СТАТИСТИКА за четврта година

Постојат повеќе сфаќања за потеклото на зборот статистика. Етимолошкото потекло на зборот статистика не е со сигурност утврдено. Најверојатно потекнува од латинскиот збор „status“ што значи состојба, положба, држава. Почетоците на статистиката како наука, треба да се бараат во собирањето на податоците за државата, за состојбата на населението и имотот. Првите собирања на бројки и нивно прикажување се појавуваат со потребата на војсководците да соберат податоци за бројната состојба на населението, за бројот на способни мажи кои можат да учествуваат во битките, за бројот на даночните обврзници, големината на имотот итн., бидејќи владетелите отсекогаш сакале да знаат колкава е нивната воена и финансиска моќ. Такви собирања на податоци се вршени во Кина уште во 4.000 година пред нашата ера и во Египет во 3.000 година пред нашата ера. Најважни статистички акции во Стариот век се изведувале во Стариот Рим и тие опфаќале пребројување на жителите и нивниот имот. Ова пребројување се извршувало на секои пет години.

Појавата на електронските сметачи овозможила брз развој на статистиката. Статистичките податоци можеле да се обработуваат со голема брзина, а со тоа статистичката работа станала ефикасна и точна. Ова придонело за сè поголема примена на статистиката во изучувањето на голем број појави. Денес, статистиката не е лимитирана само на нумерички информации потребни за државата, туку е навлезена во сите пори на општествените и природните науки.

Статистиката се дели на дескриптивна статистика, статистичко заклучување и статистичка теорија.

- Дескриптивната статистика е множество на методи кои се употребуваат за собирање, обработка и претворање на податоците во информации.
- Статистичкото заклучување опфаќа збир на методи кои се користат за предвидување и прогнозирање како и претворање на податоците и информациите во знаења. На тој начин се дава податок за масата врз основа на проучувањето на примерок.
- Статистичката теорија пронаоѓа нови статистички методи, истите ги објаснува, докажува и усовршува.

✓ **Статистиката е наука која ги проучува масовните појави во природата и општеството.** Масовна појава е онаа појавакоја се јавува во голем број на индивидуални случаи. Јавувањата во масовната појава се разликуваат по многу свои особини, поради различните услови во кои се јавуваат. Поради тоа, секоја масовна појава е полна со внатрешни варирања и разликувања, што создава потреба таа да биде статистички проучена, за да може да се добие точна и правилна претстава за неа. Статистиката не смее да се поистоветува со евиденцијата, која има за цел да ги регистрира сите јавувања во масовната појава. Статистиката има многу пошироко подрачје на дејствување. Таа се занимава со собирање, обработка, презентирање, анализирање и интерпретирање на податоците за масовните појави.

Под податок подразбираме факти и броеви кои се собираат, анализираат, сумираат, презентираат и интерпретираат. На пример, еден купувач, за да одлучи да купи телевизор, треба да собере бројни податоци за видовите телевизори, нивниот квалитет, нивната цена, сервисирање на

телевизорите, гарантен рок и слично. Купувачот треба да ги среди податоците, да ги анализира, сумира и објасни и дури тогаш да донесе конечна одлука за купување на одреден вид телевизор.

Масовните појави се појави кои се јавуваат во голем број на индивидуални случаи. На пример, масовна појава (маса) се: населението, врнежите, земјотресите, производството, продажбата, компаниите, различните видови дејности и други.

Секое одделно јавување во статистичката маса се вика статистичка единица или елемент на статистичката маса.

Особеностите по кои статистичките единици се разликуваат меѓу себе се викаат белези на статистичките единици. Белегот може да се јави во повеќе видови наречени модалитети на белегот.

Постојат две групи на белези:

- квалитативни белези;
- и квантитативни белези.

- Квалитативните белези се белези кои изразуваат својства на статистичките единици и се изразуваат со зборови. Тие се делат на: предметни, кои уште се нарекуваат атрибутивни или описни и територијални или географски.
- Квантитативни белези се оние белези кои секогаш се изразуваат со број. Тие се делат на временски или хронолошки, и бројни или нумерички.

Масовните појави се составени од голем број одделни јавувања кои се разликуваат меѓу себе по своите белези. Поради тоа внатрешно варирање е потребно да се открие што е карактеристично во тоа множество од варијации за масовната појава.

Параметарот кој ја карактеризира централната вредност на нумеричкиот белег се нарекува средна големина. На тој начин големиот број на цифри од кои е составена една статистичка серија можат да бидат заменети само со една цифра. Поради тоа, средната вредност претставува дел од најзначајните показатели на нумеричките (квантитативните) карактеристики на серијата кој по соодветни дадени мерила ја репрезентира целата маса и овозможува споредување помеѓу различните маси.6 Во зависност од целта на конкретното проучување се пресметуваат различни видови средни големини за откривање на карактеристичното и типичното во појавата.

Сите тие видови средни големини се групираат во две големи групи и тоа:

- Пресметковни (математички, калкулативни) средни големини и
- Позициони средни големини.

Не само што постојат различни видови средни големини, туку и секоја големина се пресметува различно, во зависност од тоа дали се пресметува од негрупирани или од групирани податоци. Негрупирани податоци се оние каде што секој податок се прикажува одделно без оглед на бројот на неговите повторувања.

На пример: успехот на учениците во еден клас е следниот: 4,5,4,3,5,2,4,3,5,4,4,5,3,2,5,4,3,5,4,5,5,3,4,2,5,3,4,5,4,3,5,2,5.

Доколку еднаквите податоци ги групираме ќе добиеме групирани податоци, односно ќе ја утврдиме фреквенцијата на секој податок (колку пати секој податок се јавува). На тој начин, претходната низа од негрупирани податоци ќе го добие следниот изглед: во класот со доволен успех има 4 ученика, со добар успех 7, со многу добар 10 и со одличен успех 12 ученика.

Пресметковни средни големини се оние што се пресметуваат според утврдени правила врз основа на сите вредности на белегот на масата.

Во оваа група спаѓаат: аритметичката, хармониската и геометриската средина.

- **Аритметичката средина што се пресметува од негрупирани податоци се нарекува проста аритметичка средина**. Простата аритметичка средина претставува збир на вредностите на сите статистички единици во масата, поделен со бројот на единиците. Аритметичка средина се пресметува од целата маса и од примерок. Статистиката најчесто ги врши пресметките врз основа на примерок, така што сите формули и пресметки што следат ќе се вршат врз основа на примерок.

Аритметичката средина се пресметува како збир на сите вредности на примерокот, поделен со вкупниот број единици. Всушност, аритметичката средина на целата статистичка маса и аритметичката средина на примерокот се пресметува на ист начин, само се користат различни симболи. Аритметичката средина на примерокот се бележи со x . За пресметка се користи следнава формула:

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_N}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Каде што:

x̄ = аритметичка средина на примерокот x1 ,x2 , x3 ……

xn = вредности на белегот

n = бројот на единици во примерокот (број на податоци).

- **Аритметичката средина пресметана од групирани податоци се нарекува пондерирана аритметичка средина**. Се пресметува кога вредноста на белегот ќе се помножи со бројот на неговите јавувања. На тој начин добиените збирови од овие производи се делат со вкупниот број фреквенции. Симболите за проста и за пондерирана аритметичка средина се исти, но се различни формулите за пресметка. Пондерирана аритметичкасредина се пресметува со примена на следнава формула:

$$\bar{x} = \frac{f_1 x_1 + f_2 x_2 + ... + f_k x_k}{f_1 + f_2 + ... f_k} = \frac{\sum_{i=1}^{k} f_i x_i}{\sum_{i=1}^{k} f_i}$$

Каде што:

x̄ = аритметичка средна вредност (просек)

xi = вредност за белегот

f i = бројот на јавување на секоја податок (фреквенција)

k = бројот на различни модалитети на белегот.

ПРИМЕРИ:

1. Во табелата дадени се месечните приходи на работниците во една компанија. Да се пресметаат просечните месечни приходи во компанијата.

Приходи на работниците во компанијата

Број на работници (f)	Приход во денари (x)	Пондерирани износи (f*x)
6	10 800	64 800
10	14 100	141 000
15	14 200	213 000
17	15 500	263 500
13	15 000	195 000
11	15 200	167 200
10	16 000	160 000
Σ 82		Σ 1 204 500

2. Да се пресмета просечниот износ на штеден влог за детско штедење во Стопанска банка во 2008 година.

Штедни влогови за детско штедење

Штеден влог во евра	Број на штедачи f	Среден износ на штедниот влог x	Пондериран износ x* f
2-50	6 310	26	164 060
51-100	4 184	75	313 800
101-300	1 032	200	206 400
301-800	74	550	40 700
Вкупно	**11 600**		**724 960**

- **Хармониската средина** што се пресметува од негрупирани податоци се нарекува проста хармониска средина. Се пресметува кога вкупниот број на податоци ќе се подели со збирот од реципрочните износи на вредностите на белегот.

Хармониска средина е специфична средна големина, која се користи во исклучителни случаи, кога вредностите на белегот се во обратно пропорционален однос со неговата големина.

Хармониската средина се обележува со симболот Mh. Се пресметува со примена на следнава формула:

$$Mh = \frac{N}{\frac{1}{x_1} + \frac{1}{x_2} + \frac{1}{x_3} ..... + \frac{1}{x_n}} = \frac{N}{\sum_{i=1}^{k} \frac{1}{x_i}}$$

Каде што:

Mh = хармониска средна големина

N = бројот на податоци

x = вредност на белегот

ПРИМЕР:

1.Шест работника произведуваат ист вид производ. Времето потребно за изработка на една единица производ по одделен работник е дадено во табелата. Да се пресмета просечното потребно време за изработка на единица производ.

Време потребно за изработка на една единица производ по одделен работник

Работник	Потребно време за изработка на единица производ во минута (x)
А	10
Б	8
В	7
Г	9
Д	12
Е	11

Хармониската средина пресметана од групирани податоци се нарекува хармониска пондерирана средина. Се пресметува кога вкупниот број на фреквенции ќе се подели со збирот на поделените износи меѓу бројот на фреквенциите и вредноста на белегот.

Формулата за пресметка на хармониската пондерирана средина е следнава:

$$Mh = \frac{\sum_{i=1}^{k} f_i}{\sum_{i=1}^{k} \frac{f_i}{x_i}}$$

каде што:

Mh = хармониска средна големина

f i = бројот повторувања на податоците

x = вредност на белегот.

ПРИМЕР:

Во компанијата ПП за производство на една единица од производот А се троши следново време:

Број на работници (f)	34	30	12	8	5	1
Потрошено работно време (x)	24	26	28	30	32	34

Да се пресмета просечното потрошено време за изработка на една единица од производот А.

- **Геометриската средина** претставува облик на средна вредност со која се изедначуваат релативните или пропорционалните разлики во вредностите на белегот. Геометриската средина е корисна во пресметување на просек од проценти, индекси или стапки на растеж.

Геометриската средина што се пресметува од негрупирани податоци се нарекува проста геометриска средина. Формулата за нејзина пресметка е следнава:

$$Mg = \sqrt[N]{x_1 * x_2 * ..... * x_n}$$

каде што:

Mg = геометриска средина

x1 = првиот податок

xn = последниот податок

N = број на податоци

ПРИМЕР:

Процентите на заработка од четири видови продадени производи во вкупниот профит на компанијата се 3%, 2%, 4% и 6%. Да се пресмета геометриската средина.

Геометриската средина пресметана од групирани податоци се нарекува геометриска пондерирана средина. Формулата за нејзиното пресметување е следнава:

$$Mg = \sqrt[N]{x^{f_1} * x^{f_2} * ..... * x^{f_n}}$$

каде што:

Mg = геометриска средина

x1 = првиот податок

xn = последниот податок

f 1,2,3....n = број на фреквенции

N = број на податоци

ПРИМЕР:

Врз основа на податоците од табелата да се пресмета геометриската средина на прометот.ж

Промет на фирмите во една општина

Промет во милиони денари	Број на фирми (f)	просечен промет	$\log x$	$f \times \log x$
10-20	2	15	1,17609	2,35218
20-30	7	25	1,39794	9,78558
30-40	5	35	1,54407	7,72035
40-50	3	45	1,65321	4,95963
50-60	2	55	1,74036	3,48072
60-70	1	65	1,81291	1,81291
Вкупно	**20**			**30,11137**

- **Медијаната** е вредноста од низата нумерички податоци која ја дели низата на два еднакви делови. Ако низата содржи парен број на податоци, медијаната е аритметичка средина од двете вредности.

  Пример: 6, 6, 7, 7, 7, 8, 8, 9, 9, 9

  Медијаната е погоден податок за средната вредност кога нема многу податоци. Во случај кога постои можност за субјективно оценување медијаната е подесена, затоа што е неосетлива на екстремни вредности.

- **Мода** е вредноста која најчесто се појавува во низата на податоци. Таа е резултатот кој се јавува најголем број пати, што значи резултатот, вредноста со најголема фреквенција. Може да постојат повеќе моди, а може и да не постои.

Хи квадрат тест

Хи- квадрат тест

Хи квадрат тестот спаѓа во групата на непараметарските тестови. Овој тест е еден од постарите статистички тестови. Тестот го разработил Карл Пирсон во 1900-тите години, па познат е и под називот Пирсонов тест. Тестовите кои се засновани на x2 распоредот опфаќаат цела низа проблеми кои можат да се однесуваат на модалитетите на еден или повеќе белези.

Постапката наречена хи-квадрат тест се употребува во повеќето случаеви ако се работи за квалитативни податоци или ако појавата значајно отстапува од нормалата. Хи-квадрат тестот е многу практичен тест кој може особено да послужи кога сакаме да утврдиме дали некоја добиена фрекфренција отстапува од фрекфенцијата која ја очекуваме со одредена хипотеза. Кај овај тест истотака истражуваме дали постои поврзаност помеѓу две варијабли и тој ја покажува веројатноста на поврзаност, како и хомогеност на популацијата. Со други зборови кога се испитуваат два белега X и Y обично се поставува нултата хипотеза за нивната независност, при тестирањето на независноста на принципите на класификација, x2 тестот треба да покаже дали модалитетите на белезите класифицирани по одредени критериуми се зависни или независни. Така на пример, може да се тестира: Зависност на сообраќајните прекршоци од староста на возачите, Завинсост на работниците според распоредот на платите и должината на работиот стаж, Зависност на времето на задоцнување од должината на работниот стаж итн. Тестот на независноста овозможува донесување одлука во врска со прифаќањето или неприфаќањето на нултата хипотеза т.е. постоење или непостоење значајна разлика помеѓу емпириските и очекуваните фрекфенции според еден или друг критериум. Постојат два вида на хи – квадрат тестови

Хи квадрат тест во облик на распоред

Тоа е нај тест кој треба да покаже дали емпирискиот распоред статистички значајни се разликува од теоретскиот. x^2 тестот во облик на распоред се заснова на разликата помеѓу емпириските фреквенции на модалитетите (fi) и очекуваните фреквенции на тие модалитети (fit) аналогно на нивниот претпоставен распоред. Со хипотезата треба да се специфицира очекуваниот распоред (биномен,униформален,нормален),а тестот треба да покаже дали емпирискиот распоред на примерокот значајно се разликува од очекуваниот. Најпроста хипотеза е за униформален распоред. (иста френквенција за сите модалитети)

реализирана вредност $X^2=\sum_{i=1}^{r} ((fi-⟦fi⟧^{(t)})^2)/⟦fi⟧^t$
fi-емпириска фрекфренција
fit-очекувана (теоретска)фрекфренција
степени на слобода V= r-m-1
r-број на модалитети на белегот
m-параметар на распоред
критична вредност $x^2\alpha;v1$
α-ниво на значајност

x² статистиката на тестот се применува на податоците кои можат да се сведат на апсолутни фреквенции.

Кога X²α;v1 критичната вредност е поголема од реализирана вредност се прифаќа Но

H0 : Емпирскиот распоред е униформен

кога X² реализирана вредност е поголема од критичната се прифаќа H1

H1 : Емпирискиот распоред не е униформен

Пример: Бројот на гледачи на ФК Вардар на последните четири натпревари се движел на следниот начин:

Натпревар	Број на играчи fi	f_i^t	fi x f_i^t	$(fi- f_i^t)^2$	$(fi-f_i^t)^2 / f_i^t$
1	3955	4000	-45	2025	0.506
2	3863	4000	-137	18 767	4.692
3	4237	4000	237	56 169	14.042
4	3945	4000	-55	3025	0.756
	19.99	**19.99**	**19.99**	**19.99**	**19.99**

На ниво на значајност од 0.01 да се испита дали може да се испита хипотезата дека бројот на гледачи на ФК Вардар по натпревари има униформен распоред.

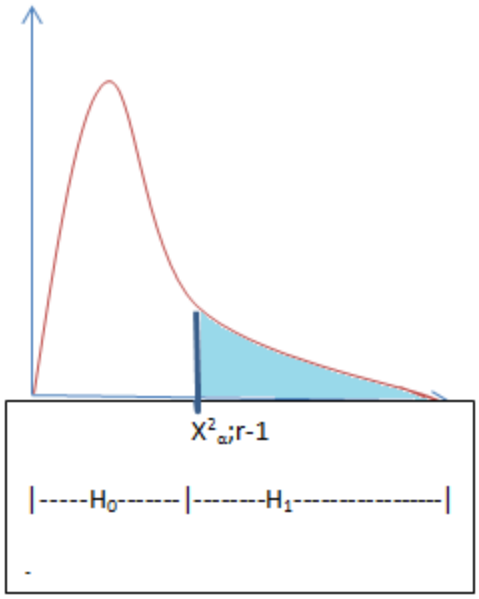
H0 : Емпирскиот распоред е униформен (бројот на гледачи на Вар дар е ист за сите натпревари)

H1 : Емпирискиот распоред не е униформен / статистички значајно се разликува

X2α;r-1= X20.01;4-1= 11.345 критична вредност

X²=∑_(i=1)^r ((fi-⟦fi⟧^(t)²))/⟦fi⟧^t= 19.99 реализирана вредност

X2 >X2α;v односно 19.99 > 11.345



H1 се прифаќа, а тоа значи дека емпириското распоред не е униформен(бројот на гледачи на натпреварите на Вардар се разликува по натпревари).

Хи квадрат тест на независност на модалитетите на два белега

При тестирањето на независноста на принципите на класификација, x² тестот треба да покаже дали модалитетите на белезите класифицирани со пределени критериуми се зависни или независни. За тестирање на независноста користиме табели на контингенција. **Табели на контингенција** се табели каде се презентирани емпириските фреквенции (fij) за ij-та комбинација на модалитетите на два белега, кои треба да се споредат со очекуваните фреквенции (fij^t) со цел да се изврши тестирање на независнос

реализирана вреедност $x^2=\sum_{i=1}^{r}\sum_{j=1}^{k}((fij-⟦fij⟧^{(t)^2}))/⟦fij⟧^t$
критична вредност $x^2\alpha;v1$
степени на слобода V=(r-1)(k-1)
r- модалитети на еден белег
к- модалитети на другиот белег

Кога $X^2\alpha;v1$ критичната вредност е поголема од реализирана вредност се прифаќа Но

H0: Модалитетите на белезите се независни

Кога X^2 реализирана вредност е поголема од критичната се прифаќа H1

H1: Модалитетите на белегзите се зависни

За мерење на интензитетот на зависност на набљудуваните модалитети се користи **коефициент на контингенција (C)**

$C=\sqrt{(x^2/(n+x^2))}$ $0<x^2<1$

$Cmax= \sqrt{((r-1)/r)}$- ако r=k

Табела на контингенција Пример:

	Брзина		**Сигурност**		**Изглед**		**Потошувачка на бензин**		
	fij	**fij^t**	**fij**	**fij^t**	**fij**	**fij^t**	**fij**	**fij^t**	
Машки	30	18.82	10	18.82	25	25.88	15	16.47	**80**
Женски	10	21.18	30	21.18	30	29.12	20	18.53	**90**
	40	**40**	**40**	**40**	**55**	**55**	**35**	**35**	

Fij	**Fij^t**	**Fij-Fij^t**	**$(Fij-Fij^t)^2$**	**$(Fij-Fij^t)^2/fij^t$**
30	18.82	11.18.	124.99	6.64.
10	18.82	-8.82	77.79	4.13.
25	25.89	-0.89	0.79	0.03
15	16.47	-1.47	2.16.	0.13
10	21.18	-11.18	124.99	5.90.
30	21.18	8.82.	77.79	3.67.
30	29.12.	0.88	0.78	0.03
20	18.53	1.47.	2.16.	0.12
170	**170**			**20.65**

H0 : Модалитетите на двата белега се независни (разликите во полот не влијаат),полот не влијае во бараните карактеристики.

H1:Модалитетите на двата белега се зависни (разликите во полот влијаат), мажите и жените имаат различни барани карактеристики.

$X2=\sum_{(i=1)}^{r} \sum_{(j=1)}^{k} ((fij-⟦fij⟧^{(t)})^2)/⟦fij⟧^t =20,65$

$X2\alpha;(r-1)(k-1)=7.8$

$X2 > X2\alpha;v$ односно 20,65 > 7,8

H1 се прифаќа, модалитетите на двата белега се зависни.

.

Мали очекувани фреквенции

Кога ќе се случи очекуваната фреквенција (⟦fij⟧^t) да е помала од 5, треба да се изврши прегрупирање на податоците- спојување на два модалитети со мали фреквенции во еден модалитет Кога примерокот е доволно голем, очекуваната фреквенција може да биде помала од 5, па дури и помала од 1.