§2. Представителни извадки. Описателни методи

1. Представителност на извадката. При всички ситуации в една или друга степен целта на статистическото изследване е да се направят изводи за състоянието на популацията според данните от извадката, което изисква извадката да отразява пропорционално основните черти на популацията. Такива извадки се наричат ***представителни***. В повечето случаи популацията е ясно дефинирана еднородна група по основен формален признак, което не изключва възможността вътре в популацията да се обособяват съществени еднородни групи по някакви други важни за изследването признаци. Например нека целта на изследването е характеризиране по някакви показатели на децата от средно училищна възраст в република България. В този случай популацията представлява напълно определена група деца на възраст между 11 и 15 години. Тази популация обаче съдържа обособени групи от момичета и момчета, както и обособени групи по териториален признак. Тук могат да се разгледат и други такива признаци. В този смисъл ***идентифицирането на популацията зависи много съществено от контекста на изследователската задача от етапа на нейното формулиране до етапа на интерпретация на резултатите***. Ако задачата е изследване отношението на горната популация към въвеждането на вечерен час, то е твърде правдоподобно да си мислим, че не съществуват съществени различия по пол и териториален признак. По този начин една представителна извадка от 300 души може да се образува на базата на случаен (непреднамерен) подбор от един достатъчно голям град. В статистиката понятията представителна и непреднамерена извадка са синоними. Непреднамереността означава, че изследователят не влага специална умисъл при съставянето на извадката.

Ако задачата се състои в изследване на евентуалните различия между момчетата и момичетата, то фактически става дума за сравняване характеристиките на две отделни популации посредством данните от две извадки – по една за популацията на момчетата и момичетата. В този случай и двете извадки трябва да бъдат представителни за своите популации.

Нека например целта на изследването е установяване отношението на населението на република България към Европейския Съюз. Тук физическата популация се състои от всички граждани в активна съзнателна възраст. Да предположим условно, че става дума за 6000000 лица. Понеже проблемът тук има преди всичко икономическа природа, редно е да разделим популацията на относително еднородни групи по икономически признак, които групи споделят приблизително едно и също отношение. Нека за простота и определеност да предположим, че въпросната група съдържа 1000000 заети в частния бизнес, 2000000 заети в държавната администрация и 3000000 наемни работници. В такъв случай една представителна извадка от 600 души трябва да съдържа 100 случайно подбрани лица от сферата на частния бизнес, 200 случайно подбрани лица от държавната администрация и 300 случайно подбрани лица – наемни работници. Истинското решаване всъщност на последната задача изисква точни данни за структурата на популацията, както и по-съдържателно структуриране.

В горния смисъл представителната извадка може да се схваща като умалено копие на популацията.

На практика липсата на явна преднамереност в дадена извадка се разглежда като достатъчен белег за нейната представителност и в много случаи това наистина е така. Както се вижда обаче от приведените примери, проблемът за представителността е по-сложен отколкото изглежда първоначално. Една извадка може да бъде представителна по отношение на дадена цел и да не бъде представителна по отношение на друга.

Основният поток на статистическите методи за анализ се отнася към физически популации с достатъчно голямо количество единици, които популации формално се интерпретират като безкрайни. Последното предположение осигурява възможността за избор на случайни независими извадки с произволно голям обем. В настоящият лекционен курс всяка дадена физическа популация ще разглеждаме като концептуално безкрайна, и следователно имаща съществено хипотетичен характер, което в крайна сметка не ограничава само по себе си общността на разсъжденията, понеже хипотетичният характер на популацията съществува фактически още при етапа на нейното обособяване. На практика безкрайността означава наличието на достатъчно много статистически единици във физическата популация, например ако съдържа повече от 10000 такива.

В много ситуации статистическите изследвания засягат специализирани групи от различно естество. Например групи военнослужещи, групи лица лишени от свобода, групи подрастващи, обучавани в специализирани заведения (в това число и елитарни) и т.н. В този случай популацията се характеризира със сравнително малко на брой физически представители, които могат да бъдат обхванати изцяло в дадено конкретно изследване, като по този начин извадката на практика съвпада с физическата популация. В такива случаи прилагането на статистически анализи, характерни за безкрайни популации не води до съществени промени в логиката на предметното разсъждение.

Всяка сравнително еднородна извадка е представителна за някаква популация. От такава гледна точка въпросът за представителността се състои в това, доколко тази популация е добре идентифицирана, което обикновено предхожда статистическия анализ и преди всичко доколко след като бъде идентифицирана тя изобщо представлява интерес за изследване. Очевидно всяко сериозно приложно изследване трябва да започва с подробно характеризиране на своя обект и своите цели, а не да се търси обект на приложения на моделите пост фактум.

Представителността на извадката в повечето случаи представлява единственият спорен пункт на дадено статистическо изследване.

2. Представяне на номинални величини. За номиналните величини взети сами по себе си е възможно да се преброят индивидите по отделните категории и резултатите да се запишат в таблица или диаграма.

Следващата диаграма 2.1 представя разпределението на индивидите по религиозна принадлежност за извадката от данните **USGSS93.**

Диаграма 2.1.

Ако направената извадка е наистина представителна, то горните пропорции носят информация за пропорциите на религиозна принадлежност за цялата популация, която в този случай представлява населението на САЩ, например можем да направим извода, че католиците съставляват около 22% от населението. Точността на направеният извод очевидно зависи от броя на наблюденията (обема на извадката). Колкото е по-голям този брой, толкова повече се увеличава сигурността на заключението.

Тези данни могат да бъдат подредени в таблица.

Таблица 2.1.

	брой	**процент**
протестанти	952	63.83
католици	333	22.30
евреи	31	2.07
атеисти	140	9.37
други	35	2.34
общо	1491	

Ако са дадени две номинални величини, можем да образуваме тяхната таблица на съвместно (взаимно) разпределение. Например величините религиозна принадлежност и брачен статус (взети от **USGSS93**), ограничени до определен брой категории, за които съответната част от извадката притежава достатъчно голям обем, имат следното взаимно разпределение.

Таблица 2.2.

	протестанти	**католици**	**общо**
женен(омъжена)	**498**	**193**	691
разведен(а)	**143**	**41**	184
никога не встъпвал(а) в брак	**157**	**66**	223
общо	798	300	1098

Ако приемем, че извадката е представителна по отношение на двете величини, то горната таблица съдържа ***цялата*** информация от опита, която има някакво отношение към статистическите закономерности в популацията, касаещи както индивидуалните характеристики така и взаимната статистическа зависимост на тези величини (с въведените изборни ограничения). Тук популацията е редуцирана до онази част от населението, която се идентифицира в посочените категории.

3. Представяне на метрични величини. Анализирането на метрични величини е от първостепенна важност в психологията, понеже най-добре структурираните и съдържателни величини пристигат като сурови балове от различни психологически тестове, за които са налице достатъчно предпоставки да бъдат определяни като метрични величини в интервална скала.

Да анализираме пример, който добре илюстрира характера на задачата в общия случай. Разглежданата величина X представлява суров бал от психологически тест за адаптация към средата на специализирана група. Тестът се състои от шест отделни субскали, които обхващат различните аспекти на адаптацията. Всяка от тези субскали се състои от известен брой тестови единици (***айтъми***). Конструкцията и основните характеристики на такива тестове ще бъдат обсъдени към края на настоящия лекционен курс. Извадката съдържа 175 лица. Най-малката наблюдавана стойност е $x_{\min} = 101$, а най-голямата $x_{\max} = 244$. Размахът (***range***) на извадката е $x_{\max} - x_{\min} = 143$, а нейният обем (***sample size***) е $n = 175$.

Статистиката се интересува не от стойностите на отделния индивид, а преди всичко от груповото поведение на извадката, което означава изследване на ***разпределението*** (***distribution***) на наблюдаваната величина. Разпределението е ***първично понятие*** в статистиката и се получава чрез групиране на наблюденията по подходящ начин, чрез разделяне числовата ос на наблюденията на определен брой (обикновено равни) интервали, за всеки от които се пресмята ***наблюдаваната честотата*** (***observed frequency***) – броят на наблюденията, попаднали в този интервал. При работа с компютър, броят на интервалите и тяхната дължина се определят автоматично, обикновено по формулата на Стържес, според която броят на интервалите се дава приблизително от формулата $1+\log_2 n$, където n е обемът на извадката, но може и да се променя по желание на потребителя.

В този пример, след групиране в 10 интервала получаваме следната ***таблица на честотите*** (***frequency table***)

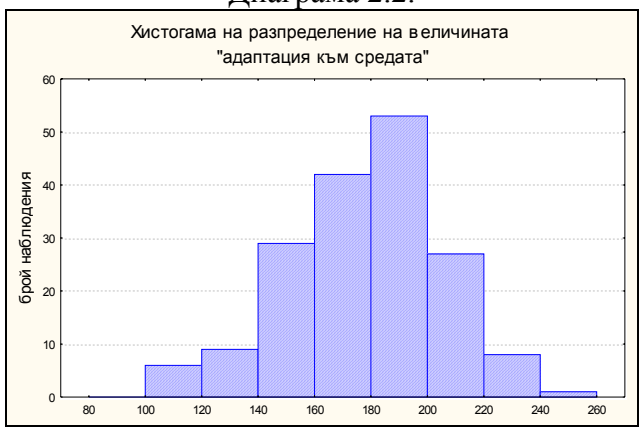
Таблица 2.3.

Интервал	Честота	Натрупана честота	Процент	Натрупан процент
80.00000<*X*<=100.0000	0	0	0	0
100.0000< *X* <=120.0000	6	6	3.43	3.43
120.0000< *X* <=140.0000	9	15	5.14	8.57
140.0000< *X* <=160.0000	29	44	16.57	25.14
160.0000< *X* <=180.0000	42	86	24.00	49.14
180.0000< *X* <=200.0000	53	139	30.29	79.43
200.0000< *X* <=220.0000	27	166	15.43	94.86
220.0000< *X* <=240.0000	8	174	4.57	99.43
240.0000< *X* <=260.0000	1	175	0.57	100

Въз основа на таблицата на честотите се изчертава ***хистограмата*** (***histogram***) на разпределението. Хистограмата е най-добрата илюстрация на разпределението на разглежданата величина. Освен илюстративно, хистограмата има важно теоретично значение относно вероятностния модел на статистическия анализ. От хистограмата могат да се забележат особености във ***формата на разпределението*** (***shape of the distribution***), например наличие на симетрия или асиметрия (лява или дясна), струпвания на наблюдения около някаква стойност и др.

В повечето случаи хистограмата не се нуждае от изрично тълкуване, а в качеството си на първично понятие се оставя да говори сама за себе си.

Диаграма 2.2.



Да отбележим специално, че изобразеното на тази диаграма разпределение трябва да се схваща като типично за повечето метрични величини.

Квантилът (***quantile***) е число върху оста на наблюденията, което разделя техния брой в определена пропорция. Когато се казва, че k_α ($0<\alpha<1$) е α-квантил (или още $\alpha.100\%$- процентов квантил) за наблюдаваната величина, се има предвид, че (приблизително) $\alpha.100\%$ от броя на наблюденията лежат вляво от k_α, а останалите $(1-\alpha).100\%$ са разположени вдясно от него. Квантилът за $\alpha=0.5$ се нарича ***медиана*** (***median***) на разпределението, $Me=k_{0.5}$. По определение медианата разделя наблюденията на две (приблизително) равни по брой части. За нашия пример имаме $Me=181$ (пресметната от компютър). Квантилът за $\alpha=0.25$ се нарича ***долен квартил*** (***lower quartile***), а квантилът за $\alpha=0.75$ се нарича ***горен квартил*** (***upper quartile***). Долният квартил, медианата и горният квартил разделят наблюденията на четири части, всяка от които съдържа приблизително 25% от броя на наблюденията. За примера имаме $k_{0.25}=160$ и $k_{0.75}=198$. Интервалът $[k_{0.25},k_{0.75}]$ съдържа около 50% от случаите. За разглеждания пример можем да направим заключението, че вероятността случайно избран индивид да има резултат по скалата между 160 и 198 е около 50%. Последното твърдение представлява характерен извод на статистическия анализ. Други специални квантили са ***процентилите*** (***percentiles***) $P_m=k_{\frac{m}{100}}$, за всевъзможните стойности $m=1,2,\ldots,99$. Програмните среди за статистическа обработка притежават заложени формули за пресмятане на различните квантили.

4. Мерки за централна тенденция и изменчивост. Когато за дадена метрична величина трябва да се покаже общото за цялата извадка чрез единствено число, се използват мерки за централна тенденция. Терминът централна тенденция е достатъчно показателен за съдържанието си. Има се предвид някоя типична обобщаваща числова характеристика на разпределението на наблюдаваната величина, която отразява адекватно количественото поведение на резултатите от цялата извадка.

Най-важната мярка за централна тенденция е ***емпиричното средно*** (***mean***)

$$\bar{x}=\frac{x_1+x_2+\cdots+x_n}{n}=\frac{\sum x_k}{n},$$

което се получава като съберем стойностите на наблюденията и разделим сумата на техния брой. За примера от предишния раздел имаме $\bar{x}=178.514$. Едва ли е необходимо да се обяснява дълго защо средното дава достатъчно обща характеристика на извадката, в съответствие с нашия опит от ежедневието, където присъства изобилие от средни величини като средни доходи, средни постижения и т.н. Като мярка за изменчивост относно средното служи ***емпиричната дисперсия*** (***variance***)

$$s_x^2=\frac{(x_1-\bar{x})^2+(x_2-\bar{x})^2+\cdots+(x_n-\bar{x})^2}{n-1}=\frac{\sum(x_k-\bar{x})^2}{n-1}$$

и свързаните с нея ***стандартно отклонение*** (***standard deviation***)

$$SD=s_x=\sqrt{s_x^2}=\sqrt{\frac{\sum(x_k-\bar{x})^2}{n-1}}$$

и ***стандартна грешка*** (***standard error***)

$$SE=\frac{SD}{\sqrt{n}}=\frac{s_x}{\sqrt{n}}.$$

За горния пример имаме $s_x^2=757.516$, $SD=27.523$, $SE=2.081$.

По-малката дисперсия означава по-голяма групиране на данните около средното и обратно. Стандартната грешка има важно значение при образуване на доверителни интервали и проверка на хипотези, което ще бъде разисквано по-нататък.

Другата основна мярка за централна тенденция е медианата $Me = k_{0.5}$, чиято роля в този смисъл също е практически очевидна. Медианата замества средното като централна тенденция в различни сравнителни анализи. В повечето случаи средното $\bar{x}$ и медианата Me имат равноценно познавателно значение. Не се определят мерки за изменчивост относно медианата. За илюстрация да пресметнем медианата на величина x със стойности $1, 5, 2, 7, 3$. Отначало подреждаме наблюденията в нарастващ ред (***вариационен ред***) и получаваме $1, 2, 3, 5, 7$. Тук медианата се средното по разположение наблюдение $Me = x_3 = 3$. Ако броят на наблюденията е четно число, например $1, 9, 5, 2, 7, 3$, то вариационният ред има вида $1, 2, 3, 5, 7, 9$, а медианата е полусборът на средните две, $Me = \frac{x_3 + x_4}{2} = \frac{3+5}{2} = 4$.

Модата (***mode***) Mo представлява мярка за централна тенденция, показваща струпването на наблюденията около някоя модална стойност. За примера имаме $Mo = 182$. Струпването около тази стойност добре се забелязва от хистограмата на диаграма 2.2. Има величини, чиито разпределения имат две (***бимодални разпределения***) или повече моди.

Разглежданият от предишния раздел пример показва, че трите описани мерки за централна тенденция не се различават много. Такова състояние на нещата е характерно в типичния случай. При малък брой наблюдения, наличието на големи отклонения в някои от стойностите обаче води до забележими промени в средното, докато медианата не се променя, което прави медианата предпочитана мярка за централна тенденция в случая на извадки с малък обем. При голям брой наблюдения, когато величината се подчинява по същество на нормално разпределение, различието между средното и медианата се получава пренебрежимо.

Мярка за изменчивост като цяло в извадката е размахът $R = x_{\max} - x_{\min}$. Полезен е също размахът между 90% и 10% квантил $D = k_{0.9} - k_{0.1}$, който представлява по-устойчива мярка в сравнение с обикновения размах, понеже при него не се разглеждат изключителните стойности (***outliers***) (много малките или много големите стойности), чиито носители обикновено са маргинални индивиди със силно отклоняващи се характеристики от общите за извадката и респективно за популацията. За нашия пример имаме $D = k_{0.9} - k_{0.1} = 213 - 143 = 70$.

5. Ординални величини. Понятието ординална величина се отнася преди всичко към начина, по който величината участва в статистическите анализи (например различни ***непараметрични*** методи), а не толкова към нейното представяне. Тя може да се зададе непосредствено чрез ранговете на наблюденията, като най-голямото по стойност получава ранг 1, а най-малкото получава ранг n (n е броят на наблюденията – обемът на извадката), но в повечето случаи в процеса на анализ нейните натурални стойности се преобразуват до рангове. Както беше отбелязано в **§1, д**околко една по природа ординална величина (зададена с натурални стойности) може да бъде интерпретирана като метрична е спорен въпрос с различни подходи за решение. Тук е полезно да се знае, че една от основните формални цели на съвременната теория на тестовете (***Item Response Theory***) е преобразуването на даден суров тестови бал, който е поне ординална по замисъл величина, до безспорно метрична величина в интервална скала. Наличието на достатъчна вариабелност в натуралните категории на величината и достатъчен по големина обем на извадката дават задоволителни обективни основания

една ординална величина да се интерпретира като метрична и върху нея да бъде прилаган богат асортимент от статистически анализи.