СӘТБАЕВ УНИВЕРСИТЕТІ

SATBAYEV UNIVERSITY

Статистика және деректерді талдау

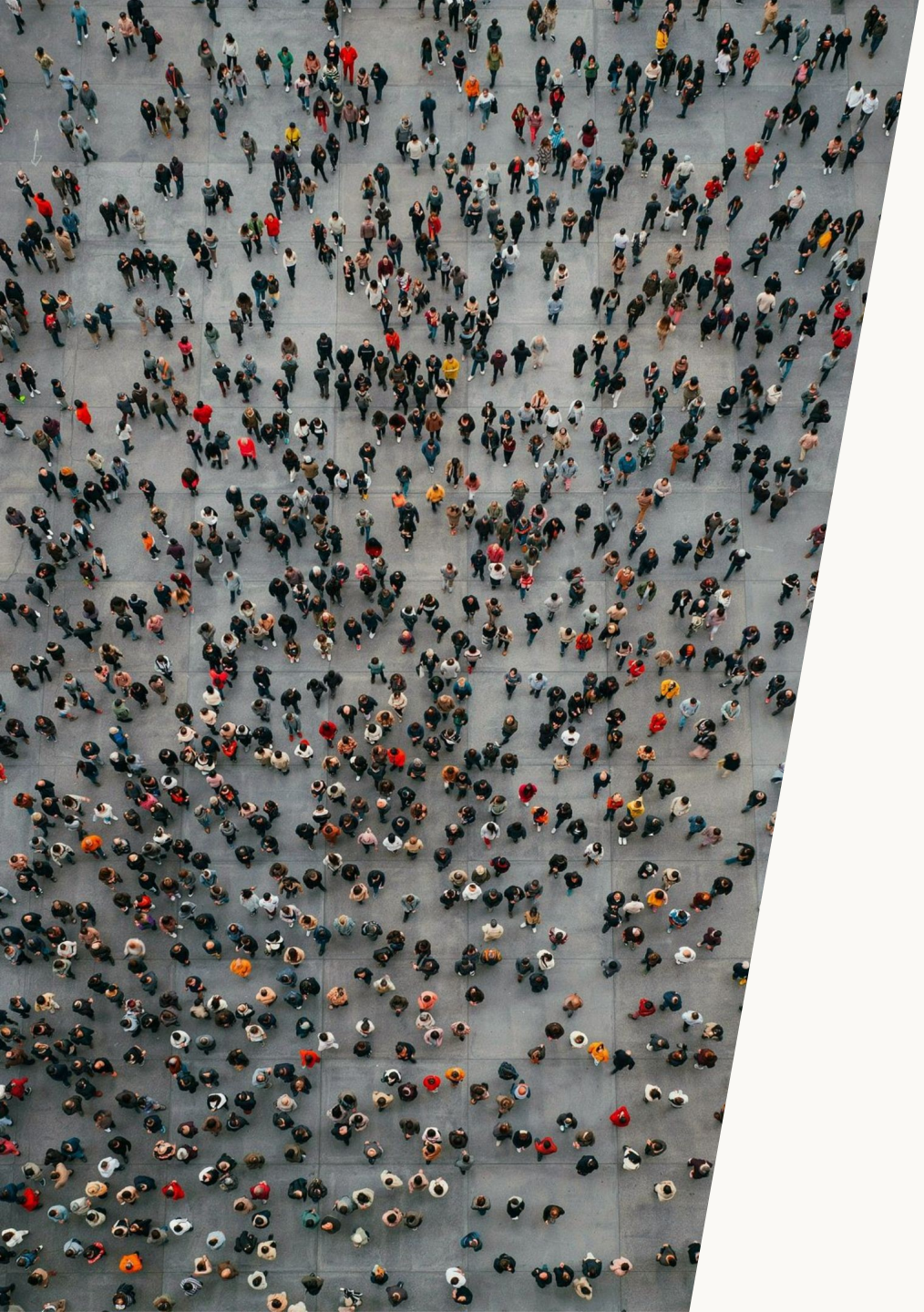
Ержан Асел Ануаркызы,Ph.D., «Ақпараттық жүйелер» кафедрасының қауымдастырылған профессоры

Assel.yerzhan@satbayev.university



Статистикалық үлестірімдер

Деректер жинақталғанда ең алдымен сұрақ туындайды: бұл сандардың артында қандай заңдылық жатыр? Статистикалық үлестірім — деректің "пішінін", шашырауын, орталығын және асимметриясын бір мезетте ашып беретін аналитикалық құрал. Осы дәріс барысында біз жиіліктен бастап ықтималдық моделіне дейінгі жолды бірге жүріп өтеміз.



Неге "үлестірім" қажет?

Статистикалық талдауда жиі жіберілетін қате — барлық деректі тек **орташа мәнмен** сипаттауға тырысу. Орташа мән — пайдалы, бірақ жалғыз болған кезде адастыратын көрсеткіш.

Орташа мән кейде өтірік айтады

Егер бір топта мәндердің 90%-ы 10–20 аралығында болса, ал 10%-ы 1000-нан асса, орташа мән 100-ден жоғары болуы мүмкін — шындықты бұрмалап.

Үлестірім толық суретті береді

Деректердің қалай таралғанын, қай жерде шоғырланғанын, қай жерде сирек кездесетінін тек үлестірім арқылы анық байқауға болады.

Шешім қабылдауда маңызы зор

Медицина, экономика, білім беру саласында дұрыс шешім қабылдау үшін орташаға ғана сүйену жеткіліксіз — үлестірімнің толық пішіні талданады.

Орташа мәнді жалғыз пайдалансаңыз — деректің жартысын ғана оқығансыз.

Үлестірім — деректің пішіні

Үлестірім дегеніміз не?

Таңдамадағы мәндер бірдей жиілікте кездеспейді. Кейбір мәндер жиі қайталанады — олар **"ортаны"** немесе **"шыңды"** құрайды. Кейбір мәндер сирек кездеседі — олар **"құйрықты"** түзеді. Осы бөліністі картаға түсіру — үлестірімді салу деп аталады.

Үлестірім — таңдаманың "портреті". Ол қай мәндердің жиі, қайсысының сирек кездесетінін, мәндердің қай жақта шоғырланатынын және ауытқулардың қай жерде орналасатынын бір мезетте суреттейді.

Үлестірімнің негізгі сипаттамалары

- **Орталық тенденция:** мәндер қай нүктеге жинақталады (мода, медиана, орташа)
- **Шашырау:** мәндер орталықтан қаншалықты алыс орналасқан (дисперсия, СА)
- **Симметрия:** үлестірім симметриялы ма, әлде бір жаққа қисайған ба
- **Шың:** үлестірімде бір ғана шың бар ма, жоқ бірнешеу ме
- **Құйрық:** сирек мәндер қандай мөлшерде бар және олар қай жақта

Бұл бес сипаттаманы бір мезетте талдай білген аналитик деректің толық "тарихын" оқи алады.

Дерек түрі: сандық пен категория

Статистикалық үлестірімді дұрыс түсіну үшін ең алдымен деректің **қандай түрде** екенін ажырата білу керек. Дерек түрі қолданылатын үлестірім моделін және талдау әдісін анықтайды.

Дискретті (санау) деректер

Тек бүтін сандармен өрнектелетін мәндер. Мысалы: отбасы мүшелерінің саны, сатылған тауар саны, апта күндері. Бұл деректер **санауға** негізделген және үздіксіз шкалада орналаспайды. Дискретті үлестірімдер: биномдық, Пуассон үлестірімдері.

Үздіксіз (өлшеу) деректер

Кез келген нақты сан мәнін қабылдайтын деректер. Мысалы: бойы, салмағы, температура, кіріс. Бұл деректер **өлшеуге** негізделген және екі мән арасында шексіз мәндер болуы мүмкін. Үздіксіз үлестірімдер: қалыпты, экспоненциалдық, бета үлестірімдері.

Категориялық деректер

Сандық мәні жоқ, топтарға бөлінетін деректер. Мысалы: жыныс, этностық топ, мамандық. Сандық үлестірім қолданылмайды; оның орнына **жиілік кестесі** мен **бағандық диаграмма** пайдаланылады.



1-ТАРАУ

Жиілік әлемі (Frequency)

Үлестірімнің алғашқы көрінісі

Деректерді алғаш қолға алғанда бірінші қарайтын нәрсе — **жиіліктер**. Жиілік дегеніміз: берілген мән немесе мәндер аралығы таңдамада неше рет кездесетіні. Жиіліктерді жинақтау үлестірімді "салудың" бірінші қадамы болып табылады.

01

Деректерді жинақтау

Барлық бақылау нәтижелері бір жерге топтастырылады. Олардың ауқымы (минимум–максимум) анықталады.

02

Аралықтарға бөлу

Деректердің ауқымы бірдей немесе ыңғайлы аралықтарға (интервалдарға) бөлінеді. Аралық ені тым кең болса — маңызды ақпарат жоғалады; тым тар болса — шу (noise) пайда болады.

03

Жиілікті есептеу

Әр аралыққа қанша бақылау мәні түсетіні санақталады. Абсолюттік жиілік (n) және салыстырмалы жиілік (%) есептеледі.

04

Жиілік кестесін түзу

Аралық, жиілік, жинақталған жиілік (CDF), салыстырмалы жиілік бағандары бар кесте жасалады.



Гистограмма: суреттегі бірінші "жауап"

Гистограмма — жиіліктер кестесін **визуалды формаға** аударудың ең тиімді тәсілі. Ол деректердің қалай таралғанын, қай интервалда мәндер жиі шоғырланатынын, ал қай жерінде сирек кездесетінін бір қарағанда байқауға мүмкіндік береді.

Гистограмманың ерекшеліктері

- X осінде — мәндердің интервалдары (аралықтары)
- Y осінде — әр интервалға түскен бақылау саны (жиілік)
- Бағандар бір-бірімен жапсарласады — үздіксіздікті білдіреді
- Биіктік жиілікке пропорционал
- Ауданы — салыстырмалы жиілікке тең

Интервал енін таңдау ережелері

Интервал саны деректер көлеміне байланысты. Жиі қолданылатын ереже:

- **Стёрджестің ережесі:** $k = 1 + 3.322 \times \log(n)$
- **Квадрат тамыр ережесі:** $k = \sqrt{n}$
- **Скотт ережесі:** $h = 3.49 \times \sigma \times n^{(-1/3)}$

Практикада 5–20 интервал жиі пайдаланылады. Мақсат: үлестірімнің пішінін айқын көрсету.

Полигон, баған және қисық: қандай жағдайға қайсысы?

Жиілік үлестірімін бейнелеудің бірнеше тәсілі бар. Олардың арасында дұрыс таңдау — деректің **табиғатына** байланысты.

Бағандық диаграмма

Дискретті немесе категориялық деректер үшін қолайлы. Бағандар арасында бос орын болады — дискреттілікті суреттейді. Жиі абсолюттік жиіліктерді көрсету үшін пайдаланылады.

Жиілік полигоны

Дискретті деректерде де, үздіксізде де қолданылады. Интервалдардың орта нүктелерін қосатын сызықтан тұрады. Бірнеше үлестірімді бір графикте салыстыруға өте ыңғайлы.

Тегістелген қисық (KDE)

Үздіксіз деректердің "шынайы" пішінін жылтыратып береді. Ядролық тығыздық бағасы (KDE) нүктелер бойынша жылжымалы орташамен есептеледі. Ең табиғи және эстетикалық нұсқа — талдаудың соңғы кезеңінде пайдаланылады.

Гистограмма

Үздіксіз деректер үшін ең табиғи таңдау. Аралықтар жапсарласады. Пішінді, симметрияны, шыңды бір қарағанда байқауға болады.

Пайда болатын сұрақтар

Гистограмманы немесе жиілік үлестірімін алғаш рет қарағанда аналитик бірден бірнеше **негізгі сұрақ** қояды. Бұл сұрақтарға жауап беру — талдаудың бастапқы кезеңінің өзегі.

Қай интервал басым?

Үлестірімнің шыңы (мода) — ең жиі кездесетін аралық. Ол бір ме, жоқ бірнешеу ме? Бір шың (unimodal) болса — процесс біртекті. Бірнеше шың болса — деректер бірнеше топтан немесе режимнен жинақталған.

Қалдықтар қай жақта?

Үлестірімнің "ұзын жағы" — оң немесе сол жаққа созылған бөлігі. Бұл асимметрия (skewness) туралы алғашқы белгі. Оң жаққа созылса — үлкен мәндер бар, солға болса — кіші мәндер.

Қаншалықты шашыраңқы?

Мәндер орталыққа жинақталған ба, жоқ кең таралған ба? Тар үлестірім — аз өзгермелілік. Кең үлестірім — жоғары өзгермелілік. Бұл — дисперсияның алғашқы визуалдық бағасы.

Құйрық (tail) — сирек оқиғалардың ізі

Үлестірімнің "құйрығы" — жиілік аз болатын шеткі аймақ. Бұл аймақ талдауда ерекше назар аударуды талап етеді, өйткені ол **сирек бірақ маңызды** оқиғалар туралы ақпарат береді.

Оң (жоғары) құйрық

Мәндердің оң жағында созылған аймақ. Табыс үлестірімінде — өте жоғары кірісті топтар. Биологияда — өте үлкен мөлшерлер. Бұл "артық" мәндер орташаны жоғары тартады.

Сол (төменгі) құйрық

Мәндердің сол жағына созылған аймақ. Кедейлік деңгейінде — өте аз кіріс алатын топтар. Өндірісте — төмен сапалы партиялар. Бұл мәндер орташаны төмен тартады.

Неге маңызды?

- Қаржыда "fat tail" — экстремалды шығын тәуекелі
- Медицинада — сирек жанама әсерлерді анықтау
- Инженерияда — материал шекті жүктемесін бағалау
- Кибербезопасқа — атаулы шабуылдар (outlier events)

Үлестірімді ерте салу арқылы осы "сирек бірақ қымбат" оқиғаларды алдын ала байқап, шаралар қабылдауға болады.

2-ТАРАУ

Орташа емес, таралу (Spread)

Үлестірімнің екінші өлшемі

Орталық тенденцияны (орташа, медиана, мода) білу жеткіліксіз. Деректердің **қаншалықты шашыраңқы** екенін білмесе, талдаушы маңызды ақпараттан айырылады. Таралу (spread немесе variability) — үлестірімнің екінші негізгі өлшемі.

Неге таралу маңызды?

Екі топтың орташа мәні бірдей болса да, шашырауы өте ерекше болуы мүмкін. Бірінші топ — тар үлестірімді (біркелкі), екіншісі — кең үлестірімді (ала-құла) болуы ықтимал.

Қандай өлшемдер бар?

Диапазон, дисперсия, стандартты ауытқу, квартильаралық аралық (IQR), пайыздық аралық (percentile range) — таралуды өлшеудің негізгі тәсілдері.

Практикадағы мысал

Екі сынып оқушысының орташа бағасы бірдей болса да, бірінші сыныпта барлығы 70–80 балл алса, екіншісінде 30-дан 100-ге дейін — нәтижелер мүлдем басқа сипатта.

Диапазон: бір қарағанда тым дөрекі

Диапазон дегеніміз не?

Диапазон — таңдамадағы ең үлкен мән мен ең кіші мән арасындағы айырым:

R = X_max − X_min

Есептеуі өте оңай — тек екі мән керек. Сондықтан алғашқы тез бағалау үшін жиі пайдаланылады.

Артықшылықтары

- Жылдам есептеледі
- Ауқымды бірден анықтайды
- Деректің шекараларын белгілейді

Негізгі кемшіліктері

Диапазон тек **екі мәнге** — минимум мен максимумға — сүйенеді. Қалған барлық бақылаулар ескерілмейді. Бұл оны **ауытқуларға өте сезімтал** етеді.

- Бір ғана ауытқу мән диапазонды үлкен өзгертуі мүмкін
- Мәндердің ортадағы таралуы туралы ештеңе айтпайды
- Үлестірімнің **пішіні** жайлы мүлде "үнсіз"

Қорытынды

Диапазон — алғашқы шолу үшін ыңғайлы, бірақ жалғыз шашырау өлшемі ретінде ешқашан пайдаланылмауы тиіс. Одан кейін дисперсия мен стандартты ауытқуға өту керек.

Дисперсия мен стандартты ауытқу

Шашырауды дәл өлшеудің ең тиімді математикалық тәсілі — әр мәннің орташадан қаншалықты алыс тұрғанын есептеу. Осы идеядан **дисперсия** мен **стандартты ауытқу** туындайды.

1

Ауытқуды есептеу

Әр бақылаудан орташаны алып тастаймыз: **$(x_i - \bar{x})$**. Теріс мәндер болады — сондықтан тікелей қосу мүмкін емес.

2

Квадратқа шығару

Теріс мәндерді жою үшін квадратқа шығарамыз: **$(x_i - \bar{x})^2$**. Енді барлық мәндер оң болды.

3

Орташалау

Квадраттардың орташасы — **дисперсия (σ^2 немесе s^2)**. Генералды жиын үшін n-ге, таңдама үшін (n−1)-ге бөлеміз.

4

Квадрат тамыр

Дисперсияның квадрат тамыры — **стандартты ауытқу (σ немесе s)**. Деректермен бірдей өлшем бірлігінде болады — интерпретациялауға ыңғайлы.

Стандартты ауытқу — шашырауды "орташа қашықтық" ретінде көрсетеді. Ол кіші болса — деректер тығыз орналасқан, үлкен болса — кең таралған.

Квартили және квартильаралық аралық

Дисперсия мен стандартты ауытқу ауытқуларға өте сезімтал. Оларға балама ретінде **квартиль** негізіндегі өлшемдер пайдаланылады — олар экстремалды мәндерге азырақ "бас ию" береді.

Квартильдер (Q1, Q2, Q3)

Квартильдер деректерді 4 тең бөлікке бөледі:

- **Q1 (25-перцентиль):** мәндердің 25%-ы осыдан төмен
- **Q2 (50-перцентиль):** медиана — мәндердің жарымы осыдан төмен
- **Q3 (75-перцентиль):** мәндердің 75%-ы осыдан төмен

Квартильдер деректердің "ортасын" ауытқуларсыз сипаттайды.

Квартильаралық аралық (IQR)

IQR = Q3 − Q1

IQR деректердің орталық 50%-ының таралуын көрсетеді. Ауытқулар (outliers) IQR есебіне кірмейді — сондықтан стандартты ауытқуға қарағанда тұрақтырақ.

Ауытқуларды анықтау ережесі

Егер мән **Q1 − 1.5×IQR** мәнінен кіші немесе **Q3 + 1.5×IQR** мәнінен үлкен болса — ауытқу (outlier) деп есептеледі. Бұл — Тьюкидің қорап диаграммасы (boxplot) ережесі.

1 минуттық тест: орташаға сенесің бе?

Егер үлестірім **симметриясыз** (қиғаш) болса, орташа мен медиана бір-бірінен алыстайды. Бұл жағдайда орташаға ғана сүйену талдаушыны адастырады.

Симметриялы үлестірім

Орташа ≈ Медиана ≈ Мода. Үшеуі де бірдей нүктеде шоғырланады. Орташа мән — дұрыс "орталық" өлшем. Мысал: қалыпты үлестірім.

Оңға қиғаш үлестірім

Орташа > Медиана > Мода. Оң жақтағы ұзын құйрық орташаны жоғары тартады. Медиана дәлірек "орталықты" береді. Мысал: табыс үлестірімі, баға деректері.

Солға қиғаш үлестірім

Орташа < Медиана < Мода. Сол жақтағы ұзын құйрық орташаны төмен тартады. Медиана тиімдірек. Мысал: өмір ұзақтығы, мамандық тәжірибесі.

⚠ Табыс туралы баяндамада "орташа жалақы" жоғары кіріс алатын азшылыққа байланысты жоғарылап кетуі мүмкін. Медиана — әділдірек өлшем.



3-ТАРАУ

Симметрия, қиғаштық, шың (Form)

Үлестірімнің "көңіл-күйі"

Үлестірімнің үшінші негізгі өлшемі оның **пішіні (form)** — симметриясы бар-жоғы, асимметрияның бағыты, және шыңдарының саны мен биіктігі. Бұл сипаттамалар деректің "мінезін" ашады.

Симметрия (Symmetry)

Үлестірімнің оң және сол жақтары айна бейнесіндей болса — симметриялы. Орташа = медиана = мода. Классикалық мысал — Гаусс (қалыпты) үлестірімі.

Қиғаштық (Skewness)

Үлестірімнің бір жаққа созылуы. Статистикалық өлшем: Пирсон немесе Фишер қиғаштығы. Оң мән — оң жаққа созылу; теріс мән — сол жаққа.

Шың биіктігі (Kurtosis)

"Шыңның" өткірлігі мен "құйрықтың" жуандығы. Жоғары kurtosis — тік шың, жуан құйрық (leptokurtic). Төмен kurtosis — жалпы, жіңішке құйрық (platykurtic).

Қиғаштық (skewness)

Қиғаштық — үлестірімнің орталығынан бір жаққа **асимметриялы таралу дәрежесін** сандық өлшейтін статистикалық көрсеткіш. Ол үлестірімнің қай жаққа "еңкейгенін" нақты анықтайды.

Оңға қиғаш (Positively Skewed)

Skewness > 0

Үлестірімнің оң жаққа ұзын құйрығы бар. Көпшілік мәндер сол жақта шоғырланған, ал бірнеше үлкен мән орташаны оңға тартады.

Мысалдар: жалақы деректері (бірен-саран миллионерлер бар), баға деректері, операциялық қателер санақтары, Пуассон үлестірімі.

Солға қиғаш (Negatively Skewed)

Skewness < 0

Үлестірімнің сол жаққа ұзын құйрығы бар. Мәндердің көпшілігі жоғары жақта шоғырланған, ал бірнеше өте төмен мән орташаны солға тартады.

Мысалдар: тест нәтижелері (оңай тест болса, барлығы жоғары алады), өмір ұзақтығы деректері, кейбір биологиялық өлшемдер.

Практикалық ереже

|Skewness| < 0.5 → шамалы қиғаштық; 0.5–1.0 → орташа; > 1.0 → күшті қиғаштық. Күшті қиғаштықта деректі трансформациялаған жөн (log, sqrt).

Мода, медиана, орташа: үш түрлі "орталық"

Үлестірімнің "орталығын" сипаттаудың үш классикалық жолы бар. Олар симметриялы үлестірімде бірдей мән береді, бірақ қисық (асимметриялы) үлестірімде бір-бірінен алыстайды. Осы айырым — деректің **бұрмаланып тұрғанының** белгісі.

Мода (Mode)

Ең жиі кездесетін мән. Категориялық деректер үшін жалғыз ыңғайлы орталық өлшем. Сандық үлестірімде шыңды (peak) белгілейді. Бірнеше мода болуы мүмкін (multimodal).

Медиана (Median)

Деректерді реттегенде ортасындағы мән. Ауытқуларға тұрақты (robust). Қиғаш үлестірімде орташаға қарағанда дәлірек "орталықты" береді. Жалақы, баға, жас талдауларында жиі пайдаланылады.

Орташа (Mean)

Барлық мәндердің қосындысын санына бөлу. Математикалық есептеулерге ыңғайлы. Симметриялы үлестірімде ең тиімді, бірақ ауытқуларға өте сезімтал. Статистикалық тесттердің негізі.

Егер орташа мен медиана арасындағы айырым үлкен болса — үлестірім асимметриялы. Бұл жағдайда "орташа" жалғыз қолданылмауы тиіс.

Пик (kurtosis): "қандай экстремум көп?"

Kurtosis — үлестірімнің **шыңының өткірлігін** және **құйрықтарының жуандығын** өлшейтін статистикалық сипаттама. Ол "шоктың" немесе "экстремалды оқиғалардың" жиілігі туралы маңызды ақпарат береді.

Мезокуртик (Mesokurtic)

Kurtosis ≈ 3 (немесе артық kurtosis ≈ 0). Қалыпты үлестірімге тән. Шың орташа биіктікте, құйрықтар орташа жуандықта. Эталон ретінде пайдаланылады.

Лептокуртик (Leptokurtic)

Kurtosis > 3. Тік өткір шың, жуан ("fat") құйрықтар. Экстремалды мәндер жиірек кездеседі. Қаржы нарықтарында жиі байқалады — "қара аққу" оқиғалары осыдан.

Платикуртик (Platykurtic)

Kurtosis < 3. Жалпақ шың, жіңішке құйрықтар. Мәндер кеңінен таралған, экстремалды оқиғалар сирек. Бірқалыпты (uniform) үлестірімге жақын.

Қаржы саласында "fat tail" — жуан құйрық — жоғары тәуекелдің белгісі. Егер үлестірімнің kurtosis мәні жоғары болса, сирек оқиғалар (дефолт, дағдарыс) болжамнан жиірек болады деген ескерту.

4-ТАРАУ

Ең маңыздысы — ықтималдық сипаттамасы

Үлестірімнің логикасы

Статистикалық үлестірімнің терең мәні тек жиіліктерді санауда ғана емес — әр мәнге **"қанша ықтимал?"** деген сұрақ қоюда. Осы сұрақтан жиілік үлестірімінен **ықтималдық үлестіріміне** өту басталады.



Дискретті үлестірім: PMF идеясы

Дискретті кездейсоқ шама үшін — яғни санауға болатын деректер үшін — **ықтималдық масса функциясы (PMF, Probability Mass Function)** пайдаланылады. PMF әр мәнге нақты ықтималдық мәнін береді.

PMF-тің анықтамасы

P(X = x) — X айнымалысы дәл x мәнін қабылдау ықтималдығы.

Негізгі шарттар:

- Барлық $P(X = x) \geq 0$
- Барлық мәндер бойынша $\Sigma P(X = x) = 1$

Бұл — ықтималдық аксиомаларынан тікелей туындайды: барлық мүмкін нәтижелердің ықтималдықтар жиыны 1-ге тең болады.

Классикалық дискретті үлестірімдер

- **Биномдық:** n тәжірибеде k жетістік саны (монета лақтыру)
- **Пуассон:** уақыт бірлігінде оқиға саны (шақыру орталығындағы қоңыраулар)
- **Геометриялық:** бірінші жетістікке дейін талап саны
- **Гипергеометриялық:** кері орналастырусыз іріктеме

Практикалық мысал

Монетаны 10 рет лақтырғанда "аяқ" 6 рет түсу ықтималдығы — биномдық PMF арқылы есептеледі: $P(X=6) = C(10,6) \times 0.5^6 \times 0.5^4 \approx 0.205$.

Үздіксіз үлестірім: PDF/аумақ

Үздіксіз кездейсоқ шама үшін нақты бір нүктенің ықтималдығы математикалық тұрғыдан **нөлге** тең. Маңыздысы — мәннің белгілі бір **аралыққа** түсу ықтималдығы. Осыдан ықтималдық тығыздығы функциясы (PDF) туындайды.

PDF (Probability Density Function)

f(x) — ықтималдық тығыздығы. Нақты мән ықтималдығы емес, тығыздық. P(a ≤ X ≤ b) = ∫[a,b] f(x)dx — аумақ ықтималдыққа тең. Барлық жерде f(x) ≥ 0 және ∫(−∞,+∞) f(x)dx = 1.

Классикалық үздіксіз үлестірімдер

Қалыпты (Normal): бойы, өлшеу қателері, табиғи процестер. **Экспоненциалдық:** күту уақыты, жабдық істен шығу мерзімі. **Бета:** ықтималдықтар мен үлестердің үлестірімі. **Гамма:** уақыт суммасы.

Гистограмма мен PDF байланысы

Деректер санаулы болса — гистограмма. Деректер жеткілікті болса және тегістелсе — PDF бағасы. n артқан сайын гистограмма PDF-ке жақындайды. Бұл — үлкен сандар заңының практикалық бейнесі.

CDF: "осыдан кіші" деген шындық

Жинақталған үлестірім функциясы (CDF, Cumulative Distribution Function) — кездейсоқ шаманың берілген мәннен кіші немесе тең болу ықтималдығын береді:

$$F(x) = P(X \leq x)$$

CDF-тің қасиеттері

- F(x) әрқашан 0-ден 1-ге дейін өседі
- Монотонды өсетін функция
- $F(-\infty) = 0$, $F(+\infty) = 1$
- Үздіксіз үлестірімде — тегіс қисық
- Дискреттіде — "баспалдақ" пішінде

Практикалық пайдасы

$P(a < X \leq b) = F(b) - F(a)$ — кез келген аралықтың ықтималдығын оңай табуға болады. Перцентильдерді де тікелей CDF арқылы оқуға болады.

CDF арқылы үлестірімдерді салыстыру

Екі топтың CDF-ін бір графикте салғанда:

- Егер бір CDF екіншісінің сол жағында тұрса — бірінші топтың мәндері жалпы кіші
- Айырым үлкен болса — үлестірімдер едәуір ерекшеленеді
- Айырымның тек белгілі аймақта болуы — топтар бір бөлікте ғана айырмашылықта

Колмогоров-Смирнов тесті CDF айырымының максимумын пайдаланып, үлестірімдер ерекшелігін тексереді.

Дерек масштабы мен "режимдер": бір үлестірім емес, бірнеше процесс

Үлестірімнің бойында **екі немесе одан да көп шың (bimodal / multimodal)** байқалса — бұл жиі дерек бірнеше ерекше топтан немесе режимнен жинақталғанының белгісі. Мұны байқамай "орташаны" есептеу — маңызды ақпаратты жасыруға алып келеді.

Өндіріс мысалы

Зауытта екі жұмыс ауысымы болса, олардың жабдықтары, жұмысшылары, температура жағдайы ерекше болуы мүмкін. Нәтижеде өлшеу мәндері екі деңгейге жинақталып, гистограммада екі шың пайда болады. "Бір орташа" — бұл деректі мүлде дұрыс сипаттамайды.

Топтарды ажыратудың қажеттілігі

Бимодальды үлестірімді анықтағаннан кейін деректерді топтарға бөліп, әр топтың жеке үлестірімін талдау керек. Топтардың анықтаушы белгілері: ауысым, жасы, аймақ, мерзім. Жеке үлестірімдер айқынырақ, ықтималдық модельдеуге жарамдырақ.

"Бір орташа" адастырады

Екі топтың орташасы 60 және 80 болса, жалпы орташа 70 шығады — бірақ бұл нәтиже ешбір топты дәл сипаттамайды. Шешім қабылдауда осындай "жалған орталық" қауіпті болуы мүмкін.

Деректегі жасырын іріктеу әсері: неге үлестірім бұрмаланады?

Үлестірім пішіні тек деректің "шынайы заңдылығын" емес, сонымен бірге **деректің қалай жиналғанын** да бейнелейді. Іріктеу тәсілі дұрыс болмаса, үлестірім бұрмаланып, жалған қорытынды беруі мүмкін.

Таңдау тәсілі үлестірімді өзгертеді

Ерікті қатысу (volunteer bias): зерттеуге тек мотивациялы адамдар қатысса — олардың үлестірімі жалпы халықтікіндей емес. Тіркелгендер ғана немесе белгілі бір шекті асқандар ғана деректе болса — құйрық қысқарады, нақты ауқым жасырын қалады.

Уақыт факторы

Процестің әртүрлі фазаларын (өсу кезеңі, тоқырау, дағдарыс) бір деректер жиынтығына қосу "кең" немесе "дұрыс емес пішіндегі" үлестірім береді. Мысалы, экономикалық өсу кезеңіндегі жалақы деректерін дағдарыс кезеңімен қосып талдасаңыз — нәтиже бұрмаланады.

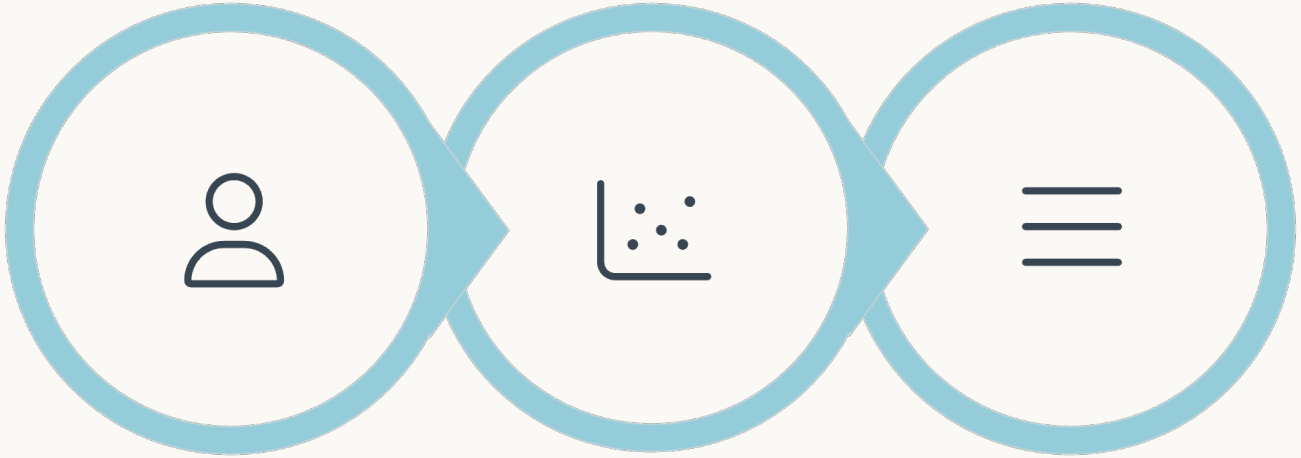
Қазақстан мысалы

Нарыққа өту кезеңінде (1990-шы жылдар) статистикалық жинау жүйесі жаңартылды: іріктемелі зерттеулер, жаңа есептеу тәсілдері енгізілді. Бұрын жинақталған деректер жаңа әдістемемен салыстырылмайтын болды — тарихи салыстырмалы талдауда мұны ескеру міндетті.

Үлестірім пішініне қарап, деректің "қалай жиналғанын" міндетті түрде сұрау — аналитиктің бірінші міндеті.

Эмпирикалық үлестірім: деректен "заңдылыққа" қарай

Деректер жинақталғанда бірінші қадам — **эмпирикалық жиілік үлестірімін** салу. Бұл — теорияға емес, нақты байқалған мәндерге негізделген "шынайы бейне". Содан кейін ол бейнені теориялық үлестірімдер арқылы модельдеу жүреді.



Эмпирикалық

Үлестірім таңдау

Параметр бағалау

Эмпирикалық үлестірімнің ерекшеліктері

- Нақты деректерге негізделген — ешқандай болжам жоқ
- Іріктеме кіші болса — "шулы" (тұрақсыз)
- n артқан сайын тұрақты бола бастайды
- Теориялық үлгімен салыстыру базасы

Ықтималдық тәсілімен байланысы

Үлкен сандар заңы бойынша: тәжірибе саны артқан сайын эмпирикалық жиілік шынайы ықтималдыққа жақындайды. Яғни n ұлғайса → гистограмма "тегістеліп" → теориялық PDF-ке ұқсас бола бастайды.

Ықтималдыққа апаратын көпір: үлестірімнен параметрге

Статистикалық талдаудың ең маңызды қадамдарының бірі — эмпирикалық деректен теориялық үлестірімге, сонан соң оның **параметрлеріне** өту. Бұл — "деректегі заңдылықты" бір немесе бірнеше санмен жинақтау процесі.

1

Деректер жинақталады

Бақылаулар жиналады, гистограмма немесе ECDF салынады.

2

Үлестірім пішіні талданады

Симметрия, қиғаштық, шың саны, құйрық жуандығы анықталады.

3

Теориялық үлгі таңдалады

Пішінге сәйкес үлестірім отбасы таңдалады (қалыпты, Пуассон, экспоненциалдық, т.б.)

4

Параметрлер бағаланады

МНК, МТП немесе байес тәсілімен параметрлер (μ, σ, λ) есептеледі.

Екі топтың гистограммалары ұқсас болса — олардың ықтималдық моделі де ұқсас болады деген қорытынды жасалады. Осы логика болжам жасаудың — "келесіде не болады?" деген сұраққа жауап берудің — математикалық негізін қалайды.

Эмпирикалық үлестірім: дерекке сүйенген "нақты бейне"

Эмпирикалық үлестірім — бақылау нәтижелерін тікелей санау арқылы алынатын "бастапқы шындық". Ол теориялық болжамсыз, тек нақты деректерге ғана сүйенеді.

Эмпирикалық үлестірімді қалай оқу керек?

- Жиі кездесетін мәндер — шоғырлану орталығы
- Сирек кездесетін мәндер — "шулы" немесе нақты ерекшелік пе?
- Симметриялы ма, қисайып тұр ма?
- Ауытқулар (outliers) байқалады ма?
- Бір шың ба, бірнеше шың ба?

Іріктеме мөлшерінің маңызы

Іріктеме **аз** болса: жиіліктер "шулы" — кездейсоқ ауытқулар бар. Әсіресе құйрық аймағы тұрақсыз. Кішкене іріктемеден модель жасау — қауіпті қорытынды.

Іріктеме **үлкен** болса: үлестірім тұрақталады. Теориялық модельмен салыстыру мағыналы бола бастайды. Статистикалық тесттер дәл жұмыс істейді.

Практикалық ереже: $n < 30$ болса — эмпирикалық үлестірімге абайлап қарау керек. $n > 100$ болса — гистограмма пішіні сенімдірек.

Үлестірімнің есебі: бақылау → санақ → ұлттық шешім

Үлестірім ұғымы тек математикалық абстракция емес — ол **ұлттық деңгейдегі шешімдердің** негізі. Халық санағы мен ұлттық статистика — нақты бөліністерді есептеудің ең іргелі тәсілі.

1999 жылғы санақ

Халық саны: 14 981,3 мың. Тәуелсіздіктен кейінгі бірінші толық санақ. Эмиграция мен демографиялық өзгерістерді бекітті.

1

2009 жылғы санақ

Халық саны: 16 009,6 мың. Өсім +1 028,3 мың. Қала/ауыл үлестірімі өзгерді. Этностық, жас, білім бөліністері жаңартылды.

2

2021 жылғы санақ

Халық саны: 19 186 015 адам. Алғаш рет онлайн форматта өтті. Урбанизация: 56,1% → 61,2%. Цифрлық деректер жинауға өтіс.

3

2025 жылғы бағыт

Деректерді автоматтандыру, іріктемелі зерттеулер. ОДМ мақсаттарымен интеграция. Дезагрегация: топтық бөліністер бойынша дәл есеп.

4

Осы деректер негізінде мемлекет бюджетін жоспарлайды, мектеп, аурухана, жол желісін орналастырады, мемлекеттік бағдарламаларды топтық бөліністерге сәйкес жасақтайды. Үлестірім — тек графика емес, саясаттың математикалық негізі.

Деректен үлестірімге дейінгі қадамдар тізбегі

Жоғары сапалы үлестірімге жету үшін деректер жинау мен дайындаудың **жүйелі процесін** сақтау өте маңызды. Бір қадамның қателігі бүкіл талдауды бұзуы мүмкін.



1. Бақылау бірлігін анықтау

Талдаудың негізгі бірлігі кім немесе не? Адам ба, үй шаруашылығы ма, кәсіпорын ба, оқиға ма? Бұл анықтамасыз деректер дұрыс жиналмайды — бақылаулар бір деңгейде болуы шарт.



2. Шикі жазбаларды тазарту

Дубликаттар, қате мәндер (мысалы: жас = -5 немесе 200), жетіспейтін мәндер (NA/null) тексеріледі. Жетіспейтін мәндерді өткізіп жіберу, орташамен толтыру немесе арнайы санат ретінде кодтау — таңдау мотивтелуі керек.



3. Айнымалыларға тип беру

Сандық пен категориялық айнымалылар ажыратылады. Өлшем бірліктері келісіледі (мысалы: теңге ме, мың теңге ме). Нақты немесе дискретті екені белгіленеді.



4. Арнаулы кодтау жасау

Категориялық айнымалылар үшін: санат атауы, реті, "басқа" немесе "жауап берген жоқ" категориясы алдын ала келісіледі. Шекаралар анық болмаса — гистограммалар сравнительдік жарамсыз болады.

Бұрмаланудың негізгі 3 көзі

Үлестірім "өтірік айтпайды" — ол деректегі заңдылықты шынайы бейнелейді. Бірақ егер деректер қате жиналса немесе өлшенсе — үлестірім де **бұрмаланған шындықты** бейнелейді. Үш негізгі бұрмалану көзін білу — аналитиктің міндеті.

1. Қате өлшеу

Зертхана жабдықтарының калибрі дұрыс болмаса — бүкіл деректер жүйелі ауытқумен жазылады. Бақылаушылар арасындағы ережені бірдей қолданбау (inter-rater inconsistency) да үлестірімді "жалпайтады". Шешімі: стандарттау, калибрация, бақылаушы нұсқаулықтары.

2. Таңдаманың қисайуы

Кей топтар деректерде жиі (over-represented), кейбіреулері сирек (under-represented) қамтылады. Интернет сауалнамасына тек жас, білімді топ жауап берсе — үлестірім жалпы халықты бейнелемейді. Шешімі: салмақтау (weighting), стратификациялық іріктеме.

3. Санат шекараларының сәйкессіздігі

Гистограммадағы "аралықтар" (bins) бірдей ен болмаса — үлестірімнің пішіні өзгеріп кетеді. Мысалы: 0–10, 10–20, 20–50, 50–100 — соңғы екі аралық кеңірек, олар "төмен" жиілік көрсетеді. Шешімі: тең ендегі аралықтар немесе айқын жазылған анықтамалар.

PMF/PDF есептеу логикасы: бірдей үлгі, әртүрлі дерек

Дискретті және үздіксіз деректер үшін ықтималдықты есептеу логикасы ерекше. Бірақ екеуінің де негізінде бір ой жатыр: **мән немесе аралықтың салыстырмалы жиілігін ықтималдыққа айналдыру**.

Дискретті PMF логикасы

Қадам 1: Барлық мүмкін мәндерді анықтау (x_1, x_2, ..., x_n)

Қадам 2: Әр мәннің жиілігін есептеу: count(x_i)

Қадам 3: Жалпы санға бөлу: p(x_i) = count(x_i) / n

Тексеру: Σ p(x_i) = 1 болуы шарт

Гистограммада: X осі мәндер, Y осі ықтималдық. Бағандар арасында бос орын — дискреттілік белгісі.

Үздіксіз PDF логикасы

Мәселе: Нақты бір нүкте ықтималдығы 0 — өйткені үздіксіз шкалада шексіз нүктелер бар.

Шешім: Аралықты пайдалану. P(a ≤ X ≤ b) = ∫[a to b] f(x)dx

Практикада: Мәндерді "қораптарға" (bins) жіктеп, салыстырмалы жиілікті тіктөртбұрыш ауданымен теңестіреміз. Бин ені × жиілік тығыздығы = аралық ықтималдығы.

Ереже: Гистограмма — PDF-тің дискретті жуықтауы. n артқан сайын жуықтау дәлдігі артады.

Эмпирикалық үлестірім: "дерек айтқан шындық" қалай алынады

Эмпирикалық үлестірімді есептеудің нақты алгоритмі бар. Бұл алгоритмді түсіну — статистикалық бағдарламалық жасақтаманың (R, Python, SPSS) ішінде не болып жатқанын ұғуға мүмкіндік береді.

01

Позиция табу

Әр бақылауда айнымалының мәні бойынша позиция анықталады. Деректер реттелген болса — рангі табылады; реттелмеген болса — тікелей мән пайдаланылады.

02

Жиілік жинақтау

Мәндер аралықтарға (bins) орналастырылады немесе нақты мәндер бойынша санақталады. Нәтиже: frequency table — кесте түрінде жиіліктер.

03

Салмақтау

Тең салмақ (барлық бақылауға бірдей 1/n) немесе ерекше салмақ (stratified sampling кезінде) қолданылады. Салмақтау іріктеме қисаюын түзетеді.

04

Салыстырмалы жиілікке айналдыру

P(үлес) = жиілік / жалпы сан. ECDF үшін: $\hat{F}(x)$ = (x-тен кіші немесе тең мәндер саны) / n. Нәтиже 0-ден 1-ге дейінгі шкалада — ықтималдыққа дайын.

CDF-тің күші: бір графикте "толық көрініс"

Жиілік гистограммасы жақсы визуализация болса да, **CDF (жинақталған үлестірім функциясы)** бірқатар жағдайда одан артық ақпарат береді. Бір графикте — үлестірімнің толық сипаттамасы.

CDF-тің артықшылықтары

- "X-тен кіші мән" ықтималдығы — тікелей оқылады
- Медиана: F(x) = 0.5 болатын нүкте
- Квартильдер: F(x) = 0.25 және 0.75
- Перцентильдер: кез келген деңгей оқылады
- Ауытқулар: CDF-тің кенеттен өсетін жерлері

Шыңдар мен құйрықтар CDF-та

Шың бар жерде CDF тік өседі (сол аралықта мәндер жиі). Құйрық аймағында CDF жайлап өседі (мәндер сирек). Осыдан "бірқалыпты өсуден ауытқу нүктелерін" табуға болады.

Екі үлестірімді CDF арқылы салыстыру

Екі топтың CDF-ін бір графикте салғанда:

- Егер бір CDF үнемі екіншісінің солында болса — бірінші топтың мәндері жалпы кіші (бірінші доминанттылық)
- Айырым тек бір аймақта болса — топтар тек ол диапазонда ерекшеленеді
- CDF-тер қиылысса — бір топ кіші мәндерде, екіншісі үлкен мәндерде артық

Колмогоров-Смирнов статистикасы — екі CDF арасындағы максималды тік қашықтық. Ол неғұрлым үлкен болса, үлестірімдер соғұрлым ерекше.

Үлестірім параметрін іздеу: медиана/квартилиден үлгіге дейін

Эмпирикалық үлестірімнің пішінін анықтаған соң, оны сипаттайтын **параметрленген теориялық үлгіні** таңдау кезеңі басталады. Бұл — "дерекке жақын модель" іздеу процесі.

1

Пішінді зерттеу

Симметриялы ма? Қиғаш па? Бір шың ба? Құйрықтар қалай? Осыдан үлестірім отбасы (қалыпты, экспоненциалдық, бета...) шектеледі.

2

Параметрленген үлгі таңдау

Таңдалған отбасының параметрлері (μ, σ, λ, α, β) деректен МТП немесе сәйкестіру тәсілімен бағаланады.

3

Сәйкестікті тексеру

Хи-квадрат, Колмогоров-Смирнов немесе Андерсон-Дарлинг тесті арқылы үлгі мен деректің сәйкестігі тексеріледі. Қай аймақта айырым бар — құйрық па, ортасы ма?

Жақсы модельдің белгілері

- Орталық аймақта теориялық және эмпирикалық жиіліктер жақын
- Құйрықтар орынды модельденген
- Сәйкестік тесттері мәнді ауытқу көрсетпейді
- Параметрлер интерпретацияланатын мағынаға ие

Мақсат: болжам мен салыстыру

Жақсы модель жасалғаннан кейін: болжам жасауға (жаңа бақылаулар қай аралыққа түседі?), топтарды статистикалық тестпен салыстыруға (екі топтың үлестірімдері ерекше ме?), тәуекел бағалауға (сирек оқиғалардың ықтималдығы қандай?) мүмкіндік туады.

Қазақстандағы халық санағы: үлестірімдерге кіретін "үлкен бастапқы сүзгі"

Халық санағы — ұлттық деңгейдегі ең іргелі деректер жинау процесі. Қазақстанда санақ **10 жылда бір рет** жүргізіледі және ол бүкіл демографиялық, әлеуметтік, экономикалық талдаулардың негізін қалайды.

19.2M

2021 жылғы халық

19 186 015 адам — 2021 жылғы санақ бойынша Қазақстанның жалпы халқы.

61.2%

Қала үлесі

2021 жылы қалалық халықтың үлесі 56,1%-дан 61,2%-ға дейін өсті — урбанизация жеделдеді.

3

Санақ саны

1999, 2009 және 2021 жылдары — тәуелсіздік кезеңіндегі үш толық ұлттық санақ.

2021 жылғы санақ алғаш рет **онлайн форматта** да өтті (1 қыркүйек–30 қазан). Цифрлық жинау — деректердің жылдамдығы мен дәлдігін арттырды. Санақ нәтижелері бюджет жоспарлауда, білім, денсаулық, инфрақұрылым бөліністерінде пайдаланылды — яғни үлестірімдер тікелей мемлекеттік шешімге айналды.



Дезагрегация: "орташа" емес, әр топтың үлестірімі

"Жалпы орташа" жиі жеткіліксіз. Теңсіздікті, мүмкіндіктер айырмасын және нақты топтардың жағдайын түсіну үшін деректерді **дезагрегациялау** — бөліктерге бөліп, әр топтың үлестірімін жеке талдау — міндетті талап болып отыр.

Жынысы бойынша бөлу

Жалақы, білім деңгейі, еңбек нарығына қатысу — жыныс бойынша бөлінген үлестірімдер теңсіздікті анық байқатады. Жалпы орташа алшақтықты жасырады.

Аймақ/география бойынша

Қалалық және ауылдық аймақтар арасындағы айырым — экономикалық, денсаулық, білім көрсеткіштерінде кедергілерді анықтайды. Региональдық үлестірімдер ресурстарды дұрыс бөлуге негіз береді.

Жасы мен мүгедектігі бойынша

Қазақстанның ОДМ жүйесінде "ешкім артта қалмауы" үшін барлық осалды топтардың деректері бөлек жиналады. 2030-ға дейінгі мақсаттарға жету — топтық үлестірімдерсіз мүмкін емес.

Іс жүзінде қиындық: деректі уақытылы және бөлек жинау — ресурс сыйымдылығы жоғары. Бірақ бұл инвестиция — дәл бағытталған мемлекеттік саясаттың алғышарты.

Дезагрегация — статистикалық үлестірімнің саяси-практикалық парасаты. Деректер жинақталғанда қандай топтар бойынша бөлу жоспарланды? Осы сұрақ зерттеудің басында қойылуы тиіс — соңында емес.