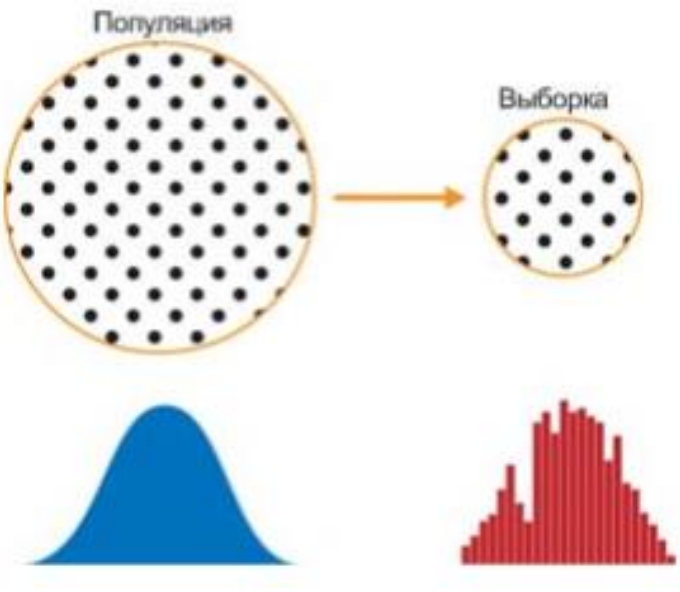
Деректер мен үлгілерді бөлу

Танымал пікір үлкен деректер дәуірі іріктеу қажеттілігін жоққа шығарады деп қате айтады. Шын мәнінде, өзгермелі сапа мен өзектілік деректерінің тез таралуы әр түрлі мәліметтермен тиімді жұмыс жасау және жылжуды азайту құралы ретінде іріктеу қажеттілігін күшейтеді. Үлкен мәліметтерге негізделген жобада да, болжамды модельдер, әдетте, үлгілердің көмегімен жасалады және сыналады. Үлгілер әртүрлі сынақтарда да қолданылады (мысалы, баға белгілеу, веб-өңдеу). Суретте. 2.1 осы тараудың тұжырымдамаларын қолдайтын схема көрсетілген. Сол жақта статистикада негізгі, бірақ белгісіз үлестірімге бағынатын халық бар. Қол жетімді жалғыз нәрсе - бұл деректер тақтасы және оның оң жақта көрсетілген эмпирикалық таралуы. Сол жақтан оң жаққа "жету" үшін таңдау процедурасы қолданылады (көрсеткі көрсетілген). Дәстүрлі статистика негізінен сол жаққа бағытталған, популяция хактері туралы байыпты болжамдарға негізделген теорияны қолданады. Қазіргі статистика оң жаққа көшті, онда мұндай болжамдар қажет емес.



2.1.-сурет. Популяцияны іріктеумен салыстыру

Жалпы алғанда, деректерді талдаушылар сол жақтың теориялық табиғаты туралы алаңдамауы керек; оның орнына таңдау процедуралары мен қолда бар мәліметтерге назар аудару керек. Рас, кейбір маңызды нәрселер бар ерекшеліктер. Кейде деректер модельдеуге болатын физикалық процестен жасалады. Ең қарапайым мысал — монетаны лақтыру: ол биномдық үлестіруге бағынады. Кез - келген нақты биномдық жағдай (ку - ішу немесе сатып алмау, алаяқтық немесе алаяқтық емес, басу немесе басу) монетамен тиімді түрде модельденуі мүмкін (әрине, бүркіттің қону ықтималдығымен). Мұндай жағдайларда популяция туралы түсінігімізді қолдану арқылы қосымша ақпарат алуға болады.

Кездейсоқ таңдау және аралас іріктеу

Үлгі дегеніміз-статистикада популяция немесе популяция деп аталатын үлкен мәліметтер жиынтығынан алынған мәліметтер жиынтығы. Статистикадағы Популяция биологиядағы популяциямен бірдей емес - бұл үлкен, берілген, бірақ көбінесе теориялық немесе қиялды мәліметтер жиынтығы.

Негізгі терминдер

Үлгі (sample)- үлкенірек деректер жиынтығының ішкі жиынтығы.

Популяция (population) үлкен мәліметтер жиынтығы немесе мәліметтер жиынтығы идеясы. Синонимі: бас жиынтық.

N (n) -популяция мөлшері (іріктеме).

Кездейсоқ таңдау (random sampling) элементтерді кездейсоқ ретпен іріктеу.

Стратификацияланған іріктеу (stratified sampling) популяцияны страталарға бөлу және әрбір стратадан элементтерді кездейсоқ іріктеу.

Жай кездейсоқ іріктеу (simple random sampling)- таралымды страталарға Бөлмей, кездейсоқ іріктеу нәтижесінде алынатын іріктеу.

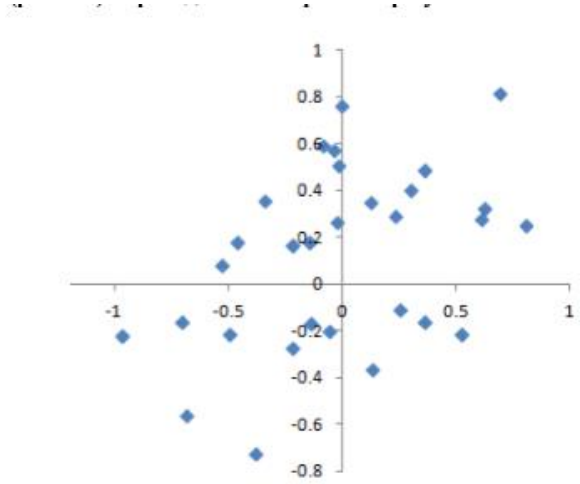
Офсеттік үлгі (sample bias) - бұл популяцияны бұрмаланған түрде көрсететін үлгі.

Кездейсоқ таңдау дегеніміз-іріктеуден өткен халықтың әр қол жетімді мүшесінің әр алу кезінде іріктеуге тең мүмкіндігі бар процесс. Алынған үлгі қарапайым кездейсоқ үлгі деп аталады. Іріктеу әр ойықтан кейін болашақта мүмкін болатын қайта іріктеу үшін бақылаулар популяцияға қайтарылған кезде қайтарумен жүзеге асырылуы мүмкін. Альтер сияқты-Деректер мен үлгілерді жергілікті таңдау қайтарусыз жүзеге асырылуы мүмкін және бұл жағдайда таңдалған бақылаулар болашақ ойықтар үшін қол жетімді емес.

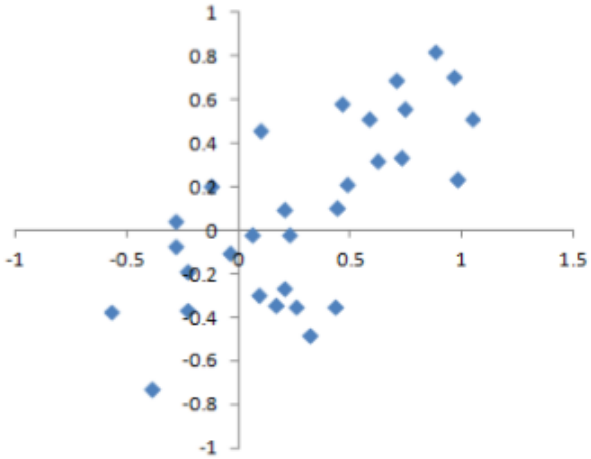
Деректер сапасы көбінесе бағалау жүргізілгенде немесе үлгі негізінде модель жасалғанда олардың санынан үлкен мәнге ие болады. Деректер ғылымындағы деректердің сапасы жеке мәліметтер нүктелерінің толықтығымен, форматының реттілігімен, тазалығымен және дәлдігімен байланысты. Статистика ғылымы мұнда өкілдік ұғымын қосады.

Орын ауыстыру.

Статистикалық ығысу дегеніміз жүйелі және деректерді өлшеу немесе іріктеу процесінде пайда болатын өлшеу және таңдау қателерін білдіреді. Кездейсоқ мүмкіндікке байланысты пайда болатын қателер мен бұрмалануға байланысты қателер арасында айырмашылық жасау керек. Мылтықты нысанаға атудың физикалық процесін қарастырыңыз. Мақсаттың абсолютті орталығын жеңу әрдайым болмайды немесе тіпті болмайды. Объективті емес процесс қате жібереді, бірақ ол кездейсоқ болады және кез-келген бағытқа бет бұрмайды (сурет. 2.2). Суретте келтірілген. 2.3 нәтижелер офсеттік процесті көрсетеді-х бағытында да, у бағытында да кездейсоқ қате бар, бірақ одан тыс орын ауыстыру бар. Кадрлар жоғарғы оң жақ квадрантқа түсу тенденциясын көрсетеді.



2.2.-Сурет. Реттелген көрінісі бар мылтықтың шашырау диаграммасы



2.3.-Сурет. Мылтықтың атыс шашырауы диаграммасы

Ауыстыру әр түрлі формада болады және көрінетін немесе көрінбейтін болуы мүмкін. Егер нәтиже шынымен бұрмалануды білдірсе (мысалы, эталонға немесе нақты мәндерге сүйене отырып), бұл статистикалық немесе машиналық оқыту моделі дұрыс орнатылмағанын немесе маңызды айнымалы ескерілмегенін білдіреді.

Жүйелі таңдау қатесі

Бейсбол ойыншысы йога Берраны (йога Берра) парафраздау үшін "егер сіз не іздеп жүргеніңізді білмесеңіз, мұқият қарап шығыңыз және сіз оны таба аласыз". Іріктеудің жүйелі қателігі (немесе іріктеу кезіндегі бейтараптық, selection bias) деректерді таңдау тәжірибесіне жатады - саналы немесе бейсаналық. Осылайша, ол алдамшы немесе қысқа мерзімді қорытындыға әкеледі.

Негізгі терминдер

Ауыстыру (bias) жүйелік қате.

Деректерді тарақтау (data snooping) қызықты нәрсені табу үшін деректерді егжей-тегжейлі тексеру.

Шексіз іздеу эффектісі (fast search effect) деректерді бірнеше рет модельдеуден немесе алдын - ала айнымалылардың көп мөлшерімен деректерді модельдеуден туындайтын орын ауыстыру немесе репродуктивтілік.

Егер сіз гипотезаны анықтап, оны тексеру үшін жақсы тәжірибе жасасаңыз, онда сіз алынған тұжырымға сенімді бола аласыз. Алайда, көбінесе олай емес. Оның орнына, үлгілерді көру үшін қол жетімді деректерге жиі назар аударыңыз. Бірақ үлгі нақты ма, әлде бұл жай ғана деректерді тарайтын өнім ме, яғни қызықты нәрсе пайда болғанша деректерді егжей-тегжейлі қайта қарау керек пе? Статистиктер арасында: "Егер сіз деректерді ұзақ уақыт бойы азаптасаңыз, онда ерте ме, кеш пе олар мойындалады"деген сөз танымал. Гипотезаны эксперимент арқылы тексерген кездегі құбылыс пен қол жетімді деректерді іздейтін құбылыс арасындағы айырмашылықты келесі ойлау экспериментімен түсіндіруге болады.

Біреу сізге бүркіт салған монетаны қатарынан 10 рет лақтыруға мәжбүр етуі мүмкін деп Айтады делік. Сіз қоңырауды қабылдайсыз (эксперимент эквиваленті), сынаушы монетаны 10 рет лақтыра бастайды және әр кезде бүркіт жоғары қарай қонады. Сіз бұл адамға қандай да бір ерекше талантты қосқаныңыз анық - монетаның 10 лақтыруы нәтижесінде ол жай кездейсоқ бүркітке айналады, ол 1000-нан 1 — ге жетеді. Енді стадиондағы диктор барлық қатысушылардан 20 мың сұрайды делік. ер адам монетаны 10 рет лақтырып, стадион қызметкеріне хабарлады, егер олар қатарынан 10 бүркіт құлап кетсе. Стадиондағы біреудің 10 бүркітке жету мүмкіндігі өте жоғары (99% — дан астамы - 1 минус, ешкім 10 бүркітке ие болмайды). Әрине, стадионда 10 бүркіт алатын адамның (немесе адамдардың) ретроактивті таңдауы оның ерекше таланты бар дегенді білдірмейді-бұл жай ғана сәттілік. Үлкен деректер жиынтығын қайта қарау деректер ғылымында негізгі құндылық ұсынысы болғандықтан, жүйелі таңдау қатесіне мұқият назар аудару керек. Деректерді талдаушылар үшін ерекше маңызы бар жүйелі іріктеу қатесінің нысанын Elder Research компаниясының негізін қалаушы, деректерді терең талдау саласындағы құрметті консалтингтік компания Джон Элдер шексіз іздеудің әсері деп атайды. Егер сіз әртүрлі модельдерді бір рет жасамасаңыз және ауыр деректер жиынтығы жағдайында әртүрлі сұрақтар қойсаңыз, онда сіз қызықты нәрсе таба аласыз. Табылған нәтиже шынымен назар аударуға тұрарлық нәрсе ме, әлде бұл кездейсоқ шығарылу ма? Бұған қарсы қорғаныс шараларын кейінге қалдырылған деректері бар бақылау жиынтығын (holdout), ал кейде бірнеше бақылау жиынтығын қолдану арқылы қабылдауға болады, соның негізінде нәтижелілікті растауға болады. Сонымен қатар, Элдер сонымен қатар деректерді терең талдау моделін ұсынатын болжамды қауымдастықтардың сенімділігін тексеру үшін мақсатты араластыру (мақсатты араластыру — бұл пермутациялық тест) деп аталатын нәрсені қолдануды жақтайды. Статистикада жүйелі іріктеу қатесінің типтік формалары, шексіз іздеу әсерінен басқа, кездейсоқ емес таңдауды (үлгілерді іріктеу кезінде орын ауыстыру), кремді кетіру қағидаты бойынша іріктеу нәтижесінде алынған мәліметтерді, белгілі бір статистикалық әсерді баса көрсететін уақыт аралықтарын таңдауды және нәтижелер "қызықты" болып көрінетін экспериментті тоқтатуды қамтиды..

Орташа Регрессия

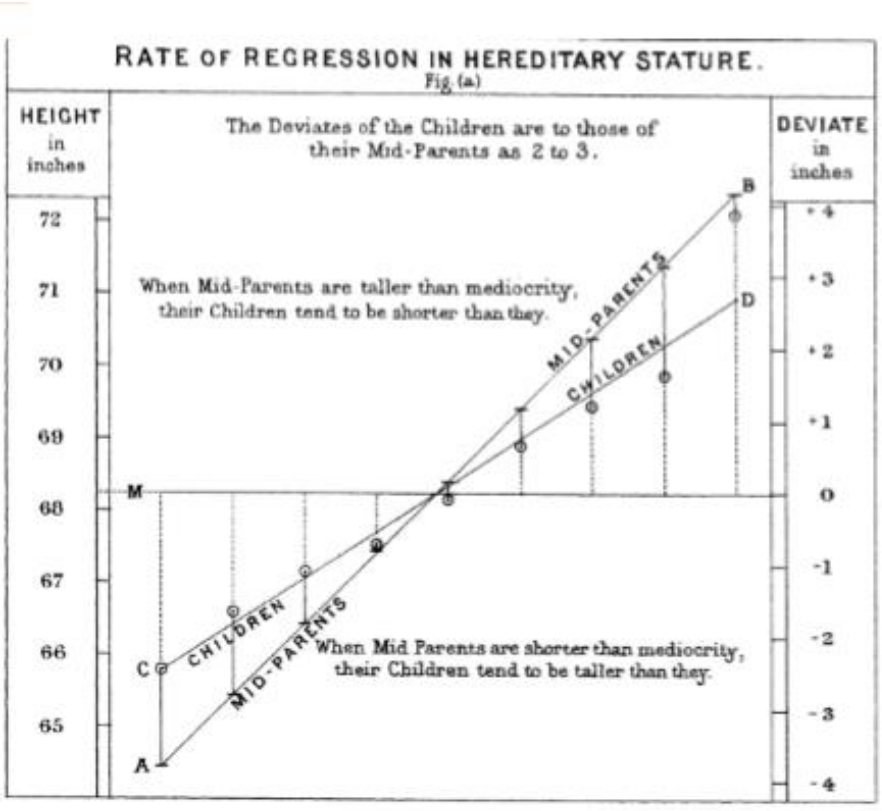
Орташа мәнге регрессия берілген айнымалының дәйекті өлшеулерімен байланысты құбылысты білдіреді: шекті бақылаулар неғұрлым орталық бақылаулармен бірге жүреді. Шекті мәнге ерекше назар мен мағына беру жүйелі таңдау қатесінің бір түріне әкелуі мүмкін.

Спорт әуесқойлары "жыл бастаушысы және бұрынғы көшбасшының дағдарысы" құбылысымен таныс. Өз мансабын белгілі бір маусымда бастаған спортшылардың арасында (жаңадан келгендер класы) әрқашан барлық негізгі спортшыларға қарағанда нәтижелі болады. Әдетте, бұл "жыл бастаушысы" келесі жылы бірдей нәтижелерге қол жеткізе алмайды. Неліктен? Спорттың барлық дерлік түрлерінде, кем дегенде доппен немесе шайбамен командалық жарыстарда, жалпы өнімділікте маңызды рөл атқаратын екі элемент бар:

шеберлік;

сәттілік.

Орташа мәнге Регрессия-бұл жүйелік іріктеу қатесінің белгілі бір формасының салдары. Біз ең жақсы өнімділікке ие жалдаушыны таңдағанда, шеберлік пен сәттілік бұған ықпал етуі мүмкін. Келесі маусымда ол әлі де орнында болады, бірақ көп жағдайда сәттілік болмайды, сондықтан оның тиімділігі төмендейді - ол қалпына келеді. Бұл құбылысты алғаш рет Фрэнсис Галтон 1886 жылы анықтаған [Galton - 1886], ол оны генетикалық тенденцияларға байланысты сипаттаған; мысалы, өте ұзын еркектердің балалары әкелері сияқты ұзын болмауға бейім (сурет. 2.5).



Сурет. 2.5. Галтонның зерттеу жұмысы, онда ол регрессия құбылысын ортаға анықтады

Жүйелі таңдау қатесінің негізгі идеялары

- *Гипотезаны анықтау және одан әрі деректерді жинау, рандомизация және кездейсоқ таңдау қағидаттарына сәйкес, орын ауыстырудан қорғауды қамтамасыз етеді.*
- *Деректерді талдаудың барлық басқа нысандары деректерді жинау/талдау процесінде туындайтын бұрмалану қаупіне ұшырайды (деректерді терең талдауда модельдерді бірнеше рет орындау, Статистикалық зерттеу кезінде деректерді тарақтау және қызықты оқиғаларды ретроактивті түрде таңдау).*

Статистиканы іріктеп бөлу

Статистиканың "іріктемелі үлестіру" термині белгілі бір іріктемелі статистиканың, яғни іріктемелі статистикалық шаманың бір популяциядан шығарылатын іріктеменің ауырсыну санына бөлінуін білдіреді. Классикалық статистика ғылымының едәуір бөлігі (шағын) үлгілерден және (өте үлкен) популяциялардан статистикалық тұжырымдар алумен айналысады.

Негізгі терминдер

Үлгі статистикасы (sample statistics) үлкен популяциядан алынған деректерді таңдау үшін есептелетін метрикалық көрсеткіш. Синонимі: іріктемелі статистикалық шама.

Деректерді бөлу (data distribution) деректер жиынындағы жеке мәндердің жиіліктік таралуы.

Іріктемелі үлестіру (sampling distribution) көптеген іріктемелерде немесе қайталанған іріктемелерде іріктемелі статистиканың жиіліктік таралуы.

Орталық шектік теорема (central limit theorem) іріктемелі үлестіру тенденциясы іріктеме көлемі ұлғайған сайын қалыпты форманы қабылдау. Синонимі: CPT.

Стандартты қате (standard error) көп санды үлгілердегі үлгі статистикасының өзгермелілігі (стандартты ауытқу) (деректердің жеке мәндерінің өзгергіштігіне жататын стандартты ауытқумен шатастырмау керек).

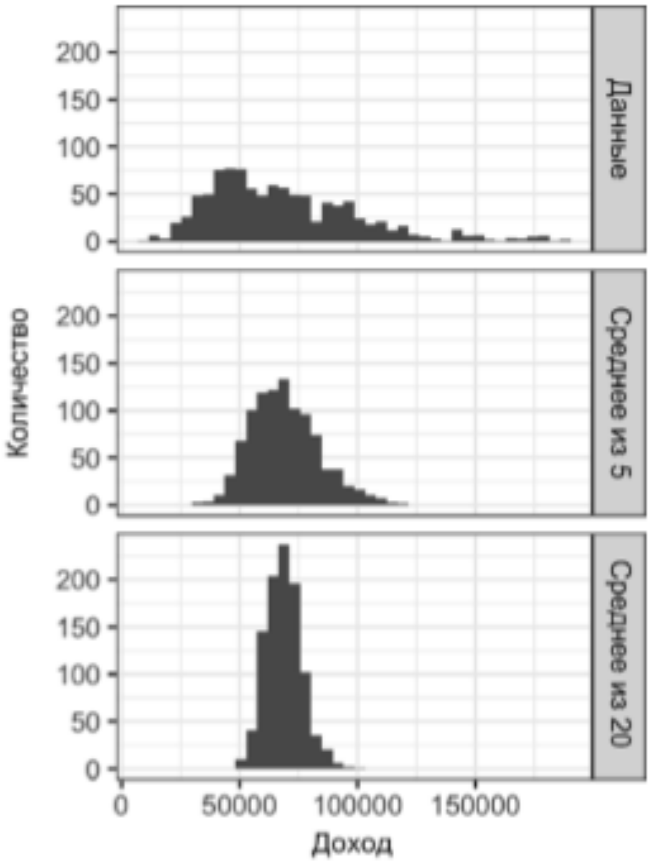
Әдетте, үлгі бір нәрсені өлшеу үшін (үлгі статистикасын қолдана отырып) немесе бір нәрсені модельдеу үшін (статикалық немесе машиналық оқыту моделін қолдана отырып) алынады. Біздің бағалауымыз немесе моделіміз үлгіге негізделгенін ескере отырып, ол ауытқуларға, яғни статистикалық қателіктерге ие болуы мүмкін; егер біз басқа үлгіні алып тастауды шешсек, ол басқаша болуы мүмкін. Біз оның қаншалықты өзгеше болуы мүмкін екенін білуге қуаныштымыз - негізгі мәселе — іріктеменің өзгергіштігі, яғни бағалау сынамалар арасында қаншалықты өзгереді. Егер бізде көп деректер болса, онда біз қосымша үлгілерді алып, үлгі статистикасының таралуын тікелей байқай аламыз. Әдетте, біз өз бағалауымызды немесе моделімізді қол жетімді деректерді қолдана отырып есептейміз, сондықтан популяциядан қосымша үлгілерді алу мүмкіндігі әрдайым бола бермейді.

Деректерді бөлуге қарсы үлгіні бөлу.

Деректерді бөлу деп аталатын жеке деректер нүктелерінің таралуын және үлгіні бөлу деп аталатын үлгі статистикасының таралуын ажырату маңызды

Үлгі статистикасының таралуы, мысалы, орташа, деректердің өзінен гөрі тұрақты және қоңырау тәрізді болуы мүмкін. Статистика неғұрлым көп болса, соғұрлым ол орынды болады. Сонымен қатар, үлгі неғұрлым көп болса, үлгі статистикасын бөлу соғұрлым тар болады.

Мұны өтініш берушілердің соты үшін жылдық кірісті қолдана отырып мысал келтіреді несие клубы Lending Club (қараңыз: разд. "Шағын мысал: несие қайтарылмайтынын болжау" 6-тарау, онда мәліметтер сипатталған). Осы мәліметтерден үш үлгіні алайық: 1000 мәні бар Үлгі, 5 мәннен 1000 орташа үлгі және 20 мәннен 1000 орташа үлгі. Содан кейін сурет салу арқылы әр үлгінің гистограммасын саламыз. 2.6.



2.6. -сурет. Өтініш берушілердің 1000 сотының жылдық кірістерінің гистограммасы (жоғарыда), содан кейін n = 5 (ортасында) және N = 20 (төменде)өтініш берушілердің саны бар 1000 орташа

Жеке деректер мәндерінің гистограммасы кеңінен орналастырылған және кірістер туралы мәліметтермен күтілгендей жоғары мәндерге кесілген. 5 және 20 мәндерінің орташа

гистограммаларының екеуі де ықшам және қоңырау тәрізді. Төменде ggplot2 визуализация пакетін қолдана отырып, осы гистограммаларды құратын R-дегі код үзіндісі келтірілген.

```
library(ggplot2) # взять простую случайную выборку

samp_data <- data.frame(income=sample(loans_income, 1000), type='data_dist')
# взять выборку средних из 5 значений

samp_mean_05 <- data.frame( income = tapply(sample(loans_income, 1000*5), rep(1:1000, rep(5, 1000)), FUN=mean),

 type = 'mean_of_5')

# взять выборку средних из 20 значений samp_mean_20 <- data.frame( income = tapply(sample(loans_income, 1000*20),

 rep(1:1000, rep(20, 1000)), FUN=mean),

 type = 'mean_of_20')

# связать кадры данных data.frames и конвертировать тип в фактор income <- rbind(samp_data, samp_mean_05, samp_mean_20) income$type = factor(income$type,

 levels=c('data_dist', 'mean_of_5', 'mean_of_20'),

 labels=c('Данные', 'Среднее из 5', Среднее 20'))

# построить гистограммы ggplot(income, aes(x=income)) + geom_histogram(bins=40) + facet_grid(type ~ .)
```

Орталық шекті теорема

Орталық шекті теорема деп аталатын құбылыс көптеген үлгілерден алынған орташа мәндер таныс омыртқа тәрізді қалыпты қисыққа ұқсайды дейді (қараңыз. "Қалыпты үлестіру" осы тарауда келтірілген), егер бастапқы популяция қалыпты бөлінбесе де, егер үлгілердің мөлшері жеткілікті үлкен болса және деректердің қалыптыдан ауытқуы тым жоғары болмаса. Орталық шекті теорема статистикалық тұжырым үшін үлгіні бөлуді есептеуде қолданылатын t-бөлу, атап айтқанда сенімділік интервалдары және статистикалық гипотезаларды тексеру сияқты қалыпты үлестірумен жуықтау формулаларын қолдануға шақырады. Дәстүрлі статистикалық тексерулерде орталық шекті теоремаға көп көңіл бөлінеді, өйткені ол сенімді интервалдар мен статистикалық гипотезаларды тексеру механизмінің негізінде жатыр, олар өздері осындай мәтіндердің мазмұнын алады. Деректер талдаушылары оның рөлі туралы білуі керек, бірақ гипотезаларды ресми тексерудің және деректер ғылымындағы сенімділік аралықтарының рөлі аз болғандықтан және қандай да бір жолмен әрдайым жүктеу бар болғандықтан, баға - тральды шекті теорема деректер ғылымының тәжірибесінде ерекше маңызды рөл атқармайды.

Стандартты қате стандартты қате - бұл статистика үшін үлгіні бөлудегі өзгергіштікті қамтамасыз ететін жалғыз метрикалық көрсеткіш. Стандартты қатені статистиканы

қолдана отырып, үлгі мәндерінің стандартты ауытқуына және N үлгінің мөлшеріне сүйене отырып бағалауға болады:

$$\text{Стандартная ошибка} = \frac{s}{\sqrt{n}}.$$

Үлгі мөлшері ұлғайған сайын стандартты қате күріште байқалғанға сәйкес азаяды. 2.6. Стандартты қате мен үлгіні бөлу арасындағы байланыс кейде N - ден квадрат түбір ережесі деп аталады: стандартты қатені 2 есе қысқарту үшін үлгінің мөлшерін 4 есе көбейту керек. Стандартты қате формуласының сенімділігі орталық шекті теоремадан туындайды (алдыңғы бөлімді қараңыз). Шындығында, стандартты қатені түсіну үшін орталық шекті теоремаға сенудің қажеті жоқ. Стандартты қатені өлшеудің келесі тәсілін қарастырыңыз: 1. Популяциядан бірнеше жаңа үлгілерді алыңыз. 2. Әрбір жаңа үлгі үшін статистиканы есептеңіз (мысалы, орташа). 3. 2-қадамда есептелген статистиканың стандартты ауытқуын есептеңіз; оны стандартты қатені бағалау ретінде қолданыңыз. Іс жүзінде стандартты қателіктерді бағалау үшін жаңа үлгілерді алудың бұл әдісі әдетте мүмкін емес (және статистикалық тұрғыдан өте ысырапшыл). Бақытымызға орай, мүлдем жаңа үлгілерді алудың қажеті жоқ; оның орнына қайта жүктеу үлгілерін қолдануға болады (қараңыз: разд. "Жүктеу" бұдан әрі осы тарауда). Қазіргі статистикада жүктеу стандартты қатені бағалаудың әдеттегі әдісіне айналды. Бұл әдісті іс жүзінде кез-келген Статистика үшін қолдануға болады және Орталық шекті теоремаға немесе бөлу сипаты туралы басқа болжамдарға сүйенбейді.

Жүктеу

Статистикалық немесе модельдік параметрлердің іріктемелі үлестірімін бағалаудың қарапайым және тиімді тәсілдерінің бірі - іріктеменің өзінен қайтарумен қосымша іріктемелерді алу және әрбір қайта іріктеу үшін статистиканы немесе модельді қайта есептеу. Бұл процедура жүктеу деп аталады (ағылшын тілінен. bootstrap-жылжыту, өзін-өзі баптау) және ол деректерді қалыпты бөлу немесе таңдаулы статистика туралы болжамдармен байланысты емес.

Негізгі терминдер

Bootstrap үлгісі (bootstrap үлгісі) бақыланатын деректер жиынтығынан қайтару арқылы алынған үлгі. Синонимі: жүктеу үлгісі.

Қайталанатын іріктеу (resampling) бақыланатын деректерден бірнеше рет іріктеу процесі; жүктеу және ауыстыру (араластыру) рәсімдерін қамтиды. Синонимдер: қайта таңдау, қайта өңдеу.

Жүктеу процесі бастапқы үлгіні мыңдаған немесе миллиондаған рет қайталау ретінде тұжырымдамалық түрде ұсынылуы мүмкін, бұл барлық білімді бастапқы үлгіге негізделген гипотетикалық популяцияны алу үшін (бұл жай ғана үлкен). Содан кейін үлгінің таралуын бағалау үшін осы гипотетикалық популяциядан үлгілерді алуға болады (сурет. 2.7).

Сурет. 2.7. Жүктеу идеясы

Іс жүзінде үлгіні бірнеше рет қайталаудың қажеті жоқ. Әр ойықтан кейін біз әр бақылауды артқа қайтарамыз; яғни таңдауды қайтару арқылы орындаймыз. Осылайша, біз шексіз популяцияны тиімді түрде жасаймыз, онда алынатын элементтің ықтималдығы кеннен кенге дейін өзгеріссіз қалады. N өлшемді іріктеу үшін орташа мәнді қайта таңдау алгоритмі келесідей болады: 1. Үлгі мәнін алып тастаңыз, оны жазып, кері қайтарыңыз. 2. N рет қайталаңыз. 3. N қайта тексерілген мәндердің орташа мәнін жазыңыз. 4. 1-3 R қадамдарын қайталаңыз. 5. R нәтижелерін пайдалану үшін: * олардың стандартты ауытқуын есептеңіз (ол стандартты үлгінің орташа қателігін бағалайды); * гистограмма немесе қорап диаграммасын құрыңыз; * сенімді аралықты табыңыз. Жүктеу процесінің R итерацияларының саны бірнеше түрде белгіленеді. Итерация неғұрлым көп болса, стандартты қатені немесе сенімділік аралығын бағалау дәлірек болады. Бұл процедураның нәтижесі - іріктеме статистикасының немесе бағалау модельдерінің параметрлерінің бутстра-повский жиынтығы, олардың қаншалықты өзгеретінін көру үшін әрі қарай тексеруге болады. R boot бағдарламалық пакеті осы қадамдарды бір функцияға біріктіреді. Мысалы, төмендегі мысалда жүктеу несие алған адамдардың кірістеріне қолданылады:

library(boot)

stat_fun <- function(x, idx) median(x[idx]) boot_obj <-

boot(loans_income, R = 1000, statistic=stat_fun)

Stat_fun функциясы IDX индексімен анықталған берілген үлгі үшін медианаңы есептейді. Нәтиже келесідей болады:

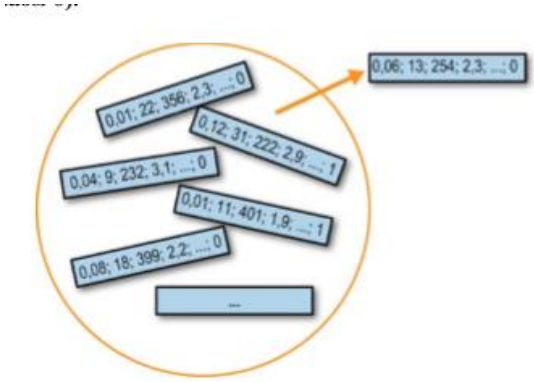
Bootstrap Statistics :

original bias std. error

t1 62000 -70.5595 209.1515*

Медианаңың бастапқы бағасы 62 мың долларды құрайды. Бутстраповтың асып кетуі бағалаудың -70 доллар және стандартты қателік 209 доллар екенін көрсетеді. Жүктеу жолын көп өлшемді деректермен қолдануға болады, онда жолдар бірлік ретінде таңдалады (сурет. 2.8). Содан кейін, жүктелген мәліметтерде модельді орындауға болады, мысалы,

модельдік жұп метрлердің тұрақтылығын (өзгергіштігін) бағалау немесе болжамды күшті жақсарту. Жіктеу және регрессия ағаштарына (шешім ағаштары деп аталады) келетін болсақ, бутстрапиялық үлгілерде көптеген ағаштарды орындау және олардың аңыздарын одан әрі орташалау (немесе жіктеу жағдайында көпшілік дауыспен шешім қабылдау, яғни көпшілік дауыспен), әдетте, бір ағашты қолданғаннан гөрі тиімді. Бұл процесс жүктеу агрегациясы немесе баггинг деп аталады (Bootstrap aggregating үшін қысқа: бөлімді қараңыз. "Баггинг және слу - шай орманы" 6-тарау).



Сурет. 2.8. Жүктеу үлгілерін көп өлшемді таңдау

Бутстрапиялық үлгілерді бірнеше рет таңдау тұжырымдамалық тұрғыдан ешқандай нәтиже бермейді, ал экономист және демограф Джулиан Саймон (Джулиан Саймон) өзінің 1969 жылғы "әлеуметтанудағы іргелі зерттеу әдістері" атты еңбегінде (Әлеуметтік ғылымдағы негізгі зерттеу әдістері, кездейсоқ үй) қайта іріктеу мысалдарының қысқаша мазмұнын жариялады. Алайда, бұл әдіс есептеу жағынан да үлкен және есептеу қуаты кең таралғанға дейін физикалық тұрғыдан мүмкін емес болып қала берді - менің мүмкіндігім. Ол Брэдли Эфронның (Bradley Efron) Стэндфорд статистикасы кітабын шығарғаннан кейін және 1970 жылдардың аяғы мен 1980 жылдардың басында журналдарда бірнеше мақалалар жариялағаннан кейін өз атын алды және танымал болды. және математикалық жуықтау оңай қол жетімді емес метрикалық көрсеткіштермен немесе модельдермен пайдалануға арналған. Орташа үлгіні бөлу 1908 жылдан бастап жақсы жұмыс істеді, бұл көптеген басқа метрикалық көрсеткіштердің селективті таралуына қатысты айту мүмкін емес еді. Жүктеу үлгісін үлгінің мөлшерін анықтау үшін, n-дің әртүрлі мәндерімен тәжірибе жасау үшін, олардың үлгінің таралуына қалай әсер ететінін түсіну үшін қолдануға болады. Бутстрапиялық сынамаларды іріктеуді қайталау әдісі алғаш рет енгізілген кезде, ол айтарлықтай скептицизммен қарсы алынды; көптеген адамдар үшін бұл сабанды алтынға айналдырудың орео - сынығымен байланысты болды. Бұл скептицизм жүктеу мақсатын түсінбеуден туындады.

Қайта іріктеу және жүктеу.

Кейде "қайта таңдау" термині жалпы түрде ұсынылған "жүктеу" терминінің синонимі ретінде қолданылады. Көбінесе "қайта таңдау" термині пермутация процедураларын да қамтиды (бөлімді қараңыз. "3-тараудың "пермутациялық тест"), онда көптеген үлгілер біріктіріліп, таңдау қайтарылмай жасалуы мүмкін. Қалай болғанда да, "жүктеу" термині әрқашан қайталанатын мәліметтер жиынтығынан іріктеуді білдіреді

Жүктеу үшін негізгі идеялар •

Жүктеу (қайтару арқылы мәліметтер жиынтығынан таңдау) - таңдаулы статистиканың өзгергіштігін анықтайтын қуатты құрал. •

Жүктемелерді әр түрлі жағдайларда бірдей қолдануға болады, олар таңдамалы үлестірулердің математикалық жуықтауын кең талдаусыз. •

Бұл әдіс сонымен қатар математикалық жуықтау әзірленбеген Статистика үшін үлгіні үлестіруді бағалауға мүмкіндік береді. •

Бұл әдіс болжамды модельдерге қолданылған кезде, жүктемелік үлгілерге (баггинг) негізделген көптеген болжамдарды біріктіру тиімділігі жағынан жалғыз модельден асып түседі.