасауға болады. Бірақ бұл қарапайым емес: қандай бөлік алынады, қалай алынады, алынған бөлік бүкілді "өкіле" ала ма — осы сұрақтардың барлығы статистиканың жүрегіне тікелей қатысты.

> Таңдама дұрыс алынбаса, мың адамнан жиналған деректің өзі нәтижені бұрмалауы мүмкін. Ал дұрыс алынса, жүз адам да үлкен жиынтық туралы сенімді ақпарат береді.

Жиынтық ұғымы: бастау нүктесі

Статистикалық жиынтық дегеніміз не?

Сапалық немесе сандық белгілермен сипатталатын нысандардың жиыны — **статистикалық жиынтық** деп аталады. Бұл нысандар адамдар, өнімдер, оқиғалар, өлшемдер — кез келген зерттеу объектісі бола алады.

Маңызды шарт: жиынтықтағы барлық нысандар кем дегенде бір ортақ белгімен сипатталуы тиіс. Мысалы, «Қазақстан азаматтары», «2024 жылы шыққан өнімдер» немесе «университет студенттері» — бұлардың барлығы белгілі бір критерий бойынша бірігетін нысандар жиыны.

Жиынтық белгілерінің түрлері

- **Сандық белгілер:** бой, салмақ, жалақы, температура — сандармен өрнектелетін, есептеуге болатын мәндер
- **Сапалық белгілер:** жыныс, тіл, мамандық, ұлт — санаттарға бөлінетін, бірақ сандық мән беру міндетті емес белгілер
- **Дискретті белгілер:** тек белгілі мәндерді қабылдайтын сандық белгілер (балалар саны, дауыс саны)
- **Үздіксіз белгілер:** кез келген аралықтағы мән қабылдай алатын белгілер (биік, жылдамдық)

Жиынтықты дұрыс анықтамай, статистикалық зерттеу жүргізу мүмкін емес. Зерттеудің нысаны, шекарасы және белгілері алдын ала нақты айқындалуы керек — тек содан кейін таңдама алуға кіріседі.

Бас жиынтық: толық "карта"

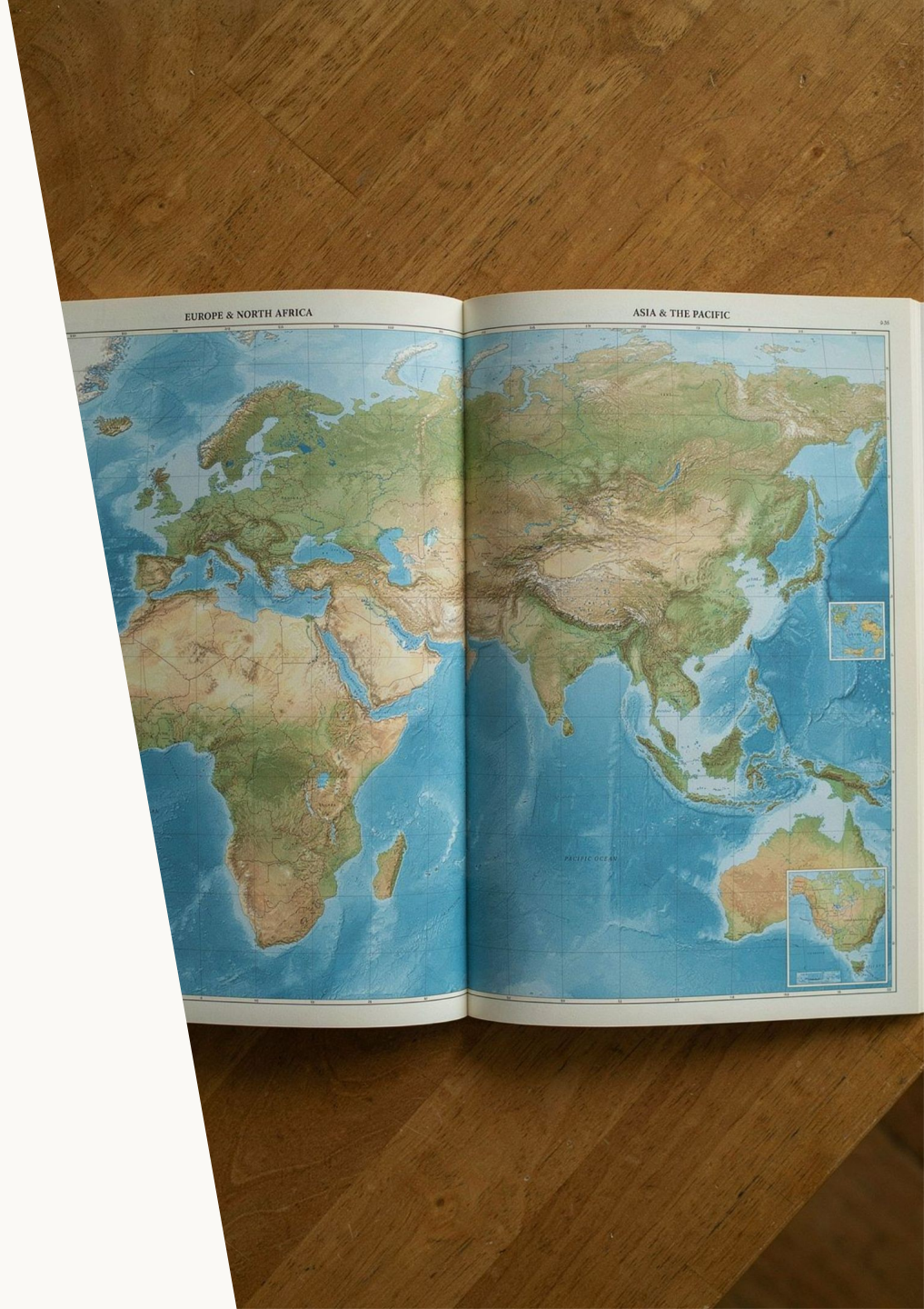
Анықтама

Тексерілуге жататын — ең болмағанда теория жүзінде — **барлық нысандардан тұратын жиынтық** бас статистикалық жиынтық деп аталады. Бұл зерттеушінің «білгісі келетін» барлық нәрсенің жиыны.

«Теория жүзінде» деген сөз маңызды: практикада бас жиынтықты толық тексеру көбінесе мүмкін емес — тым үлкен, тым қымбат немесе тым уақыт қажет.

Мысалдар

- Қазақстандағы барлық 18 жастан асқан сайлаушылар — сайлауалды сауалнаманың бас жиынтығы
- Зауыттан шығарылған барлық шамдар — сапа бақылауының бас жиынтығы
- Белгілі бір университеттің барлық студенттері — оқу нәтижелерін зерттеудің бас жиынтығы
- Бүкіл ел аумағындағы барлық топырақ үлгілері — экологиялық зерттеу жиынтығы



Бас жиынтықты толық зерттеу — **перепись** (санақ) деп аталады. Мемлекеттік халық санағы осының мысалы. Бірақ бұл өте сирек және өте қымбат тәжірибе.

Таңдама жиынтық: дәл сол "үлгі"

Таңдама — бас жиынтықтың шағын көрінісі

Бас жиынтықтан **кездейсоқ түрде алынған нысандар жиыны** таңдама (немесе таңдама жиынтық) деп аталады. Таңдаманың мақсаты — бас жиынтықты "өкілеу", яғни оның негізгі қасиеттерін өзінде сақтау.

Таңдама — бас жиынтықтың толық көшірмесі емес, бірақ оның репрезентативті "проекциясы" болуы тиіс. Дәл осы айырмашылық статистиканың ең маңызды мәселесі болып табылады.

Практикалық маңызы

Таңдаманың нәтижелері бас жиынтыққа "жобаланады": таңдамада есептелген орташа мән, дисперсия немесе пропорция бас жиынтықтың сәйкес параметрін бағалауға қолданылады. Бұл "жобалау" процесі статистикалық қорытындының негізін құрайды.

Негізгі терминдер

Бас жиынтық

Зерттеу объектісінің барлық нысандары. Параметрлері: μ, σ, P (грек әріптері)

Таңдама

Бас жиынтықтан алынған нысандардың жиыны. Статистикалары: $\bar{x}$, s, $\hat{p}$ (латын әріптері)

Бағалаушы

Таңдама деректері бойынша бас жиынтық параметрін есептеу формуласы

Көлем және өлшем: *n*, *N* неге керек?

N

Бас жиынтық көлемі

Бас жиынтықтағы барлық элементтердің жалпы саны. Кейде миллиондаған немесе шексіз болуы мүмкін.

n

Таңдама көлемі

Бас жиынтықтан іріктеліп алынған нысандардың саны. Зерттеуші анықтайды.

n/N

Іріктеу үлесі

Таңдама көлемінің бас жиынтыққа қатынасы. Репрезентативтілік есебінде қолданылады.

Көлем неліктен маңызды?

Таңдама көлемі — зерттеу дәлдігінің негізгі факторларының бірі. **n** неғұрлым үлкен болса, таңдама нәтижелерінің бас жиынтыққа жақындығы соғұрлым жоғары болады. Алайда, бұл тек кездейсоқ ауытқуға қатысты: жүйелік қате бар болса, **n**-ді ешқандай ұлғайту пайда бермейді.

Шексіз жиынтық

Кейбір жиынтықтар теориялық тұрғыдан шексіз болады: мысалы, болашақта өндірілетін барлық өнімдер немесе болуы мүмкін барлық өлшемдер. Мұндай жағдайда **N** анықталмайды, ал іріктеу үлесі нөлге жақындайды — бірақ таңдама логикасы өзгермейді.



1-ТАРАУ

Таңдама әдісі — қорытынды механизмі

Таңдама жиынтықты зерттеу арқылы бас жиынтық туралы қорытынды жасау — таңдама әдісі деп аталады. Бұл статистикалық зерттеудің іргелі логикасы: "бөліктен — бүтін туралы" ой қорыту.

1

Бас жиынтық

Барлық нысандар жиыны. Параметрлері белгісіз немесе тексерілмеген.

2

Таңдама алу

Кездейсоқ іріктеу арқылы n нысан таңдалады. Репрезентативтілік сақталуы міндетті.

3

Статистика есептеу

Таңдама бойынша орташа мән, дисперсия, пропорция есептеледі.

4

Қорытынды жасау

Таңдама статистикасы бас жиынтық параметрін бағалауға қолданылады.

Репрезентативтілік шарты — ең маңызды талап

Репрезентативтілік дегеніміз не?

Таңдама бас жиынтықтың қасиеттерін **"бейне ретінде"** дәл көрсетуі керек. Яғни таңдамадағы нысандардың белгілері бас жиынтықтағы нысандардың белгілеріне жуық болуы тиіс.

Репрезентативтілік — таңдаманың ең маңызды сапалық сипаттамасы. Таңдама репрезентативті болмаса, ол қаншалықты үлкен болса да, нәтиже бас жиынтықты дұрыс бейнелемейді.

Репрезентативтілік қашан бұзылады?

- Іріктеу кездейсоқ емес, белгілі топтарды артық немесе кем қамтиды
- Жауап беру деңгейі төмен (кей топтар сауалнамаға қатыспайды)
- Іріктеу рамасы бас жиынтықты толық қамтымайды
- Өлшеу немесе тіркеу жүйелі қатені тудырады

Классикалық мысал

1936 жылғы АҚШ президенттік сайлауы алдында "Literary Digest" журналы 2,4 миллион адамнан сауалнама жүргізіп, Лэндонның жеңетінін болжады. Нәтиже керісінше болды — Рузвельт жеңді. Себебі: сауалнама тек телефоны мен автомобилі бар байлардан алынды, ал олар бас жиынтықты репрезентативті түрде өкілемеді.

Қателеспейтін іріктеу қандай?

Кездейсоқтық принципі

Бас жиынтықтың **әр элементінің таңдалу мүмкіндігі бірдей** болуы тиіс. Бұл ережені сақтамай жасалған іріктеу — ерікті іріктеу деп аталады және статистикалық қорытынды жасауға жарамсыз.

Кездейсоқтық деген — стихиялы немесе бейберекет дегенді білдірмейді. Бұл — математикалық ықтималдыққа негізделген, арнайы рәсімдермен қамтамасыз етілетін процесс (кездейсоқ сандар генераторы, лотерея тәртібі т.б.).

Тәуелсіздік принципі

Таңдамадағы бір бақылаудың нәтижесі басқа бақылаудың нәтижесіне **ешқандай ықпал етпеуі керек**. Бұл — тәуелсіздік шарты.

Мысалы: бір үйдегі барлық мүшелерді сұрау — тәуелсіздікті бұзады, өйткені олардың пікірлері бір-біріне ықпал еткен болуы мүмкін. Дұрысы — әр үйден бір адам немесе мүлдем басқа іріктеу схемасы қолдану.

Толық қамту принципі

Іріктеу рамасы бас жиынтықтың **барлық мүшелерін** қамтуы тиіс. Егер белгілі топтар іріктеу рамасынан тыс қалса, олардың таңдалу мүмкіндігі нөлге тең болады — бұл жүйелік қате тудырады.

Мысалы: телефон бойынша сауалнама телефоны жоқтарды автоматты түрде шығарып тастайды.

2-ТАРАУ

Вариациялық қатардың тілі

Варианта және вариациялық қатар

Таңдамадағы белгінің жеке өлшенген мәні **варианта** деп аталады. Барлық варианталарды реттік тәртіппен (өсу немесе кему бойынша) орналастырсақ, **вариациялық қатар** аламыз.

Мысалы, 10 студенттің тест нәтижелері: 45, 52, 60, 60, 71, 74, 80, 85, 90, 95 — бұл вариациялық қатар. Мұнда 45 — ең кіші варианта (x_1), 95 — ең үлкен (x_n).

Статистикалық таралу

Варианталар мен олардың **жиіліктері** (яғни, әр мән неше рет кездесетіні) арасындағы байланыс — **таңдама статистикалық таралуы** деп аталады.

Таралу — таңдаманың "портреті". Ол арқылы деректердің қандай мәндер маңында шоғырланғанын, қаншалықты шашыраңқы екенін, симметриялы немесе асимметриялы екенін бірден байқауға болады.

- **Абсолютті жиілік (n_i):** варианта неше рет кездескен
- **Салыстырмалы жиілік (n_i/n):** үлес ретінде
- **Жинақталған жиілік:** x мәніне дейінгі барлық жиіліктер қосындысы

Дискретті vs интервалды логика

Дискретті белгілер

Тек бүтін немесе санаулы мәндер қабылдайды. Мысалы: балалар саны (0, 1, 2, 3…), бракталған өнім саны, дауыс саны.

Таралу: **нүктелік** — әр мән үшін жиілік жеке көрсетіледі. Полигон немесе баған диаграммасы (бөлек бағандар) қолданылады.

Математикалық ерекшелік: мәндер арасындағы аралықтарда мән болмайды, демек жиілік нөл.

Үздіксіз (интервалды) белгілер

Кез келген нақты санды қабылдай алады. Мысалы: бой, салмақ, температура, кіріс деңгейі.

Таралу: **интервалдық** — мәндер аралықтарға (класстарға) топтастырылады. Гистограмма қолданылады — бағандар жанасып тұрады, өйткені аралықтар үздіксіз.

Практикалық мәселе: интервал санын дұрыс таңдау (тым аз — ақпарат жоғалады, тым көп — шу артады). Стёрджес ережесі: $k \approx 1 + 3.322 \cdot \log(n)$

Белгінің табиғаты — дискретті ме, үздіксіз бе — таралу кестесін қалай құру керектігін, қандай график таңдау керектігін және орташа мәнді қалай есептеу керектігін тікелей анықтайды. Осы айырмашылықты ескермеу — жиі кездесетін статистикалық қате.

Мода мен медиана: есептеусіз сипаттама

Мода (Mo)

Ең үлкен жиілікке сәйкес варианта — мода деп аталады. Яғни таңдамада ең жиі кездесетін мән.

Дискретті таралуда мода — жиілігі ең жоғары нақты мән. Интервалды таралуда мода — жиілігі ең жоғары интервалдың ортасы ретінде бағаланады.

Қолдану: "Ең танымал мән қандай?" деген сұраққа жауап береді. Мысалы, ең көп сатылатын аяқ киім өлшемі — мода.

Ескерту: таралу бимодальды (екі шыңы бар) болуы мүмкін — бұл жиынтықтың іштей бөлінгенін білдіруі ықтимал.

Медиана (Me)

Вариациялық қатарды тең екі бөлікке бөлетін варианта — медиана. Жартысы медианадан төмен, жартысы жоғары болады.

Тақ n үшін: медиана — ортасындағы элемент. Жұп n үшін: медиана — екі ортаңғы элементтің орташасы.

Медиананың артықшылығы: ол **шеткі мәндерге (выбростарға) сезімтал емес**. Сондықтан кіріс деңгейін, баға сипаттамасын немесе асимметриялы таралуды сипаттауда медиана орташа мәннен тиімдірек болады.

Орташа мән ($\bar{x}$)

Барлық варианталар қосындысын олардың санына бөлу арқылы алынатын классикалық орталық тенденция өлшемі.

Орташа мән — математикалық тұрғыдан ыңғайлы және дисперсия есептеуінің негізі. Бірақ шеткі мәндерге өте сезімтал: бір "аномалия" бүкіл орташаны ысырып жіберуі мүмкін.

Таңдаудың ережесі: симметриялы таралуда үшеуі де дерлік тең. Асимметриялы таралуда медиана қолданымдылық жағынан басым.

Орта мән, ауытқу және симметрия идеясы

Симметрия мен асимметрия

Егер таралу симметриялы болса, онда: **Mo = Me = $\bar{x}$**. Бұл тепе-теңдік белгісі — таңдама "дұрыс" таралған, шеткі мәндер жоқ.

Оң асимметрия (оңға созылған "құйрық"): **Mo < Me < $\bar{x}$**. Мысалы, кіріс таралуы: аздаған өте бай адамдар орташа мәнді жоғары қарай ысырады.

Сол асимметрия (солға созылған "құйрық"): **$\bar{x}$ < Me < Mo**. Мысалы, емтихан нәтижелері: аздаған өте нашар нәтиже орташаны төмендетеді.

Ауытқу және шашырау өлшемдері

Орталық тенденция (мода, медиана, орташа) таралудың "орталығын" сипаттайды. Бірақ толық сурет үшін **шашырауды** да өлшеу керек — нысандар орталықтан қаншалықты алшақ орналасқан?

- **Дисперсия (s^2):** орташадан ауытқулардың квадраттарының орташасы
- **Стандарттық ауытқу (s):** дисперсияның квадрат түбірі, бастапқы бірліктерде
- **Вариация коэффициенті (CV = s/$\bar{x}$):** салыстырмалы шашырау — бірліктерден тәуелсіз
- **Амплитуда ($R = x_{max} - x_{min}$):** ең қарапайым шашырау өлшемі

Стандарттық ауытқу — ең жиі қолданылатын шашырау өлшемі. Ол орташаның "сенімділігін" де анықтайды: ауытқу аз болса, орташа мән деректерді жақсы сипаттайды.



3-ТАРАУ

Таңдау түрлері: қалай іріктейміз?

Жәй кездейсоқ іріктеу

Бөлшектерге бөлуді талап етпейтін ең қарапайым іріктеу түрі — **жәй кездейсоқ іріктеу**. Бас жиынтықтың әр элементінің таңдалу мүмкіндігі бірдей және тең ықтимал болады.

Қайталанбайтын іріктеу

Бір рет таңдалған нысан қайтадан таңдалуға мүмкіндік берілмейді. Нысан "жойылып" кетеді дегендей. Практикада бұл жиі кездеседі: сауалнамаға бір адам бір рет ғана қатысады.

Математикалық тиімділігі жоғары: бас жиынтық параметрін бағалаудағы дисперсия аз болады. Кішігірім бас жиынтықтарда міндетті түрде қолданылады.

Қайталанатын іріктеу

Таңдалған нысан "қайтарылады" — яғни оның қайтадан таңдалу мүмкіндігі сақталады. Теориялық есептерде жиі қолданылады, өйткені ықтималдықтарды есептеу қарапайым болады.

Шексіз немесе өте үлкен бас жиынтықта қайталанбайтын және қайталанатын іріктеу арасындағы айырмашылық елеусіз болады (**n/N < 0.05** болса, екеуін бірдей деп есептеуге болады).

Бөлшектеп іріктеу: құрылыммен жұмыс

1

Типтік (стратификацияланған) таңдау

Бас жиынтық алдын ала **біртекті қабаттарға (страталарға)** бөлінеді, кейін әр қабаттан жеке іріктеу жүргізіледі. Мысалы: 1-3-сыныптар — бір страта, 4-6-сыныптар — екінші страта, 7-11-сыныптар — үшінші страта. Бұл жиынтықтың ішкі топтарының барлығы таңдамада өкілделуін қамтамасыз етеді. Стратификация дисперсияны азайтып, бағалаудың дәлдігін арттырады — ең тиімді іріктеу тәсілдерінің бірі.

2

Механикалық (жүйелі) таңдау

Бас жиынтықтың нысандары нөмірленіп, **белгілі бір қадаммен (k)** таңдалады. Мысалы, N=1000, n=100 болса, k=10: 1-ден 10-ға дейін кездейсоқ бастапқы нүкте таңдалады (мысалы, 7), кейін 7, 17, 27, 37... деп жалғасады. Жүргізу оңай, бірақ егер бас жиынтықта қадаммен сәйкес келетін циклдік заңдылық болса, бұрмалану мүмкін.

3

Сериялық (кластерлік) таңдау

Алдымен **топтар (кластерлер)** кездейсоқ таңдалады, кейін сол таңдалған топтардың барлық мүшелері немесе ішкі үлгісі зерттеледі. Мысалы: алдымен 20 мектеп таңдалады, кейін сол мектептердің оқушылары зерттеледі. Іске асыру тиімді, бірақ кластер ішіндегі элементтер ұқсас болса (гомогенді), дисперсия артуы мүмкін.

Қатерлі сценарий: репрезентативтілік бұзылса

Жүйелік қате дегеніміз не?

Кездейсоқ емес/араласқан таңдау жағдайында таңдама бас жиынтықтан жүйелі түрде ерекшеленіп кетуі мүмкін. Бұл — **жүйелік қате (систематическая ошибка)**.

Жүйелік қатенің қауіптілігі: оны **n-ді ұлғайту арқылы жою мүмкін емес**. Қанша деректе болса да, қате қайталана береді.

Себептері: іріктеу рамасының толық еместігі, іріктеу процесіндегі субъективтілік, жауаптан бас тарту, өлшеу аппаратының жүйелі ауытқуы.

Нақты мысалдар

Интернет сауалнамасы

Тек интернет пайдаланушыларды қамтиды → ауыл тұрғындары, қарттар кем өкілденеді → "барлық халық" туралы қорытынды бұрмаланады

Ерікті қатысу

Өз еркімен қатысушылар ерекше мотивацияланған топты құрайды → олардың пікірі бас жиынтықтан жүйелі түрде өзгеше болуы мүмкін

Тарихи артта қалу

Ескі тізімнен іріктесе (10 жыл бұрынғы деректер негізінде), бас жиынтық өзгеріп кеткен → іріктеу рамасы ескірген

4-ТАРАУ

Таралуды графикамен көру

Жиілік полигоны

Нүктелік немесе дискретті деректерге арналған графиктің негізгі түрі. Әр варианта (немесе интервалдың ортасы) горизонталь осінде, оның жиілігі вертикаль осінде белгіленеді, нүктелер түзу сызықтармен қосылады.

Полигон таралудың **жалпы формасын** — симметриялы, оң немесе сол қисық, шыңдар саны — дереу байқатады. Бірнеше таңдаманы бір графикте салыстыруға өте ыңғайлы.

Полигон астындағы аудан = 1 (барлық жиіліктер қосындысы), бұл ықтималдықтарды геометриялық түрде интерпретациялауға мүмкіндік береді.

Гистограмма

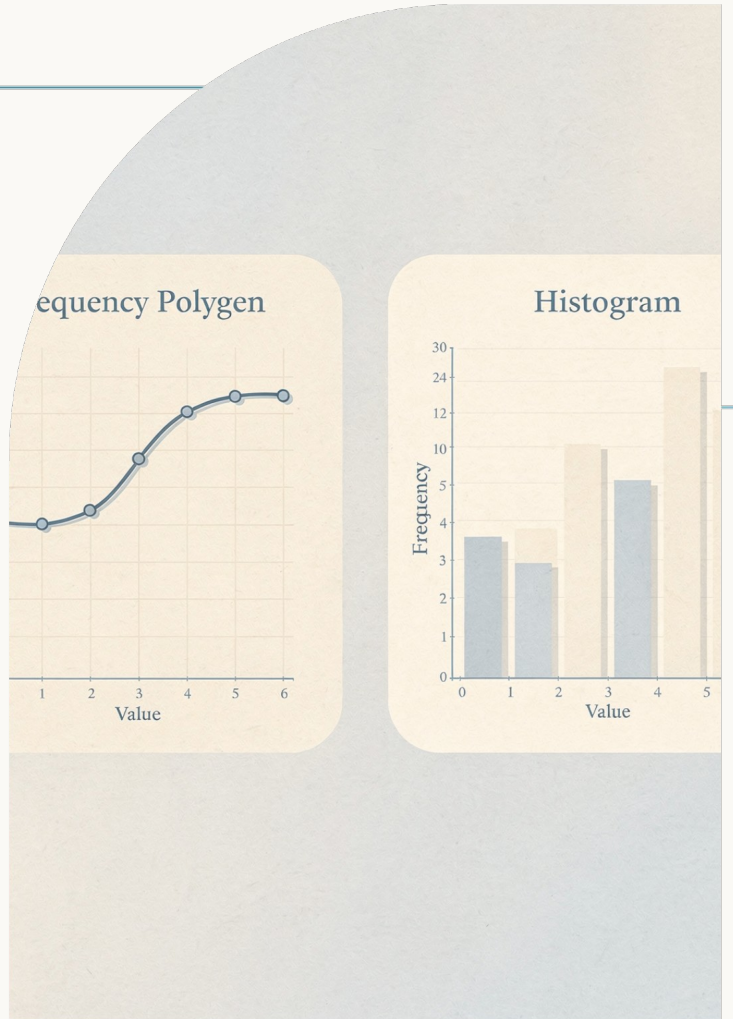
Үздіксіз (интервалды) деректерге арналған. Горизонталь осіне интервалдар (класстар), вертикаль осіне жиілік (немесе жиілік тығыздығы) белгіленеді. Бағандар жанасып тұрады — аралықтар үздіксіз.

Жиілік тығыздығы: $\mathbf{f = n_i / (n \cdot h)}$, мұнда h — интервал ені. Бұл интервал ені бірдей болмаса да салыстыруға мүмкіндік береді.

Гистограмма астындағы аудан да 1-ге тең. Баған ені кішірейген сайын гистограмма үздіксіз таралу функциясына жақындайды.

Жиілік полигоны

Дискретті деректер — қосылған нүктелер арқылы.



Гистограмма

Үздіксіз деректер — жанасатын бағандар.

Эмпирикалық функция: "нақты бақылаудан" ықтималға

Эмпирикалық таралу функциясы F*(x)

Бас жиынтықта теориялық таралу функциясы **F(x)** болса, таңдама бойынша оны **эмпирикалық функция F*(x)** арқылы бағалаймыз.

Анықтамасы: **F*(x) = m(x) / n**, мұнда **m(x)** — мәні x-тен кіші байқаулар саны, **n** — таңдама көлемі.

Бұл функция: монотонды өсетін, $x \to -\infty$ болса $F^*(x) \to 0$, $x \to +\infty$ болса $F^*(x) \to 1$ болатын баспалдақты (ступенчатая) функция.

Эмпирикалық функцияны оқу

$F^*(x_0) = 0.7$ дегенді қалай оқуға болады? Бұл — таңдамадағы бақылаулардың 70%-ы x_0-ден кіші екенін білдіреді.

Мысалы, студенттердің тест нәтижелері бойынша $F^*(80) = 0.65$ болса, студенттердің 65%-ы 80 баллдан төмен нәтиже алды дегенді білдіреді.

Маңыздылығы: n өскен сайын F*(x) теориялық F(x)-ке жақындайды (**Гливенко-Кантелли теоремасы**). Бұл — эмпирикалық функцияның статистикалық негіздемесі.

- Таралудың кез келген квантилін анықтауға болады
- Медиана: $F^*(Me) = 0.5$ болатын мән
- Квартилдер: $F^*(Q_1) = 0.25$, $F^*(Q_3) = 0.75$

Дискретті статистикалық таралу қатары

Таңдамадағы әр мәннің жиілігі — дискретті таралу кестесінің негізгі мазмұны. Ол таңдамадағы деректердің n ішінде қалай үлестірілгенін нақты, санды түрде көрсетеді.

Варианта (x_i)	**Абс. жиілік (n_i)**	**Сал. жиілік (n_i/n)**	**Жинақталған жиілік**
x_1	n_1	n_1/n	n_1/n
x_2	n_2	n_2/n	$(n_1+n_2)/n$
x_3	n_3	n_3/n	$(n_1+n_2+n_3)/n$
Барлығы	**n**	**1**	—

Кесте мысалы: дүкендегі сатылым

Сатылым (дана)	**Күндер саны**	**Үлес (%)**
0	5	10%
1	12	24%
2	18	36%
3	10	20%
4	5	10%
Барлығы	**50**	**100%**

Мода = 2 (ең жоғары жиілік), Медиана = 2, орташа мән ≈ 1.96. Таралу шамамен симметриялы.

Репрезентативтілік өлшемі: "үлгі қателігі" қайдан шығады?

Таңдама әрқашан кездейсоқ алынатындықтан, бірдей зерттеуді бірнеше рет жүргізсек те, нәтижелер аздап өзгеше шығады. Бұл — **таңдаманың тұрақсыздығы** немесе іріктеу қателігі.

Қате туындайтын екі негізгі себеп

Таңдаманың шектеулі көлемі

n шексіз болмағандықтан, кез келген таңдама бас жиынтықтан аздап ауытқиды. Бұл — **кездейсоқ (стохастикалық) қате**. Оны азайтудың жалғыз жолы — n-ді ұлғайту немесе стратификацияны пайдалану.

"Адам факторы" — өлшеу қателігі

Деректерді жіктеудегі қате, анкетадағы түсініксіз сұрақтар, тіркеу процесіндегі абайсыздық — бұлардың барлығы қосымша, кейде жүйелік сипатқа ие болатын қате тудырады.

Практикалық қорытынды

"Ең жақсы үлгі" деген жоқ — барлық таңдамалар белгілі бір дәрежеде қатемен жұмыс жасайды. Мақсат — **қателікті түсініп, оны бақылауға алу**. Статистикалық зерттеудің кәсіби деңгейі — нәтижені нүктелік мәнмен ғана емес, оның белгісіздік шегімен бірге ұсыну арқылы өлшенеді.

Ең маңызды ереже: **кездейсоқ қателікті n арқылы азайтуға болады, бірақ жүйелік қателікті тек дизайн арқылы жоюға болады.** Бұл айырмашылықты тану — статистикалық ойлаудың іргелі дағдысы.

Таңдама көлемі неге әсер етеді: сигнал vs шу

Үлкен n — тар ауытқу аймағы

Таңдама көлемі артқан сайын кездейсоқ ауытқулар азаяды: таңдама орташасы бас жиынтықтың нақты орташасына тұрақтана береді. Бұл математикалық тұрғыдан **үлкен сандар заңымен** (заң больших чисел) түсіндіріледі.

Формула: орташаның стандарттық қателігі **SE = σ / √n**. n 4 есе артса, қате 2 есе азаяды. n 100 есе артса, қате 10 есе азаяды. Яғни жетістіктер азайып бара жатады — үлкен n-ге жетудің маржинальды пайдасы азаяды.

Тәуекелді есепке алу

Аз таңдамамен де "дұрыс" шешім мүмкін — бірақ бұл тәуекелдің өсуімен бірге жүреді. Кішігірім n болса, нәтиже кездейсоқ ауытқудан айтарлықтай бұрмалануы ықтимал.

Практикалық ережелер

- Зерттеу алдында **"қаншалықты дәл болу керек?"** деген талап қойылады. Бұл — дәлдік шегі (margin of error) мен сенімділік деңгейін (confidence level) анықтау.
- Медициналық зерттеулер, сайлауалды сауалнамалар, сапа бақылауы — әрқайсысының өзіндік **ең аз n** талабы бар.
- Жалпы ереже: **n ≥ 30** болса, Орталық Шектілік Теорема жұмыс істейді — орташаның таралуы нормальға жақындайды.
- Стратификацияланған іріктеуде бірдей n қарапайым іріктеуден тиімдірек — қабаттар ішіндегі вариация азаяды.

Ықтималдық тілі: сенім аралығы түсінігі

Нүктелік бағадан диапазонға

Нүктелік баға (мысалы, орташа мән $\bar{x}$) — бір ғана сан. Бірақ бұл сан шынайы жиынтық параметрімен (**μ**) дәл тең болуы ықтимал емес. Дұрысы — **диапазон ойлау**: шынайы жиынтық мәні сол аралықта болуы мүмкін.

Сенім аралығы (доверительный интервал): **$[\bar{x} - E, \bar{x} + E]$**, мұнда **E** — қателік шегі (margin of error). Бұл аралықтың шынайы параметрді қамтуының белгілі бір ықтималдығы (сенімділік деңгейі) бар.

Сенімділік деңгейін оқу

"95% сенімділік деңгейі" дегенді дұрыс оқу: егер бірдей әдіспен 100 рет таңдама алып, 100 рет сенім аралығы есептелсе, оның 95-і шынайы параметрді қамтиды деп күтіледі.

Аралықты кеңейтетін факторлар

n кіші болса немесе деректегі өзгергіштік (σ) жоғары болса, аралық кеңейеді — белгісіздік артады

Аралықты тарылтатын факторлар

n үлкен болса немесе сенімділік деңгейін төмендетсе (99% → 90%), аралық тарылады

Іріктеу қатесі мен жүйелік қате: бір-біріне ұқсамайды

Іріктеу (кездейсоқ) қатесі

Кездейсоқтықтан пайда болады: үлгі кездейсоқ ауытқиды. Бір бағытқа ысырмайды — кей жолы асып, кей жолы кем шығады. Таңдаманы қайталаса, орташаланып "жойылады". **n ұлғайту арқылы азайтуға болады.**



Жүйелік қате (систематическая ошибка)

"Дұрыс емес жерден алу" нәтижені бір бағытқа ысырып жібереді — репрезентативтілік бұзылған. Таңдаманы қанша рет қайталаса да, қате қайталанады. **n ұлғайту жүйелік қатені жоймайды.** Тек дизайн арқылы жою мүмкін.

Нақты мысал

Бір аудандағы адамдардан ғана сауалнама алынып, барлық қалаға "орташа" баға берілсе — бұл жүйелік қате. Тіпті сол аудандан 10 000 адамды сұрасаңыз да, барлық қала туралы қорытынды жасауға болмайды. Себебі: ол аудан — бас жиынтықты репрезентативті түрде өкілемейді.

Екі қатені бірге ескеру

Кәсіби статистикалық зерттеу екі қате түрін де бір мезетте ескереді. Ең тиімді стратегия — **дизайн кезеңінде жүйелік қатені жою** (іріктеу рамасын толық жасау, стратификацияны пайдалану), **n арқылы кездейсоқ қатені минимизациялау.**

Таңдама дизайны: қалай таңдасаң, солай қорытындылайсың



Қарапайым кездейсоқ іріктеу

Әрбір элементтің түсу мүмкіндігі бірдей. Іске асыру қарапайым, бірақ жиынтықтың ішкі құрылымын ескермейді. Кішігірім, біртекті жиынтықтарда тиімді.



Стратификация

Жиынтықты ішкі топтарға (қабаттарға) бөліп, әр қабаттан үлеске сай таңдайсың. Дисперсия азаяды, бағалау тұрақтанады. Жиынтықта анық ішкі топтар болса — ең тиімді тәсіл.



Кластерлік іріктеу

Алдымен топтарды таңдап, кейін сол топ ішінде элементтерді аласың. Үлкен географиялық зерттеулерде тиімді. Кластер ішіндегі гомогенділік жоғары болса, нәтиже сапасы төмендейді.

Нәтиже сапасын жоспар дизайны шешеді, тек n емес. Дұрыс таңдалған іріктеу тәсілі бірдей n-мен анағұрлым дәл нәтиже береді.

Дерек сапасы: өлшеу қателігі таңдамадан да қауіпті болуы мүмкін

Репрезентативтілік жеткіліксіз

Таңдама идеал дизайнмен алынса да, **өлшеу әдісі қате болса**, жиынтық туралы түсінік бұрмаланады. Бұл — зерттеушілер жиі елемейтін қауіп.

Мысалдар:

- Салмақты дәлдігі төмен таразымен өлшесе, орташа мәннің вариациясы өседі — шынайы вариация мен өлшеу қателігі шатасады
- Анкетадағы сұрақ екіұшты болса, респонденттер оны әртүрлі түсінеді — деректе жасанды вариация пайда болады
- Интервьюер пікірі респондентке ықпал етсе — өлшеу субъективтілігі (интервьюер эффектісі)

Категорияны дұрыс жіктемеу

Тағы бір қауіп: дискретті/интервалды логика шатасуы. Егер үздіксіз белгіні (мысалы, кіріс) шамадан тыс аз санды категорияға бөлсе, ақпараттың маңызды бөлігі жоғалады.

Практикалық нұсқаулар:

- Өлшеу аспаптарын алдын ала калибрлеу
- Анкета сұрақтарын алдын ала сынақтан өткізу (пилоттық зерттеу)
- Интервьюерлерді стандартты нұсқаулармен оқыту
- Қайталанатын өлшемдер арқылы сенімділікті тексеру (test-retest reliability)

Таңдамадағы "жасырын" қате: іріктеу негізінің бұзылуы

Іріктеу рамасы (sampling frame) дегеніміз не?

Іріктеу рамасы — бас жиынтықтың барлық мүшелерінің тізімі немесе картасы, яғни іріктеу жүргізілетін практикалық нысандар тізімі. Теориялық бас жиынтық пен іріктеу рамасы **сәйкес келмейде** — бұл "жасырын" репрезентативтілік проблемасы.

Іріктеу рамасы қате болса (кей топтар мүлде кірмесе), репрезентативтілік тек "көрінбей" бұзылады. Деректе ешқандай сигнал жоқ — бірақ қорытынды жүйелі түрде бұрмаланған.

Мысалдар мен салдарлары

- **Телефон тізімі:** телефоны жоқ немесе нөмірін жариялаудан бас тартқандар іріктеу рамасынан тыс қалады → байлар/кедейлер теңсіз өкілденеді
- **Тұрғын үй тізімі:** үйсіздер, жатақхана тұрғындары, жаңа қоныстанушылар кірмеуі мүмкін
- **Сауалнама тек қала орталығында жүргізілсе:** ауыл тұрғындарының ықтималдық үлесі нөлге түседі → "барлық халық" туралы қорытынды жасауға болмайды

Шешім: іріктеу рамасын мүмкіндігінше бас жиынтыққа жақындату, "шығып қалған" топтарды анықтап, есепке алу немесе қосымша іріктеу жүргізу (oversampling).

Стратификация және кластер: бірдей көлем емес, бірдей ақпарат



Стратификация тиімділігі

Стратификациялы іріктеуде **әр ішкі топтан (стратадан) жеке үлгі** алынады. Мұның негізгі артықшылығы: страта ішіндегі нысандар бір-біріне ұқсас болса, жалпы дисперсия күрт азаяды.

Математикалық дәлел: жалпы дисперсия = (**страта ішіндегі дисперсия** орташасы) + (**стратааралық дисперсия**). Стратификация стратааралық дисперсияны "жойып", тек страта ішіндегі дисперсияны ғана қалдырады. Бағалау тұрақтанады.



Кластерлік іріктеудің логикасы

Кластерлік іріктеуде алдымен **топтар (кластерлер)** кездейсоқ таңдалады, кейін сол топ ішіндегі барлық немесе таңдамалы элементтер зерттеледі.

Мысалы: мектептерді кластер қылып, алдымен 20 мектеп таңдалады, кейін тек сол мектептердің оқушылары сұрастырылады. Бұл тәсіл **логистикалық жағынан тиімді**, бірақ кластер ішіндегі элементтер ұқсас (гомогенді) болса, статистикалық тиімділік төмендейді.



Дизайн шешеді

Екі тәсілдің статистикалық салдары өзгеше. Стратификация бағалаудың дисперсиясын азайтады — бірдей n-мен жоғары дәлдік береді. Кластер дисперсияны арттырады — логистика ыңғайлылығы үшін статистикалық тиімділіктен бас тартылады.

Нәтиже сапасын жоспар дизайны шешеді — тек n ғана емес. Мақсат, бюджет және жиынтықтың ішкі құрылымына байланысты оңтайлы тәсіл таңдалуы тиіс.

Бейтарап емес бағалаушының "тізбегі": таңдау → статистика → ықтималдық



Іріктеу Статистика Ықтималдық

Екі сатылы логика

Статистикалық қорытынды екі кезеңнен тұрады:

1. **Таңдама қалай алынды?** — іріктеу дизайны (кездейсоқтық, тәуелсіздік, репрезентативтілік)
2. **Сол таңдамадан қандай статистика есептелді?** — орташа ма, медиана ма, пропорция ма, және ол бас жиынтық параметрін бейтарап бағалай ма?

Екі кезең де дұрыс орындалғанда ғана қорытынды статистикалық тұрғыдан негізделген болады.

Қате көздері мен бағалаушының маңызы

Статистиканы дұрыс таңдамау да нәтижені бұрмалайды. Мысалы, асимметриялы таралуда орташа мәнді қолдану — бас жиынтық "орталығын" дұрыс бейнелемейді. Медиана немесе трімдік орташа (trimmed mean) тиімдірек болуы мүмкін.

Бейтарап бағалаушы (несмещённая оценка): таңдама статистикасының математикалық күтімі бас жиынтық параметріне тең болуы тиіс. Мысалы, дисперсияны есептегенде **(n-1)** қолдану — бейтараптықты қамтамасыз ету үшін (Бессель түзетуі).