

## Статистикалық эксперименттер және маңыздылығын тексеру.

Эксперименттерді жоспарлау іс жүзінде зерттеудің барлық салаларында қолданбалары бар практикалық статика стикасының негізі болып табылады. Мұндай жоспарлаудың мақсаты- Гипотезаны растайтын немесе қабылдамайтын эксперимент жасау. Деректер аналитиктері үздіксіз эксперименттер жүргізу қажеттілігіне тап болады, әсіресе пайдаланушы интерфейсі мен маркетингке қатысты. Бұл тарауда дәстүрлі эксперименттік жоспарлауға шолу жасалады және деректер ғылымындағы бірнеше жалпы міндеттер талқыланады. Ол сондай - ақ статистикалық тұжырымда жиі келтірілген кейбір ұғымдарды қарастырады және олардың мағынасы мен өзектілігі туралы түсініктеме береді (немесе деректер ғылымына қатысты болмауы).

Статистикалық маңыздылық, t-статистика немесе P-мәнге негізделген тексеру туралы айтылған сайын, бұл әдетте статистикалық қорытындының классикалық "құбыры" контекстінде болады (сурет. 3.1). Бұл процесс гипотезадан басталады ("а препараты — қолданыстағы стандартты препараттан Жақсы", "а бағасы — қолданыстағы в бағасынан тиімдірек"). Эксперимент (Бұл А/В сынағы болуы мүмкін) сенімді нәтижелерді қамтамасыз ету үшін құрылған гипотезаны тексеруге арналған. Деректер жиналады және талданады, содан кейін қорытынды жасалады. "Қорытынды" термині үлкен процеске немесе популяцияға шекті деректер жиынтығымен байланысты эксперименттік нәтижелерді қолдану ниетін көрсетеді.

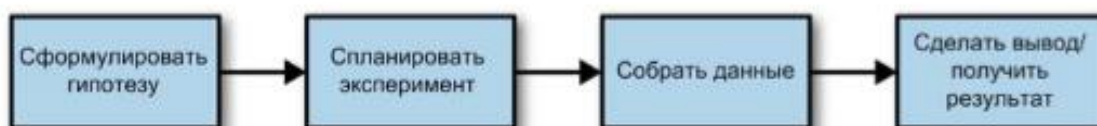


Рис. 3.1. Классический конвейер статистического вывода

### 3.1. - Сурет. Классикалық статистикалық Шығыс құбыры

#### А / В тестілеу

А / В сынағы-бұл өнімдердің, процедуралардың, емдеудің және басқа салыстырылатын объектілердің екі нұсқасының қайсысы жақсы екенін анықтауға арналған екі топтық эксперимент. Көбінесе екеуінің бір нұсқасы стандартты бар нұсқа болып табылады немесе мүлдем жоқ. Егер стандартты нұсқа қолданылса (немесе ол жоқ болса), онда ол бақылау деп аталады. Әдеттегі гипотеза-ұсынылған нұсқа бақылаудан гөрі жақсы.

## ***Негізгі терминдер***

*Эксперимент нұсқасы (treatment) дәрі, баға, веб-тақырып, яғни тестілеу мақсатында субъектіге ұсынылатын нәрсе.*

*Сынақ тобы (treatment group) белгілі бір нұсқа ұсынылатын субъектілер тобы.*

*Бақылау тобы (бақылау тобы) стандартты емес опция ұсынылатын немесе ұсынылмайтын субъектілер тобы.*

*Рандомизация (randomization) субъектілерді нұсқаларға ерікті түрде тағайындау процесі.*

*Субъектілер (субъектілер) опциялар ұсынылатын субъектілер (веб-сайтқа кірушілер, пациенттер және т.б.).*

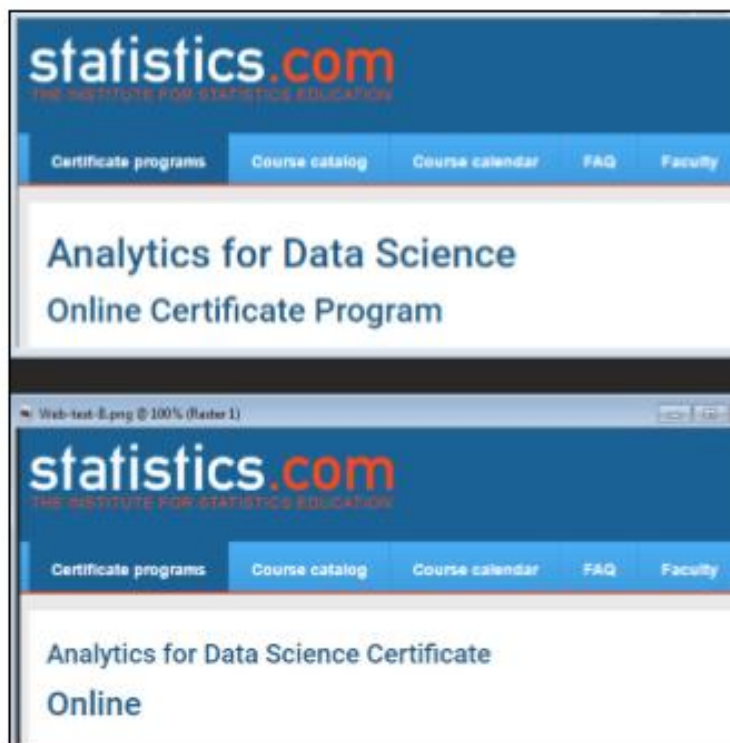
*Тексеру статистикасы (test statistic) нұсқа шарттарының әсерін өлшеу үшін қолданылатын метрикалық көрсеткіш.*

А / В тестілері веб-дизайн мен маркетингте кеңінен қабылданады, өйткені олардың нәтижелері өте оңай өлшенеді. А / В тестілеуінің кейбір мысалдарына мыналар жатады:

- қайсысы көшеттердің жақсы өнуіне әкелетінін анықтау үшін топырақты өңдеудің екі әдісін сынау;
- қайсысы тиімді екенін анықтау үшін екі емдеу әдісін сынау - бұл қатерлі ісік ауруын басады;
- екі бағаны тестілеу, қайсысы таза пайда әкелетінін анықтауға мүмкіндік береді;
- қайсысы көбірек түртуді тудыратынын анықтау үшін екі веб-тақырыпты сынау (сурет. 3.2);
- екі веб-жарнаманы тестілеу, қайсысы көбірек түрлендіруді тудыратынын анықтауға мүмкіндік береді.

Тиісті А / В тестінде белгілі бір нұсқаға жатқызылуы мүмкін субъектілер бар. Субъект адам, көшет, веб - сайтқа кіруші болуы мүмкін; ең бастысы, субъектіге эксперимент нұсқасы ұсынылады. Ең дұрысы, субъектілер ұсынылған екі нұсқа бойынша рандомизацияланады (ерікті түрде тағайындалады). Осылайша, сіз екі оқиғаның біріне байланысты сынақ топтары арасындағы кез келген айырмашылық байқалатынын білесіз:

- әр түрлі нұсқалардың шарттарының әсері;
- субъектілер нұсқаларға тағайындалатын таза кездейсоқтық (яғни, кездейсоқ тағайындау тиімдірек сынақ субъектілерінің табиғи түрде А тобында немесе ре В тобында шоғырлануына әкелуі мүмкін).



3.2. - Сурет. Маркетологтар веб-презентацияларды бір-бірімен салыстыру арқылы үнемі тексеріп отырады

Сондай — ақ, тексеру статистикасына назар аудару керек - А тобын в тобымен салыстыру үшін қолданылатын метрикалық көрсеткіш. ДДҰ мүмкін, деректер ғылымындағы ең көп қабылданған метрикалық көрсеткіш екілік айнымалы болып табылады: басу немесе басу, сатып алу немесе сатып алудан бас тарту, алаяқтық немесе алаяқтық емес және т. б. бұл нәтижелер жалпыланады  $22 \times$  өлшемді кестеде . Кестеде. 3.1 нақты бағаны тестілеу нәтижелерімен  $22 \times$  кесте берілген.

Кесте 3.1. Электрондық коммерцияны тестілеу нәтижелері бар  $2 \times 2$  кесте

Исход	Цена А	Цена В
Конверсия	200	182
Нет конверсии	23 539	22 406

Егер метрикалық көрсеткіш үздіксіз шамамен (сатып алу сомасы, пайда және т.б.) немесе санмен (мысалы, стационардағы күндер, кірген беттер) ұсынылса, онда нәтиже әртүрлі жолдармен көрсетілуі мүмкін. Егер конверсия қызықтырмаса, бірақ парақты бір рет қарауға есептегенде кіріс болса, онда кестедегі баға тестінің нәтижелері. 3.1 стандартты әдепкі бағдарламалық жасақтама жүйесінің шығысына ұқсас болуы мүмкін:

Табыс/бағамен қарау А: орташа = 3.87, СА = 51.10

кіріс / бағамен қарау В: орташа = 4.11, СА = 62.98

"СА" әр топ ішіндегі мәндердің стандартты ауытқуын білдіреді.

### Неліктен тек А / В? Неге С, D емес?

А / В сынақтары маркетинг және электрондық коммерция әлемінде танымал, бірақ бұл статистикалық эксперименттің жалғыз түрінен алыс. Қосымша опциялар қатысуы мүмкін. Субъектілер қайта өлшенетін зерттеулерге ұшырауы мүмкін. Жетіспейтін, қымбат және ұзақ уақыт қатысатын фармацевтикалық сынақтар кейде экспериментті тоқтатудың және қорытынды қорытынды алудың көптеген мүмкіндіктерімен жасалады. Статистикалық экспериментті дәстүрлі жоспарлауда көрсетілген нұсқалардың тиімділігі туралы статикалық сұраққа жауап беруге баса назар аударылады. Деректер талдаушылары " а бағасы мен В бағасы арасындағы айырмашылық статистикалық маңызды ма?" деген сұраққа қарағанда: " мүмкін болатын көптеген бағалардың қайсысы жақсы?" Ол үшін эксперименттік жоспардың салыстырмалы түрде жаңа түрі қолданылады: көп қарулы қарақшы (бөлімді қараңыз. "Көп қарулы қарақшы алгоритмі" осы тарауда).

### А / В тестілеудің негізгі идеялары

- Субъектілер екі (немесе одан да көп) топқа бөлінеді, олар қатаң түрде бірдей өңделеді, тек нұсқалар бір-бірінен ерекшеленеді.
- Ең дұрысы, субъектілер топтарға еркін тәртіпте бөлінеді.

## **Статистикалық гипотезаларды тексеру**

Статистикалық гипотезаны тексеру, маңыздылығын тексеру деп аталатын, жарияланған зерттеулерде кездесетін дәстүрлі статистикалық талдауда кеңінен қолданылды. Мұндай тексерулердің мақсаты-байқалған әсердің кездейсоқ мүмкіндіктен туындауы мүмкін екенін анықтау.

### ***Негізгі терминдер***

*Нөлдік гипотеза (null hypothesis) кездейсоқтық кінәлі деген Гипотеза.*

*Балама гипотеза (alternative hypothesis) кері болып табылады нөл (сіз дәлелдеуге үміттенетін нәрсе).*

*Бір жақты тексеру (one-way test) кездейсоқ нәтижелер тек бір бағытта есептелетін гипотезаны тексеру нұсқасы.*

*Екі жақты тексеру (two-way test) кездейсоқ нәтижелер екі бағытта есептелетін гипотезаны тексеру нұсқасы.*

А / В сынағы (бөлімді қараңыз. "Осы тараудың басында" А/В тестілеу") әдетте гипотезаны ескере отырып жасалады. Мысалы, гипотеза В бағасы жоғары пайда әкеледі деген болжам болуы мүмкін. Неліктен гипотеза қажет? Неліктен эксперименттің нәтижесіне қарап, жақсы жұмыс істейтін кез-келген нұсқаға тоқтала алмайсыз? Жауап адамның ақыл-ойының табиғи кездейсоқ мінез-құлықтың ауқымын бағаламауға бейімділігінде жатыр. Бұл бейімділіктің бір көрінісі-шекті оқиғаларды немесе "қара Аққулар" деп аталатын оқиғаларды болжай алмау (бөлімді қараңыз. 2-тараудағы" ұзын құйрықты үлестірулер"). Оның тағы бір көрінісі-кездейсоқ оқиғаларды қандай да бір маңыздылық белгілері бар деп қате түсіндіру үрдісі. Статистикалық гипотезаны тексеру зерттеушілерді кездейсоқ мүмкіндікке алданып қалудан қорғау әдісі ретінде ойлап табылды

### ***Кездейсоқтықпен қате түсіндіру.***

*Адамдардың кездейсоқтықты бағаламауға бейімділігін келесі экспериментте байқауға болады. Бірнеше достарыңыздан монетаның 50 есе лақтырылуын елестетуін сұраңыз: олар кездейсоқ О (Бүркіт) және Р (құйрық) сериясын жазсын. Содан кейін олардан монетаны 50 рет лақтырып, нәтижелерін жазуын сұраңыз. Олар нақты лақтырудың нәтижелерін орналастырсын монета-сен бір тізімге және ойдан шығарылған нәтижелер-екіншісіне. Қандай НӘТИЖЕЛЕР нақты екенін ажырату оңай: нақты нәтижелер О немесе Р-ден тұратын ұзағырақ болады. 50 нақты*

*лақтырулар жиынтығында қатарынан бес-алты О немесе R-ді көру ерекше емес. Алайда, көпшілігіміз монетаның кездейсоқ лақтырылуын елестетіп, қатарынан үш-төртеуін алсақ, біз өзімізге айтамыз: серия кездейсоқ болып көрінуі үшін, біз P-ға ауысқанымыз жөн. Бұл монетаның екінші жағы, егер осылай айтуға болатын болса, онда біз алтыдан тұратын нақты эквивалентті көргенде қатарынан (мысалы, жарнаманың бір тақырыбы екіншісінің өнімділігі бойынша 10%-дан жоғары болған кезде), біз оны жай гана жасырын емес, нақты нәрсеге жатқызуға бейімбіз.*

Тиісті түрде жасалған А / В тестінде Сіз А және В нұсқалары туралы деректерді А мен В арасындағы кез - келген байқалған айырмашылық екеуінің біреуіне байланысты болуы керек етіп жинайсыз:

- субъектілерді топтарға тағайындаудағы кездейсоқ мүмкіндік;
- гипотезаларды статистикалық тексеру-бұл а/в сынағын немесе кез - келген рандомизацияланған экспериментті одан әрі талдау, егер бұл шай мүмкіндігі А мен В тобы арасындағы байқалған айырмашылықты ақылға қонымды түсіндіру болса.

### **Нөлдік гипотеза**

Гипотезаны тексеруде келесі логика қолданылады: "адам дәуірінің әдеттен тыс, бірақ кездейсоқ мінез - құлыққа жауап беруге және оны мағыналы және нақты нәрсе ретінде түсіндіруге бейімділігін ескере отырып, біздің эксперименттерімізде топтар арасындағы айырмашылық кездейсоқтықты негізді түрде тудыратыннан гөрі шекті екенін дәлелдеу қажет болады". Бұл логика опциялар эквивалентті және топтар арасындағы кез-келген айырмашылық кездейсоқ деген негізгі болжаммен байланысты. Бұл негізгі болжам нөлдік гипотеза деп аталады. Содан кейін біздің үмітіміз - біз нөлдік гипотезаның дұрыс доказстігін дәлелдей аламыз және А және В топтарының нәтижелері кездейсоқтықты тудыруы мүмкін нәрседен гөрі ерекшеленетінін көрсете аламыз. Мұны істеудің бір жолы қайта іріктеуге негізделген араластыру процедурасы арқылы өтеді, онда біз А және В топтарының нәтижелерін араластырамыз, содан кейін деректерді ұқсас өлшемдегі топтарға бірнеше рет таратамыз, содан кейін сағатты бақылаймыз. - содан кейін біз байқалған айырмашылық сияқты шекті айырмашылықты аламыз. Қосымша мәліметтер алу үшін бөлімді қараңыз. Осы тарауда" қайта таңдау".

### **Балама гипотеза**

Гипотезаларды олардың табиғаты бойынша тексеру тек нөлдік гипотезамен ғана емес, сонымен қатар оның орнын толтыратын балама гипотезамен де байланысты. Мұнда бірнеше мысалдар келтірілген:

- нөлдік гипотеза - "А тобы мен В тобындағы орташа мәндер арасында айырмашылық жоқ", Альтернативті гипотеза - " А В-дан өзгеше "(көп немесе аз болуы мүмкін);
- нөлдік гипотеза — " $A \leq B$ ", балама - " $B > A$ ";
- нөлдік гипотеза — "В А — дан X% - ға артық емес", балама - "В А-дан X% - ға артық".

Бірге алынған нөлдік және альтернативті гипотезалар барлық қол жетімді мүмкіндіктерді қамтиды. Нөлдік гипотезаның табиғаты гипотезаны тексеру құрылымын анықтайды.

### **Бір жақты және екі жақты гипотезаны тексеру.**

Көбінесе А / В тестінде сіз жаңа мүмкіндікті (айталық, В) әдепкі параметрге (А) қарсы тексересіз және егер жаңа мүмкіндік сөзсіз жақсы болмаса, әдепкі мүмкіндікті ұстанатындығыңызды көрсетесіз. Бұл жағдайда гипотезаны тексерген кезде сіз кездейсоқ жағдайға алданып қалмас үшін өзіңізді қорғауыңыз керек В. сіз басқа бағытта кездейсоқтыққа алданып қалудың қажеті жоқ, өйткені егер В сөзсіз жақсы болмаса, А жағында тоқтаңыз. Сондықтан сізге бағытталған балама гипотеза қажет (В жақсы А). Бұл жағдайда сіз гипотезаны бір жақты (яғни бір құйрықты) тексеруді қолданасыз. Бұл дегеніміз, алдын-ала мүмкіндік тек р - білімге қарай бір бағытты санауға әкеледі. Егер сіз гипотезаны тексеруді кез - келген бағытта кездейсоқ алдаудан қорғағыңыз келсе, онда альтернативті гипотеза екі бағытты болады (а в-дан шығады және көп немесе аз болуы мүмкін). Бұл жағдайда сіз екі жақты (немесе екі құйрықты) гипотезаны қолданасыз. Бұл шекті мүмкіндік р - мәніне қарай екі бағытты санауға әкелетінін білдіреді. Бір құйрықты гипотезаны тексеру көбінесе А / В тестілеуінде шешім қабылдау сипатына сәйкес келеді, онда шешім қабылдау міндетті болып табылады және бір мүмкіндік әдетте "әдепкі" күйге беріледі, егер екіншісі жақсы нәтиже бермесе. Дегенмен, бағдарламалық қамтамасыз ету жүйелері, соның ішінде R, әдетте, әдепкі бойынша екі жақты тексеруді (екі құйрықты) қамтамасыз етеді және көптеген статистика мамандары дау - дамайды болдырмау үшін консервіленген екі жақты тексеруді қалайды. Бір құйрықты немесе екі құйрықты тексерулер арасындағы жеке тұлғаның тақырыбы өте түсініксіз және р-мәнін есептеу дәлдігі өте маңызды емес деректер ғылымына тікелей қатысы жоқ.

## ***Статистикалық гипотезаларды тексеруге арналған негізгі идеялар***

- *Нөлдік гипотеза - бұл ерекше ештеңе болған жоқ деген тұжырымды қамтитын логикалық құрылым және сіз байқаған кез-келген әсер кездейсоқ мүмкіндікке байланысты болады.*
- *Статистикалық гипотезаны тексеру нөлдік гипотезаның ақиқат екенін, "нөлдік модельді" (ықтималдық моделін) құратынын және сіз көріп отырған әсердің осы модельдің ақылға қонымды нәтижесі екендігіне сенетіндігін болжайды.*

## **Қайта іріктеу**

Статистикадағы қайта іріктеу статистикалық шамадағы кездейсоқ вариацияны анықтаудың жалпы мақсатымен бақыланатын деректерден іріктеме мәндерін бірнеше рет алуды білдіреді. Бұл әдісті кейбір машиналық оқытылатын модельдерді диагностикалау және дәлдігін жақсарту үшін де қолдануға болады (мысалы, бірнеше рет жүктелген деректерге негізделген шешім ағаштарына негізделген модельдердегі болжамдар, олар деп аталатын процестің нәтижесінде орташалануы мүмкін. "Бэггинг және кездейсоқ орман" б-тарау). Қайта таңдау процедураларының екі негізгі түрі бар: жүктеу және ауыстыру. Bootstrap бағалаудың сенімділігін анықтау үшін қолданылады; бұл әдіс алдыңғы тарауда талқыланды (бөлімді қараңыз. "Жүктеу" 2-тарау). Ауыстыру сынақтары гипотезаларды тексеру үшін қолданылады және әдетте екі немесе одан да көп топтармен біріктіріледі және біз оларды осы бөлімде талқылаймыз.

## ***Негізгі терминдер***

*Ауыстыру сынағы (пермутация сынағы) екі немесе одан да көп үлгілерді біріктіру процедурасы және бақылауларды қайталама үлгілерге ерікті (немесе толық) қайта бөлу. Синонимдер: рандомизация сынағы, ерікті ауыстыру сынағы, дәл сынақ.*

*Қайтарумен немесе қайтарусыз (with  $W$  without replacement) элемент келесі ойыққа дейін қайтарылған немесе қайтарылмаған кезде элементтерді таңдау опциясы.*

## **Ауыстыру сынағы (тесті)**

Ауыстыру процедурасы екі немесе одан да көп үлгілерді қамтиды, әдетте A/B тестіндегі немесе басқа гипотезаны тексерудегі топтар. Ауыстыру дегеніміз-мәндердің ретін өзгерту немесе олардың пермутациясы. Гипотезаны тоқтата тұрудың алғашқы қадамы A және B топтарының (және C, D, топтарының)



нәтижелерін біріктіру болып табылады .... егер олар қолданылса). Бұл нәлдік гипотезаның логикалық көрінісі-топтарға ұсынылған нұсқалар әр түрлі емес. Содан кейін біз бұл гипотезаны осы біріктірілген жиынтықтан топтарды ерікті түрде алу арқылы тексереміз және олардың бір-бірінен қаншалықты ерекшеленетінін көреміз. Ауыстыру процедурасы келесідей:

1. Әр түрлі топтардың нәтижелерін бір деректер жиынтығына біріктіріңіз.
2. Біріктірілген деректерді араластырыңыз, содан кейін А тобымен бірдей өлшемдегі қайта іріктеуді (қайтарусыз) кездейсоқ ретпен алыңыз.
3. Қалған деректерден в тобымен бірдей мөлшердегі қайталама үлгіні (қайтарусыз) ерікті тәртіппен алып тастаңыз.
4. С, D және т. б. топтары үшін де солай жасаңыз.
5. Бастапқы үлгілер үшін есептелген статистикаға немесе бағалауға байланысты (мысалы, топтардың үлестік айырмашылығы) енді оны қайта іріктеу үшін есептеп, жазыңыз; бұл ауыстырудың бір итерациясы болады.
6. Тексеру статистикасын қайта бөлу үшін алдыңғы R қадамдарын бір рет қайталаңыз.

Енді топтар арасындағы байқалған айырмашылыққа оралып, оны ауыстыру айырмашылықтарының жиынтығымен салыстырыңыз. Егер байқалған айырмашылық пермутациялық айырмашылықтар жиынтығының шегінде сенімді түрде жатса, онда біз ештеңе дәлелдеген жоқпыз — байқалған айырмашылық кездейсоқтықты тудыруы мүмкін диапазонда. Біреуі, егер байқалған айырмашылық қайта бөлу үлестірімінің көп бөлігінен тыс болса, онда біз кездейсоқтық жауапты емес деген қорытындыға келеміз. Техникалық тілмен айтқанда, айырмашылық статистикалық маңызды (бөлімді қараңыз. "Статистикалық маңыздылығы және р-мәндері" осы тарауда).

### **Мысал: веб-беттердің жабысқақтығы**

Салыстырмалы түрде қымбат қызметті сататын Компания екі веб-презентацияның қайсысы сатылымда жақсы жұмыс істейтінін тексергісі келеді. Сатылатын қызметтің қымбаттығына байланысты сатылымдар сирек кездеседі және сату циклі ұзаққа созылады; алдын - ала қараудың қайсысы жоғары екенін білу үшін жеткілікті сатылымды жинақтау тым көп уақытты қажет етеді. Сондықтан компания қызметті сипаттайтын веб-сайттың егжей-

тегжейлі ішкі бетін қолдана отырып, нәтижелерді эрзац айнымалысы арқылы өлшеуге шешім қабылдайды.

Біздің компания үшін ықтимал эрзац айнымалыларының бірі - егжей-тегжейлі қону бетіндегі басу саны. Одан да жақсы — адамдар бетте қанша уақыт жүреді. Адамдардың назарын ұзақ уақытқа созатын веб-презентация (бет) көбірек сатылымға әкеледі деп сену орынды. Демек, а бетін в парағымен салыстырған кезде біздің метрикалық қазандық сеанстың орташа уақыты болады. Бұл арнайы мақсаттағы ішкі бет туралы, сондықтан ол көптеген келушілерді ала алмайды. Сондай-ақ, біз сеанс уақытын өлшейтін Google Analytics (GA) қызметі клиенттің параққа соңғы рет кіру сеансының уақытын өлшей алмайтынын атап өткен жөн. Бұл сеансты деректерден шығарудың орнына, GA оны 0 деп жазады, сондықтан деректер осы сеанстарды жою үшін нақтылауды қажет етеді. Нәтижесінде екі түрлі презентация үшін барлығы 36 сеанс пайда болады: А парағы үшін 21 сеанс және В парағы үшін 15 сеанс. ggplot бағдарламалық пакетін қолдана отырып, жұптастырылған қорап диаграммаларын қолдана отырып, сеанстардың уақытын көзбен салыстыруға болады:

```
ggplot(session_times, aes(x=Page, y=Time)) +  
geom_boxplot()
```

3.3 - суретте ұсынылған. қорап диаграммасы в беті А бетіне қарағанда ұзағырақ сеанстарға әкелетінін көрсетеді, әр топ үшін орташа мәндерді келесідей есептеуге болады:

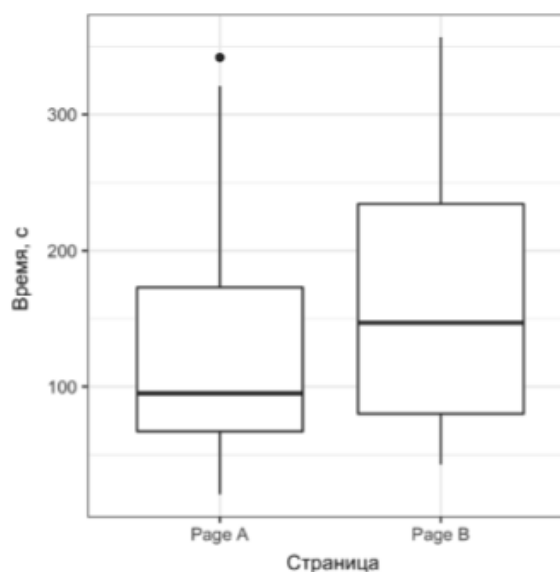
```
mean_a <- mean(session_times[session_times['Page']=='Page A', 'Time'])  
mean_b <- mean(session_times[session_times['Page']=='Page B', 'Time'])  
mean_b - mean_a
```

```
[1] 21.4
```

В бетінің сеанстарының уақыты А парағынан айырмашылығы орта есеппен 21,4 секундқа ұзағырақ, мәселе мынада: бұл айырмашылық кездейсоқ мүмкіндік тудыруы мүмкін ауқымның ішінде ме, әлде, керісінше, статистикалық маңызды ма. Бұл сұраққа жауап берудің бір жолы-ауыстыру тестін қолдану-барлық сеанс уақыттарын біріктіру, содан кейін оларды бірнеше рет араластырып, 21 элементтен тұратын топтарға бөлу (А беті үшін 21 n=) және 15 элементтен ( в үшін 15 n=).

Ауыстыру тестін қолдану үшін бізге 21 элементтен тұратын топқа (А беті) және 15 элементтен тұратын топқа (В беті) 36 сеанс уақытын ерікті түрде тағайындайтын функция қажет:

```
perm_fun <- function(x, n1, n2)
{
  n <- n1 + n2
  idx_b <- sample(1:n, n1)
  idx_a <- setdiff(1:n, idx_b)
  mean_diff <- mean(x[idx_b]) - mean(x[idx_a])
  return(mean_diff)
}
```



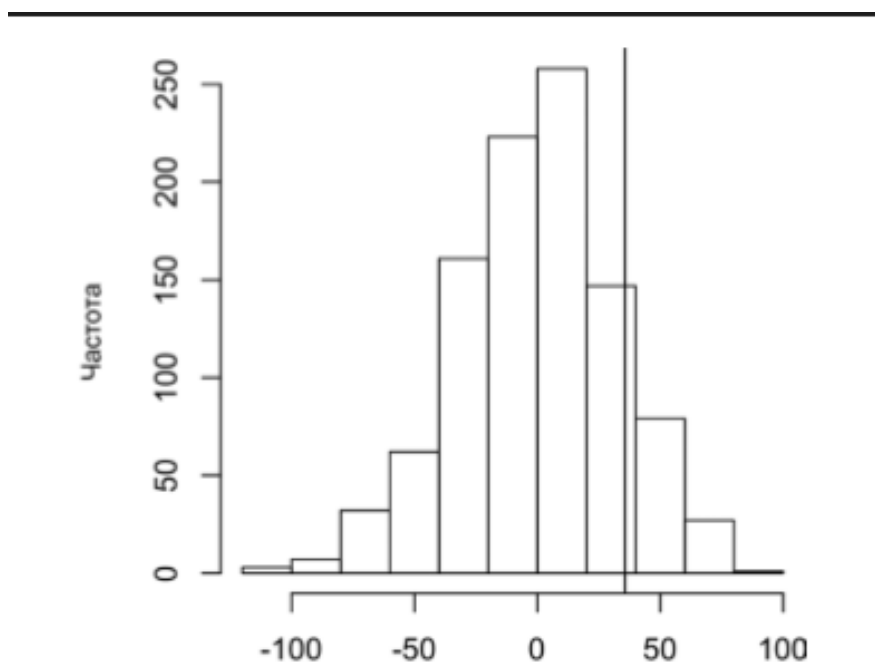
### 3.3. -Сурет. А және В веб-беттеріне арналған сеанс уақыттары

Бұл функция N2 индекстерін қайтарусыз таңдау және оларды В тобына жатқызу арқылы жұмыс істейді; қалған N1 индекстері А тобына тағайындалады. Бұл функцияны 1000 R = рет шақыру және N2=15 және N1 = 21 орнату қайта бөлуге әкеледі сеанс уақытының айырмашылығы, оны гистограмма түрінде көрсетуге болады.

```
perm_diffs <- rep(0, 1000)
for(i in 1:1000)
  perm_diffs[i] = perm_fun(session_times[, 'Time'], 21, 15)
hist(perm_diffs, xlab='Разницы во времени сессий, с')
abline(v = mean_b - mean_a)
```

3.4-суретте көрсетілген Гистограмма. Ерікті ауыстырулардың орташа айырмашылығы сеанс уақытындағы байқалған айырмашылықтан асып түсетінін көрсетеді (вертикальды сызық). Бұл А және В беттері арасындағы сеанстар уақытындағы байқалған айырмашылық кездейсоқ вариация

диапазонында алыс екенін көрсетеді, сондықтан статистикалық маңызды емес.



3.4. - сурет. Үшін жиіліктің таралуы сеанс уақытының айырмашылығы А және В беттері арасында

### **Толық және жүктелген ауыстыру сынақтары**

Алдыңғы ерікті араластыру процедурасынан басқа, менің ерікті ауыстыру сынағы немесе рандомизация сынағы деп те аталады, ауыстыруды тексерудің екі нұсқасы бар

- толық ауыстыру сынағы;
- Bootstrap ауыстыру сынағы.

Толық ауыстыру сынағында қарапайым ерікті шөміш пен деректерді бөлудің орнына, біз оларды бөлудің барлық мүмкін жолдарын анықтаймыз. Бұл үлгілердің салыстырмалы түрде шағын өлшемдері үшін ғана мүмкін. Қайта араластырудың көп мөлшерімен кездейсоқ қайта құру тестінің нәтижелері толық қайта құру тестінің нәтижелерін жуықтайды және оларға дейін жақындайды. Толық ауыстыру сынақтары кейде статистикалық қасиеттеріне байланысты дәл сынақтар деп аталады, бұл нөлдік модель сынақтағы  $\alpha$  маңыздылық деңгейінен "маңызды" ретінде тексерілмеуін қамтамасыз етеді

(бөлімді қараңыз. "Статистикалық маңыздылығы және р-мәндері" осы тарауда).

Жүктеу ауыстыру сынағында кездейсоқ ауыстыру сынағының 2 және 3-қадамдарында қысқаша сипатталған ойықтар қайтарусыз емес, қайтарумен жасалады. Осылайша, қайта іріктеу процедурасы субъектіні нұсқаға жатқызу кезінде кездейсоқ элементті ғана емес, сонымен қатар субъектілерді популяциядан кездейсоқ таңдауды модельдейді. Екі процедура да статистикада кездеседі және олардың арасындағы айырмашылық өте күрделі және деректер ғылымының тәжірибесінде маңызды емес.

### **Статистикалық маңыздылығы және р-мәндері.**

Статистикалық маңыздылық статистика саласындағы мамандардың статистикалық экспериментті (немесе бар деректердің кез - келген статистикалық зерттеуін) қалай өлшейтінін сипаттайды: эксперимент кездейсоқтықты тудыруы мүмкін нәтижеге қарағанда шекті нәтиже бере ме? Егер нәтиже кездейсоқ вариациядан тыс болса, онда бұл статистикалық маңызды деп айтылады.

### ***Негізгі терминдер***

*Р-мәні (Р-мәні) нөлдік гипотезаны қамтитын кездейсоқ модельді ескере отырып, р-мәні бақыланатын нәтижелер сияқты әдеттен тыс немесе шекті нәтижелерді алу ықтималдығы болып табылады.*

*Альфа ( $\alpha$ , Альфа) кездейсоқ НӘТИЖЕЛЕР нақты нәтижелерді статистикалық маңызды деп санау үшін асып түсетін "әдеттен тыс" Ықтималдық шегі.Синонимдер: маңыздылық деңгейі.*

*1-ші қателік (1 — ші қате) эффект нақты деген қате тұжырым(ол кездейсоқ болған кезде).*

*2-ші қателік (2 — ші қате) эффект кездейсоқ деген қате тұжырым (ол нақты болған кезде).*

3.2 -Кестеде қарастырайық. Бұрын көрсетілген веб-тест нәтижелері.

3.2. -Кесте Электрондық коммерция экспериментінің нәтижелері үшін  $2 \times 2$  кесте

Исход	Цена А	Цена В
Конверсия	200	182
Нет конверсии	23,539	22,406

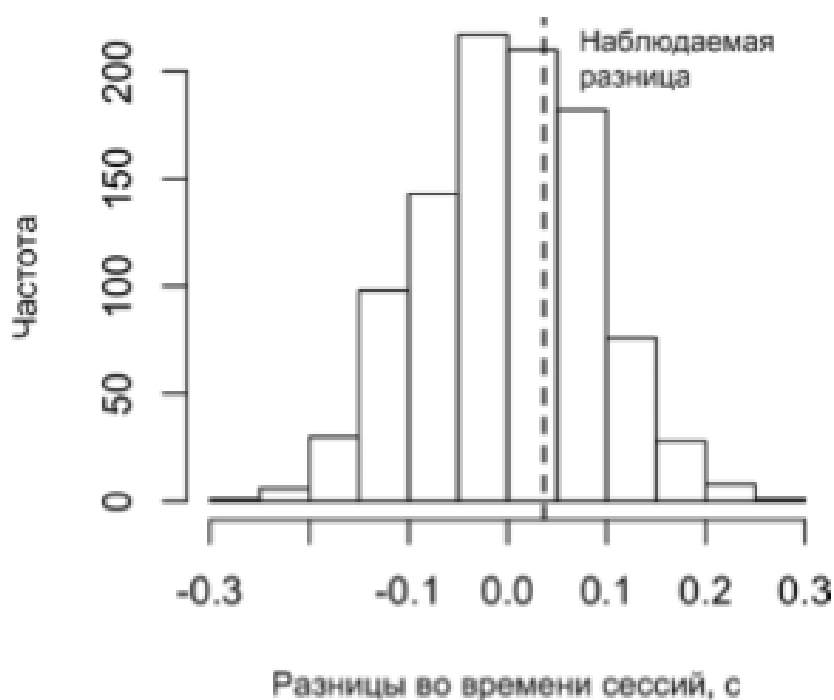
А бағасы В бағасынан 5% — ға жақсырақ (яғни келушілерді сатып алушыларға айналдырады) (0,8425% қарсы 0,8057% - айырмашылық 0,0368 пайыздық тармақ); бұл компанияның ауқымды бизнесте маңызды болуы үшін жеткілікті маңызды нәтиже. Мұнда 45000 - нан астам деректер нүктелері бар және оларды статистикалық маңыздылығын тексеруді қажет етпестен "үлкен деректер" ретінде қарастыру қызықтырады (негізінен шағын үлгілердегі таңдамалы вариацияны ескеру қажет). Алайда, конверсия деңгейі соншалықты төмен (1% - дан аз), нақты мән мәндері — конверсиялар — тек жүзден бір бөлігінде болады, ал қажетті іріктеу мөлшері осы конверсиялармен анықталады. Қайта таңдау процедурасын қолдана отырып, кездейсоқ вариация диапазонындағы А және В бағалары арасындағы айырбастау айырмашылығының бар - жоғын тексере аламыз. "Кездейсоқ вариация" деп біз конверсия деңгейлері арасында ешқандай айырмашылық жоқ деген нөлдік гипотезаны қамтитын ықтималдық моделінен туындаған кездейсоқ вариацияны айтамыз (бөлімді қараңыз. Осы тараудың басында" нөлдік гипотеза"). Келесі ауыстыру процедурасы сұрақ туғызады: егер осы екі бағаның айырбастау деңгейі бірдей болса, онда кездейсоқ вариация 5% айырмашылыққа әкелуі мүмкін бе?

1. Барлық таңдамалы нәтижелері бар қорап жасаңыз :ол 382 бірлік пен 45,945 нөлден тұратын қиялдағы бірлескен конверсия деңгейін білдіреді-0,008246 0,8246%=.
2. 23,739 өлшемді қайта іріктеуді араластырыңыз және алыңыз (N Саны а бағасымен бірдей) және бірліктер санын жазыңыз.
3. Қалған 22,588-де бірліктер санын жазыңыз (N Саны В бағасымен бірдей).
4. Бірліктердің үлесі түріндегі айырмашылықты жазыңыз.
5. 2-4 қадамдарды қайталаңыз.  $\geq 0,0368$  айырмашылығы қаншалықты жиі байқалды?

Бөлімде анықталған `perm_fun` функциясын қайтадан пайдалану. "Мысал: веб-беттердің жабысқақтығы" осы тараудың басында сіз конверсия

деңгейіндегі кездейсоқ өзгертілген айырмашылықтардың гистограммасын жасай аласыз:

```
obs_pct_diff <- 100*(200/23739 - 182/22588)
conversion <- c(rep(0, 45945), rep(1, 382))
perm_diffs <- rep(0, 1000)
for(i in 1:1000)
  perm_diffs[i] = 100*perm_fun(conversion, 23739, 22588)
hist(perm_diffs, xlab='Разницы во времени сессий, с')
abline(v = obs_pct_diff)
```



3.5. - сурет. А және В беттері арасындағы конверсия деңгейлерінің айырмашылығы үшін жиіліктің таралуы

Бір 3.5-суретте қайта сыналған 1000 нәтиженің гистограммасын қарастырайық. белгілі болғандай, бұл жағдайда байқалған 0,0368% айырмашылық кездейсоқ вариация диапазонында орналасқан.

### р-мәні

Қарапайым графикті тексеру статистикалық маңыздылықты өлшеудің өте дәл әдісі емес, сондықтан р мәні үлкен қызығушылық тудырады. Бұл кездейсоқ модель байқалған нәтижеге қарағанда шекті нәтижеге әкелетін жиілік. Ауыстыру сынағы байқалған айырмашылыққа тең немесе одан үлкен

айырмашылықты тудыратын уақыттың үлесін алу арқылы біздің ауыстыру сынағымыздан  $p$  мәнін бағалай аламыз:

```
mean(perm_diffs > obs_pct_diff)
```

```
[1] 0.308
```

$p$ -мәні 0,308 құрайды, яғни біз уақыттың 30% - дан астамы осы сияқты шекті нәтижеге немесе кездейсоқ мүмкіндіктің арқасында шекті мәнге жетеді деп күтеміз. Бұл жағдайда  $p$  мәнін алу үшін ауыстыру тестін қолданудың қажеті жоқ еді. Биномдық үлестіру болғандықтан, біз қалыпты үлестіру арқылы  $p$  мәнін жуықтай аламыз. R кодының үзіндісінде біз мұны функция арқылы жасаймыз

```
prop.test:
```

```
> prop.test(x=c(200,182), n=c(23739,22588), alternative="greater")
```

```
2-sample test for equality of proportions with continuity correction
```

```
data: c(200, 182) out of c(23739, 22588)
```

```
X-squared = 0.14893, df = 1, p-value = 0.3498
```

```
alternative hypothesis: greater
```

```
95 percent confidence interval:
```

```
-0.001057439 1.000000000
```

```
sample estimates:
```

```
prop 1 prop 2
```

```
0.008424955 0.008057376
```

$X$  аргументі - әр топ үшін сәттілік саны,  $n$  аргументі-сынақ саны. Қалыпты жуықтау 0,3498  $p$  мәнін береді, бұл ауыстыру сынағынан алынған  $p$  мәніне жақын.

### **R мәні қандай?**

Соңғы жылдары  $p$ -мәнін қолдану белсенді полемикамен қоршалған. Психология журналдарының бірі оған берілген мақалаларда  $p$  - мәндерді қолдануға " тыйым салғаны " соншалық, тек  $p$ -мәніне негізделген жариялау шешімдері сапасыз зерттеу жұмыстарының жариялануына әкелді.  $P$ -мәнінің шын мәнінде нені білдіретінін анық көрсетпейтін тым көп зерттеушілер деректерге және тексерілетін әртүрлі ықтимал гипотезалар арасында тиісті  $p$ -білімге, демек, жариялауға қолайлы жұмысқа әкелетін комбинацияны тапқанша қазып алады.

Нақты мәселе-адамдар  $p$ -мәнінен ауырсыну мағынасын алғысы келеді. Міне, біз  $p$ -мәнінде не болғанын қалаймыз:



Нәтиженің кездейсоқ болу ықтималдығы.

Біз төмен мәнге үміт білдіреміз, сондықтан бір нәрсе дәлелденді деп қорытынды жасай аламыз. Көптеген журнал редакторлары  $p$ -мәнін осылай түсіндіреді. Бірақ міне,  $p$  мәні нені білдіреді:

Кездейсоқ модельді ескере отырып, бақыланатын нәтижелермен бірдей алдын - ала нәтижелерге қол жеткізу ықтималдығы.

Айырмашылық нәзік, бірақ нақты. Мағыналы  $p$  мәні сізді "дәлелдеуге" апармайды, өйткені ол уәде еткендей. "Статистикалық маңыздылық" туралы қорытынды жасаудың логикалық негізі  $p$  - мәнінің нақты мағынасын түсінген кезде біршама әлсірейді.

2016 жылдың наурызында американдық статистикалық қауымдастық (American Statistical Association, ASA) ұзақ ішкі пікірталастардан кейін  $p$  мәндеріне қатысты түсінбеушілік дәрежесін көрсетті, өйткені ол оларды пайдалану туралы ескерту шығарды.

ASA мәлімдемесінде зерттеушілер мен журнал редакторлары үшін алты қағидаға баса назар аударылды.

1.  $p$ -мәндері деректердің берілген статистикалық модельге қаншалықты сәйкес келмейтінін көрсетуі мүмкін.
2.  $p$  мәндері зерттелетін гипотезаның ақиқат болу ықтималдығын немесе деректердің тек шай мүмкіндігіне байланысты пайда болу ықтималдығын өлшемейді.
3. Ғылыми тұжырымдар мен бизнес немесе стратегиялық шешімдер тек  $p$  мәнінің белгілі бір шекті мәннен өтуіне негізделмеуі керек.
4. Тиісті қорытынды толық есеп пен ашықтықты қажет етеді.
5.  $p$ -мәні немесе статистикалық маңыздылығы әсер өлшемін немесе нәтиженің маңыздылығын өлшемейді.
6. Осылайша,  $p$  мәні модельге немесе гипотезаға қатысты жақсы дәлелдеу өлшемін қамтамасыз етпейді.

### **1-ші және 2-ші қателер.**

Статистикалық маңыздылықты анықтауда қателіктердің екі түрі болуы мүмкін:

1 - ші қателік, егер сіз эффект нақты деп қате тұжырымдасаңыз, ал іс жүзінде ол тек кездейсоқ болады;

2-ші қателік, егер сіз эффект нақты емес (яғни кездейсоқ) деген қате тұжырым жасасаңыз, ол шын мәнінде нақты. Нысандағы 2-ші қателік-бұл қате емес, бірақ үлгінің өлшемі әсерді анықтау үшін тым кішкентай деген пікір. Р мәні статистикалық маңыздылықтан алыс болған кезде (мысалы, ол 5% - дан асады), біз "әсер дәлелденбеген" деп айтқымыз келеді. Үлкен үлгі р мәнін төмендетуі мүмкін. Маңыздылықты тексерудің (немесе гипотезаны тексеру деп аталатын) негізгі функциясы — экспериментатордың кездейсоқ мүмкіндікке алданып қалуынан қорғау; демек, олар әдетте 1 - ші типтегі қателерді азайту үшін жасалады.

### **Деректер туралы ғылым және Р-мәндер**

Деректерді талдаушылар жүргізетін жұмыс, әдетте, ғылыми журналдарда жариялауға арналмаған, сондықтан р-мағынасы туралы пікірталас академиялық емес. Деректерді талдау үшін р мәні-бұл модельдік нәтиженің қызықты және пайдалы болып көрінетін қалыпты кездейсоқ өзгергіштік диапазонында жатқанын білгіңіз келетін жағдайларда пайдалы метрикалық көрсеткіш. Экспериментте шешім қабылдау құралы ретінде р мәні шешуші көрсеткіш емес, шешімге әсер ететін ақпараттың тағы бір нүктесі ретінде қарастырылуы керек. Мысалы, р мәндері кейде кейбір статистикалық немесе машиналық оқытылатын модельдерге аралық кірістер ретінде пайдаланылады - белгі оның р мәніне байланысты модельге қосылуы немесе алынып тасталуы мүмкін.

### *Статистикалық маңыздылық пен р-мәнге арналған негізгі идеялар*

- *Маңыздылықты тексеру бақыланатын әсердің нөлдік гипотеза моделі үшін кездейсоқ вариация диапазонында жатқанын анықтау үшін қолданылады.*
- *р мәні - бұл нөлдік гипотезаны ескере отырып, бақыланатын нәтижелер болуы мүмкін шекті нәтижелерге әкелетін ықтималдық.*
- *А мәні немесе маңыздылық деңгейі - бұл нөлдік гипотезаның кездейсоқ моделіндегі "ерекше" шегі.*
- *Маңыздылықты тексеру деректер ғылымына қарағанда (бірақ маңыздылығы соңғы кездері тіпті біріншісі үшін де жоғала бастаған) ресми зерттеу есептері үшін әлдеқайда маңызды болып қала береді.*

### **Т-статистика негізінде тексеру**

Маңыздылықты тексерудің көптеген түрлері бар, олар деректердің сандық немесе өлшенуіне, қанша Борок бар екеніне және не өлшенетініне байланысты. Ең көп қабылданған-студенттің  $t$ -үлестірімінің атымен аталған  $t$ -статистикалық тексеру (немесе  $t$  - тест), бастапқыда В. С. Госсет бір таңдаулы орташа үлестірімді жуықтау үшін әзірлеген

### ***Негізгі терминдер***

*Тексеру статистикасы (test statistic) мақсатты айырмашылықтың немесе әсердің метрикалық көрсеткіші. Синоним: критерий статистикасы.*

*$t$ -Статистика ( $t$ -statistic) тексеру статистикасының стандартталған нұсқасы.*

*$t$ -үлестіру ( $t$ -distribution) бақыланатын  $t$ -статистикасын салыстыруға болатын анықтамалық үлестіру (бұл жағдайда нөлдік гипотезадан алынған).*

Барлық маңыздылықты тексеру сізден тексеру статистикасын (отандық әдебиеттерде — критерий статистикасы) анықтауды талап етеді, ол сізді қызықтыратын әсерді өлшейді және бұл байқалған эффект қалыпты кездейсоқ вариация диапазонында екенін анықтауға көмектеседі. Қайта іріктеу тексеруінде (ауыстыруды талқылау үшін бөлімді қараңыз. Осы тараудың басында "ауыстыру тесті") деректер шкаласы маңызды емес. Сіз деректердің өзінен тікелей анықтамалық (нөлдік гипотеза) үлестіруді жасайсыз және дәлелді статистиканы сол күйінде қолданасыз. 1920-30 жылдары статистикалық гипотезаларды тексеру жүргізіліп жатқан кезде, сынамаларды қайта іріктеу негізінде тексеру жүргізу мақсатында деректерді бірнеше мың итерациямен ерікті түрде араластыру міндеті орындалмады. Статистика мамандары ауыстыру (араластыру) үлестірімінің жақсы жуықтауы Госсеттің  $t$ -үлестіріміне негізделген  $t$  - statistic негізіндегі тексеру екенін анықтады. Ол өте кең таралған екі үлгіні салыстыру үшін қолданылады —  $A/B$ -тест, онда деректер сандық болып табылады. Бірақ  $t$ -таратуды қандай да бір нақты деректер шкаласына байланыстырмай пайдалануға болатындай етіп, тексеру статистикасының стандартталған нысаны пайдаланылуы тиіс. Осы кезеңдегі классикалық статистикалық мәтінде олардың құрамында Госсеттің таралуы бар және сіздің деректеріңізді стандартты  $t$  - таратумен салыстыру мақсатында стандарттаудың қалай орындалатынын көрсететін әртүрлі формулалар көрсетілген болар еді. Бұл формулалар мұнда берілмейді, өйткені барлық статистикалық бағдарламалық

жүйелер, сондай-ақ R және Python, осы формулаларды жүзеге асыратын командаларды қамтиды. R де бұл функция

t.test:

```
> t.test(Time ~ Page, data=session_times, alternative='less')
```

Welch Two Sample t-test

data: Time by Page

t = -1.0983, df = 27.693, p-value = 0.1408

alternative hypothesis: true difference in means is less than 0

95 percent confidence interval:

-Inf 19.59674

sample estimates:

mean in group Page A mean in group Page B

126.3333          162.0000

Балама гипотеза-А парағының орташа сеанс уақыты в парағына қарағанда аз, бұл жаңа тесттің 0,124 р мәніне жақын (осы тараудың басында "мысал: веб-беттердің жабысқақтығы" бөлімін қараңыз). Қайта іріктеу режимінде біз деректерді ондық немесе екілік, үлгі өлшемдері теңдестірілген бе, жоқ па, үлгі дисперсиялары бар ма, жоқ па немесе басқа да көптеген факторлар туралы алаңдамай, таңдалған деректер мен тексерілетін гипотезаны көрсету үшін шешімді құрылымдаймыз. Формула әлемінде көптеген вариациялар өздерін бейнелейді және олар таңқаларлық болуы мүмкін. Статистика мамандары осы әлем бойынша саяхаттап, оның картасын зерттеуі керек, бірақ деректер талдаушыларына бұл қажет емес — олар, әдетте, ғылыми жұмысты жариялауға дайындайтын зерттеуші сияқты гипотеза тексерулерінен және сенімді аралықтардан мәліметтерді сығумен айналыспайды.

### ***T-статистика негізінде тексеруге арналған негізгі идеялар***

- *Компьютерлер пайда болғанға дейін қайта іріктеу негізінде тексерулер практикалық болмады, ал статистика мамандары стандарттық анықтамалық үлестірулерді қолданды.*

- *Тексеру статистикасы бұдан әрі стандартталуы және анықтамалық үлестіріммен салыстырылуы мүмкін.*
- *Осындай кеңінен қолданылатын стандартталған статистиканың бірі-t-статистика.*