До г-жа Ваня Григорова
Изпълнителен директор
на Изпълнителната агенция по околна среда

ДОКЛАД

от д-р Румяна Костова

за извършена работа договор № 1962/13.04.2011. По проект BG0052 „Разработване на информационна система към национална система за мониторинг на биологичното разнообразие в България”

София

13.06.2011

В изпълнение на възложената ми задача представям извършената работа по точка 1 от договора:

СЪДЪРЖАНИЕ:

1. Планиране на наблюденията по методиките за мониторинг в рамките на НСМБР

1.1 Изисквания към рутинните методи на събиране на информация от гледна точка на репрезентативност на извадките и използване на статистически методи.....5

1.2. Алгоритъм на провеждане на изследвания, подлежащи на статистически анализ.....12

1.2.1 Избор на размер и форма на индивидуалните наблюдения (извадъчни единици).....13

- Определяне на подходящи индивидуални наблюдения.....13

- Определяне на минималния необходим размер на индивидуалните наблюдения.....15

- Оценка на необходимия минимален брой индивидуални наблюдения (обем на извадката).....19

- Брой на необходимите повторения и избягване на псевдоповторенията.....20

1.2.2. Схеми за съставяне на извадка.....21

1.2.2.1. За видове със средно и голямо обилие.....26

1.2.2.2. За редки видове и видове, живеещи на гъсти групи.....33

1.2.3. Изработване на мониторингови схеми.....34

1.3. Преглед и оценка на данни от проведен досега мониторинг, с оглед на пригодност за статистическа обработка.....38

1.3.1. Общи положения, засягащи събираната информация от формуляри.....38

1.3.2 Преглед на формулярите по избрани групи за мониторинг.....40
1.4. Литература.....43
2. Възможности за статистически анализ с помощта на софтуерния продукт SPSS, закупен в рамките на проект "Разработване на Информационна система към НСМБР в България".
2.1. Цел и задачи на основните статистически методи, включени в следните модули:
2.1.1. SPSS® Advanced Statistics 18.....44
2.1.1.1 Обикновена и множествена линейна регресия.....44
2.1.1.2 Общ линеен модел *(GLM)*.....49
2.1.1.3. Компонентен анализ на дисперсията.....55
2.1.1.4. Смесени линейни модели.....56
2.1.1.5. Обобщени линейни модели *(GENLIN)*.....57
2.1.1.6. Общ log-линеен модел.....63
2.1.1.7. Logit log-линеен модел.....63
2.1.1.8. Анализи на преживяемост.....64
- Статистически жизнени таблици.....64
- Kaplan-Meier анализ.....66
- Регресия на Кокс.....69
2.1.2. SPSS® Regression 18.....70
2.1.3.SPSS® Missing Values 18.....72
2.2. Примери за приложението на функциите в модулите при разрешаване на биологични проблеми.....76
2.3. Препоръки относно експорта на данните от базата данни към съответните модули на SPSS с цел статистически анализ на резултатите от мониторинга на регионално и национално ниво.....87
2.4. Литература.....88
3. Биостатистически методи за обработка на информацията от Националната база данни, за биологичните групи: висши растения, бозайници, птици.

3.1.Обобщаване на резултатите относно актуалното състояние на обектите на мониторинг по заложените параметри.....90
3.1.1. Дескриптивна статистика.....90
3.1.2. Тестове за сравняване на две и повече извадки.....95
3.1.3 Регресионни анализи.....101
3.1.4. Индекси за разнообразие.....111
3.1.5. Класификация и ординационни анализи на съобщества.....113
3.1.6. Оценка на ефективността.....116
3.2. Полова и възрастова структура на популациите на отделни видове.....116
3.2.1. Графично представяне и сравняване на полова и възрастова структура.....116
3.2.2. Жизнени таблици.....118
3.2.3. Анализ на преживяемост и оценка на риска с SPSS.....124
3.3. Литература.....132

1. Планиране на наблюденията по методиките за мониторинг в рамките на НСМБР

Най-общо методите, които се използват за оценка размерите на наблюдаваните популации са следните:

- Пълно преброяване – обикновено се използва за големи или добре видими организми, живеещи на групи или в колонии. Така напр. е определена световната популация на *Sulla bassana*, морска птица гнездяща в няколко плътно заселени колонии по северното крайбрежие на Атлантическия океан.

При такова преброяване не се налага статистическа обработка, тъй като промените се отчитат директно.

- Извадъчни методи: метод на пробните площадки (полигони, трансекти); метод на маркиране и последващ повторен улов (определена част от популацията се улавя, маркира и пуска, след което се установява делът на повторно уловените индивиди и така се оценява числеността на популацията); дистанционни методи.

1.1. Изисквания към рутинните методи на събиране на информация от гледна точка на репрезентативност на извадките и използване на статистически методи

Генералната съвкупност е термин, описващ цялата съвкупност от индивиди обект на мониторинга. Понякога може да включва цялата биологична популация или някаква част от популацията. Понякога генералната съвкупност не се състои от индивиди, а от набор индивидуални обекти, за които се правят изводи.

Извадката представлява сбор на индивидуални наблюдения (извадъчни единици) взети от генералната съвкупност (напр. дадена популация, метапопулация и т.н.). Индивидуалните наблюдения могат да бъдат условни обекти и естествени обекти. Пример за условни са полигоните, трансектните линии, няколко полигона или точки разположени по линия или в групи, водни проби с определен обем, брой уловени насекоми в почвен капан за единица време и т.н. Пример за естествени индивидуални наблюдения са дънери, постоянни локви в каменистите морски брегове, листа, отделни индивиди растения или животни или част от индивиди (напр. при растенията – брой семена, шушулки).

При съставяне на извадка трябва да се имат предвид изискванията към извадките: еднородност, обем и случайност.

Параметрите на генералната съвкупност са описателни оценки, които я характеризират и се приема, че са постоянни, но неизвестни величини, които се променят само ако генералната съвкупност се променя. Бележат се с гръцки букви.

Извадъчните показатели са описателни оценки на извадката, даващи оценка на параметрите на генералната съвкупност. Те варират в определени граници между извадките от една и съща генерална съвкупност и се променят, когато се променя генералната съвкупност. Бележат се с латински букви.

Репрезентативността на дадена извадка позволява надеждна оценка на стойностите на изследваните параметри на генералната съвкупност, от която е взета.

Данните в извадката трябва да са взети с достатъчна точност и прецизност. **Точността** е близост на измерената или изчислена стойност до истинската стойност на белега. **Прецизността** е близост между повторени измервания на една и съща величина. Ефективните схеми за съставяне на извадка, целят висока прецизност. Стандартното отклонение на средната аритметична, получена от средните аритметични на няколко извадки от една генерална съвкупност дава оценка на прецизността (по-голямо стандартно отклонение означава по-ниска прецизност).

Грешките, допускани при събиране на извадка са важен фактор за нейната репрезентативност. Те трябва да се вземат предвид и съответно да се изчистят или намалят.

Типове грешки при взимане на извадки:

Грешка на репрезентативността. Те са случайни грешки, които се получават, когато информацията от извадката не отразява истинската информация за генералната съвкупност. Причината е, че се работи само с част от генералната съвкупност, т.е. се взимат данни само за част от индивидуалните наблюдения в една генерална съвкупност. Грешката намалява с увеличаване на обема на извадката.

Другите типове грешки са свързани с човешкия фактор и обикновено не са случайни грешки:

-Методични

-Грешка на точността

-Грешка на вниманието

-Грешка на типичността

Примери:

• Субективно разполагане на индивидуалните наблюдения, заместване на индивидуалните наблюдения с по-достъпни за взимане на данни.

• Използване на индивидуални наблюдения, в които атрибутите (белезите) не могат да бъдат измерени или преброени точно. Напр. при преброяване на стъбла на тревисти растения в квадрати където се намират стотици от тях може да доведе до грешка при преброяване.

• Несъгласуван полеви опит при взимане на извадка. Такива грешки стават, ако различните членове на изследователски екипи имат различно ниво на полеви опит (напр. един изследовател брои тревисти растения отвисоко, докато друг е коленичил до квадрата) или способности (напр. един изследовател не може да чуе високочестотни звуци на птици, докато друг може).

• Грешки при запис и презаписване. Напр. когато човекът, вкарващ данните от полевия формуляр в базата данни не може да разчете правилно някои знаци, попълвани от друг член на екипа.

• Неправилна или несъгласувана идентификация на видове. Тук се включват и отклонения, зависещи от липсата на някои класове размери или цветни морфи.

Тези грешки трябва да се сведат до минимум, тъй като при съставяне на схемите за взимане на извадка и статистическите анализи се приема, че са нула.

Систематичната грешка при снемане на показателите трябва да е еднаква, т.е. сравними данни може да има само ако методиката и хората, които работят по нея събират данните по един и същи начин с един и същи вид уред.

Един от начините за оценка на риска да се получат извадъчни показатели, които силно се отличават от истинските параметри на генералната съвкупност е да се направят повторения, т.е. да се вземат няколко независими извадки от генералната

съвкупност, за да се видят разликите между тях. Ако почти всички независимо получени показатели са сходни, това означава, че схемата за съставяне на извадката е добра и има висока прецизност. Обратното, ако се различават прецизността е ниска. **Извадъчните разпределения** представляват разпределение на стойностите на извадъчните показатели на взетите независими извадки от една и съща генерална съвкупност. Графично се изобразяват чрез хистограма или полигон на честотите. При съставяне на схема за събиране на извадка целта е да се направят тези разпределения, колкото се може по-тесни (т.е извадъчните показатели за генералната съвкупност да варират колкото се може по-малко при различните извадки).

Стандартната грешка на средната е стандартно отклонение на голям брой средни аритметични на независими извадки. Тя е мярка за прецизността на средната аритметична на извадката.

$$SE = \frac{s}{\sqrt{n}}$$

Доверителните интервали на средната аритметична дават вероятността (при 95% ниво на достоверност) в този интервал да се намира истинската стойност на генералната съвкупност.

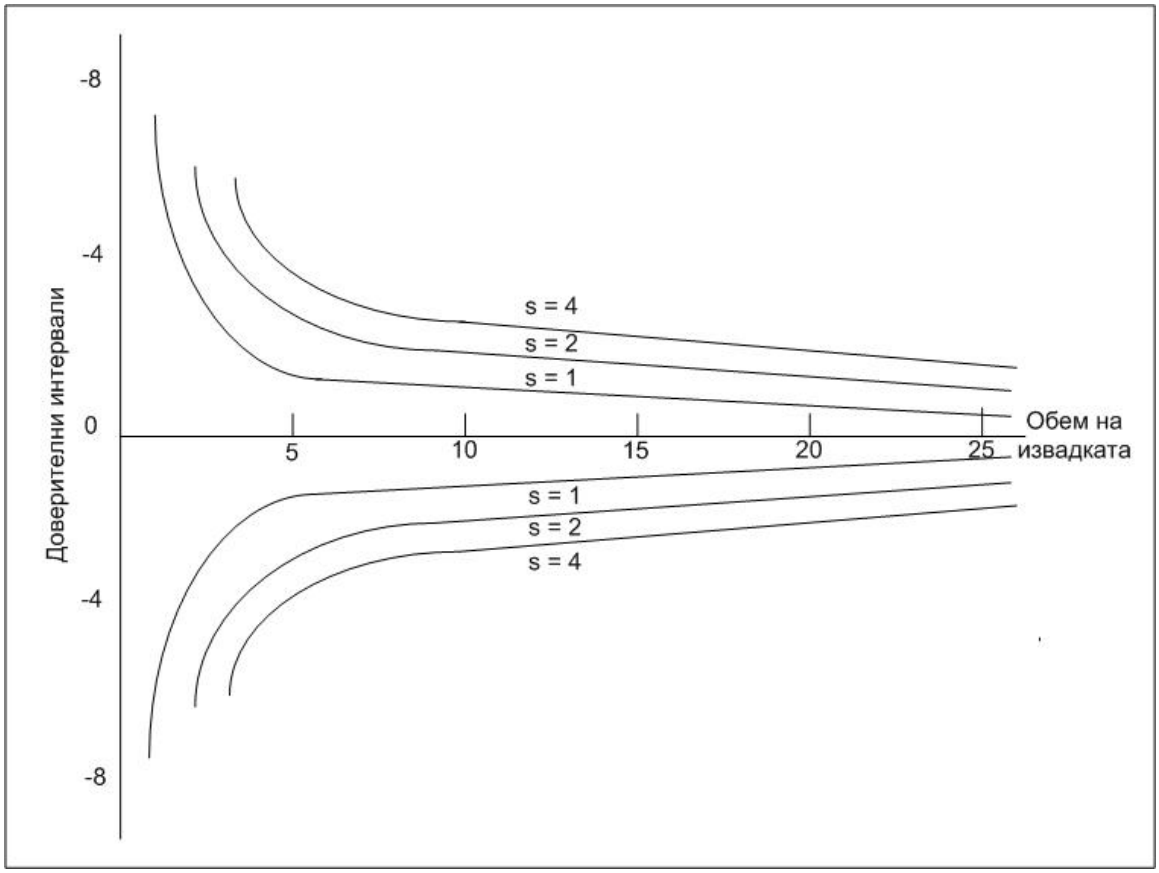
$$\bar{X} \pm t_{n-1}\left(\frac{s}{\sqrt{n}}\right)$$

При съставяне на извадка целта е да се намали стандартната грешка и да се стеснят доверителните интервали. Това става или като се увеличи обема на извадката или като се намали стандартното отклонение.

При вземане на няколко извадки с различна големина от една генерална съвкупност ще се наблюдават разлики в средните аритметични. Различията между тях (дисперсията) се дължат на грешката на репрезентативността. Тя зависи от дисперсията на генералната съвкупност и от обема на извадката.

Стойността на t е константа за даден обем на извадката и ниво на значимост. Получава се от таблиците за t -разпределението на Стюдънт. Може да се срещне като корекционен фактор за малки извадки. За големи извадки ($n \geq 30$) при 95% ниво на достоверност, t е = 1.96. За извадка от 5 стойности при 4 (df = 5-1) степени

на свобода и 95% ниво на достоверност t е = 2.776. С увеличаване на обема стойността на t се доближава до 1.96. От уравнението горе става ясно, че ширината на 95 %- доверителните интервали, мярка за прецизността на средната аритметична зависят от два фактора обема на извадката и вариацията в генералната съвкупност, оценена с извадъчното стандартно отклонение – *s*. Тъй като не е възможно дисперсията на генералната съвкупност да се променя, прецизността може да се увеличи само с увеличаване на обема на извадката. С увеличаване на обема диапазона на доверителните интервали намалява. За малки извадки намаляването на стандартната грешка става с малко увеличаване на обема (Фиг.1)



Фиг. 1. Намаляване на доверителните интервали (при 95% ниво на достоверност) с увеличаване на обема на извадката. Линиите са за различно стандартно отклонение (s). За изчислението на доверителните интервали е използвано t - разпределението на Стюдънт. Колкото повече обемът на извадката клони към безкрайност, толкова доверителните интервали клонят към 0.

Дисперсията, стандартната грешка на средната и доверителните интервали (95%) са индикатори за прецизността на извадката.

Относителната прецизност (в %) за даден параметър на генералната съвкупност е средната аритметична на разликите между доверителните интервали (95%) като процент от стойността на извадъчния показател. Въпреки че доверителните интервали може да са асиметрични относителната прецизност (PRP) се изчислява по следната формула:

$$PRP = \frac{(CL_2 - CL_1)/2}{\hat{N}} \times 100 = \frac{CL_2 - CL_1}{\hat{N}} \times 50$$

където $\hat{N}$ е обемът на генералната съвкупност определен чрез извадката, $\hat{N}$ може да се замести с всеки друг параметър който се оценява (напр. средна аритметична и т.н.). CL_1 и CL_2 са долната и горната граница на доверителния интервал.

Грешките тип α и β са свързани със ситуации, където се сравняват две или повече средни аритметични или дялове на извадки с някакъв статистически тест. Тези сравнения напр. може да са между две и повече места, два и повече периода и т.н. Напр. ако са взимани извадки от една генерална съвкупност в две различни години искаме да разберем дали за този период е настъпила промяна. Винаги, когато се използват статистически тестове се започва с допускане, наречено **нулева хипотеза H_0**. Тя гласи, че няма различия. При интерпретиране на резултатите от мониторинга може да се стигне до две решения: 1) Да решим, че е станала промяна или 2) Да решим, че не е станала промяна. И в двата случая или ще сме прави, или ще сгрешим. (Табл. 1)

Контролирането на тези грешки е изключително важно, тъй като може да се стигне до влошаване състоянието на наблюдаваната популация или пък до предприемане на ненужни действия.

Съществува известна вероятност да заключим, че е станала промяна, а в действителност да няма (т.е. да сгрешим, отхвърляйки нулевата хипотеза). Тази вероятност се бележи с **P** и е един от видовете информация, които се получават при статистическите анализи. Стойността на **P** е вероятността наблюдаваната разлика

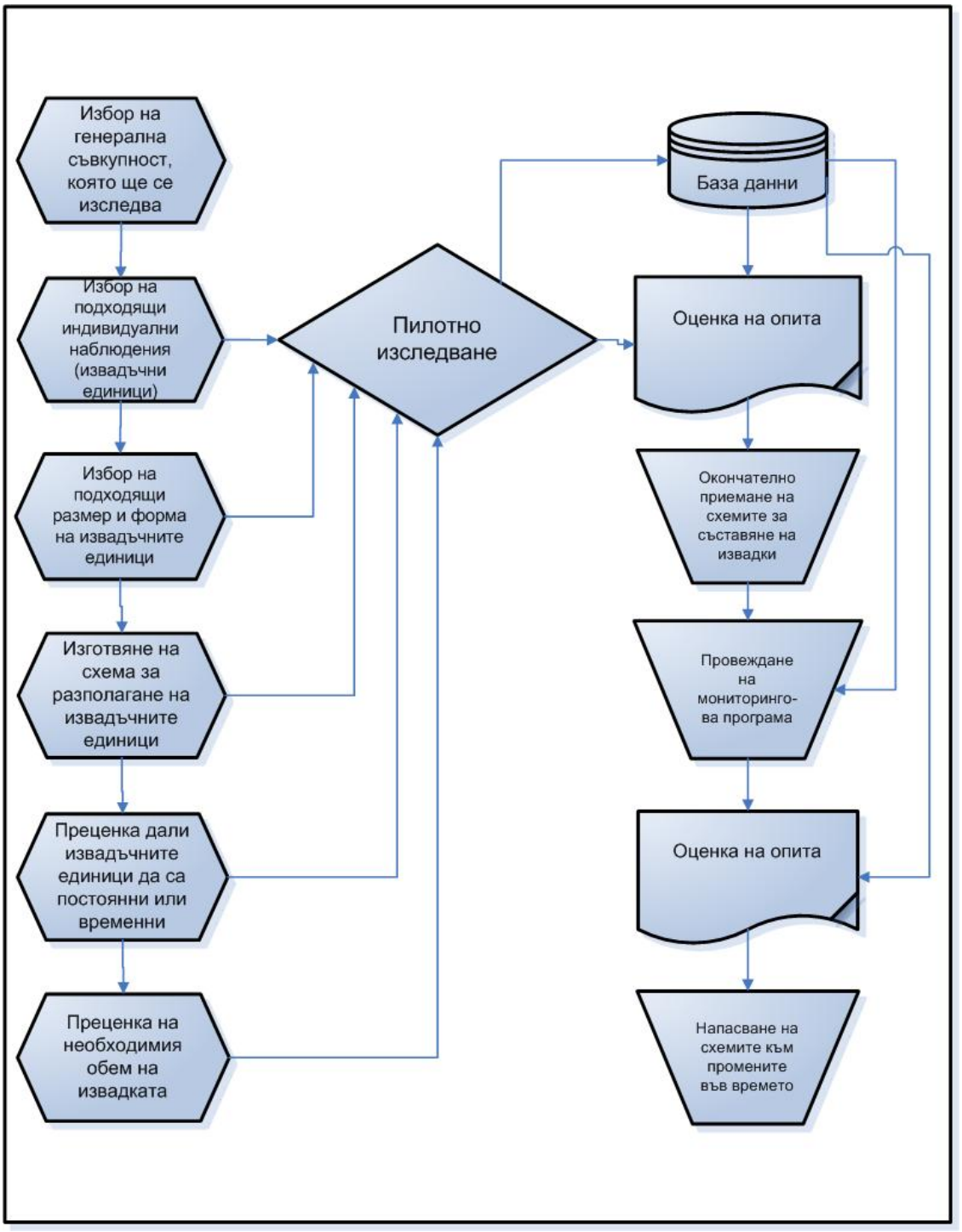
да е резултат от грешка тип **α**. За да приемем разликата за достоверна стойността на **P** трябва да е под нивото на значимост (= 0.05; = 0.01; = 0.001 - допустимата вероятност за получаване на случайни отклонения от установените с определена вероятност резултати).

Таблица. 1. Възможни изходи при провеждане на тестове за сравняване на две извадки

	В действителност няма промяна	В действителност има промяна
Открита е промяна в популацията при мониторинг	Грешно установена промяна **α - грешка** (тип I)	Вярно установена промяна Сила на теста = 1- β
Не е открита промяна в популацията при мониторинг	Вярно установена липса на промяна 1-α	Грешно установена липса на промяна **β – грешка** (тип II)

Статистическата **сила на теста** е противоположната вероятност на вероятността на грешката тип **β**. Приема се, че желаната сила на теста трябва да е ≥ 0.8, което означава малка вероятност да не се открие разлика, ако има такава. Особено важно е да се има предвид при мониторинг на застрашени видове, при които, ако не се установи достоверно намаляване на популацията във времето може да се окаже фатално.

1.2. Алгоритъм на провеждане на изследвания, подлежащи на статистически анализ.



Фиг. 2. Обща схема на дизайн на извадки с цел провеждане на мониторингова програма.

1.2.1 Избор на размер и форма на индивидуалните наблюдения (извадъчни единици)

Няма дефинирани стойности за „правилен" обем на извадката и „правилна" форма и размер на индивидуалните наблюдения. Те могат да се преценят и оценят в повечето случаи едва след провеждане на пилотно изследване и последваща оценка на ефективността.

В идеалния случай извадъчните единици трябва да са еднакви по размер и форма, но това не винаги се получава. В такъв случай, когато се комбинират стойности от различни индивидуални наблюдения трябва да се използват средни аритметични, на които се дава тежест (*weighted means*), както и подходящо обобщени стойности на дисперсията.

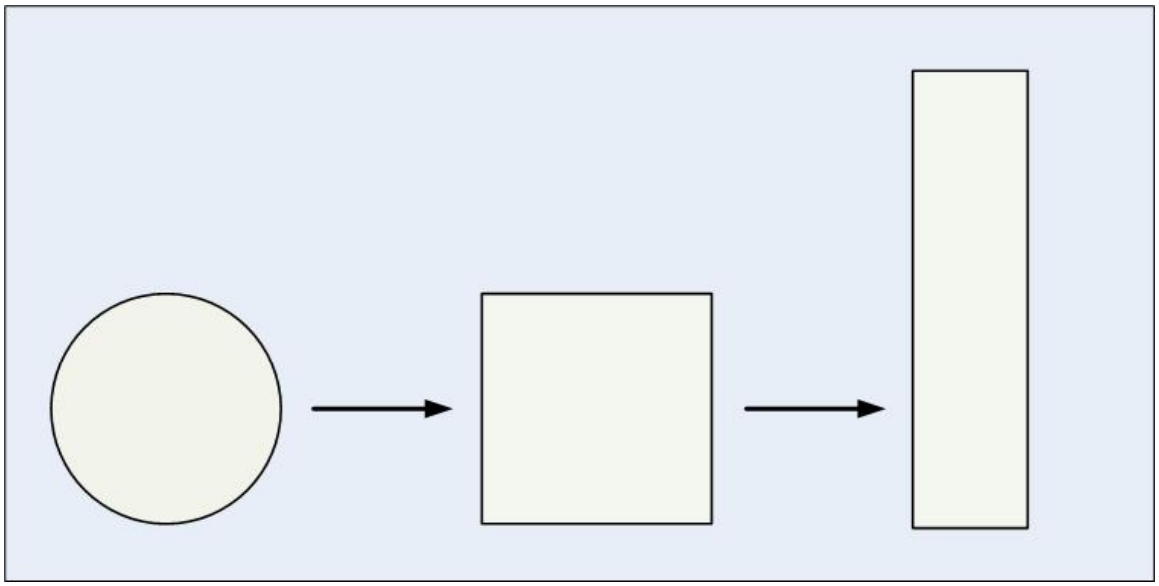
- *Определяне на подходящи индивидуални наблюдения.*

Извадъчните единици трябва да са ясно дефинирани, случайни и независими.

Изборът на извадъчни единици зависи от отношението периметър към вътрешна площ, методът за получаване на броя индивиди в него, както и от пространственото разпределение на вида – равномерно, групово или случайно. Ако не се направи внимателна преценка на размера и формата на пробните единици, много от тях може и да не пресекат (или да не покрият) находищата на вида.

Съотношението периметър към площ е мярка за т.нар. периферен ефект. **Квадратните и кръглите полигони** имат по-малко съотношение периферия-вътрешна площ сравнено с полигони с правоъгълна форма със същата площ (Фиг. 3). Силно удължените полигони се наричат "**лентови трансекти**". Този ефект е свързан с това, че в определени случаи хората, извършващи мониторинг не могат да преценят дали организмите, намиращи се по периферията на полигона са вътре или извън него, което води до грешно отчитане на броя индивиди. При по-подвижните организми този ефект е по-засилен поради това, че могат да напуснат полигона преди да бъдат преброени. Решения на проблема при такива случаи са използване на полигони с по-компактна форма, преброяване в много кратък интервал от време или заграждане на полигона с бариери. Съотношението

периферия – вътрешна площ намалява с увеличаване размера на полигоните, което води до намаляване на грешката. От друга страна, обаче изморителното пребройване на организми в големи полигони при трудни условия (напр. трудна проходимост) също може да доведе до грешки.



Фиг. 3. Три форми на полигони с еднаква площ, но различно съотношение периметър/вътрешна площ, мярка за т.нар периферен ефект. Кръглият полигон има съотношение 2.36:1, квадратният 2.67:1, а правоъгълния (лентов) 3.33:1. По-ниска стойност съответства на по-нисък потенциален периферен ефект.

Вероятността за откриване на индивидите (*detectability*) варира в зависимост от формата и размера на изследвания полигон. Дългите и тесни полигони са по-ефективни при откриване на индивиди в сравнение с квадратните и кръглите (разстоянието от централната ос до периферията е много по-малко и вероятността да бъдат пропуснати индивиди е много по-малка).

При определяне на формата и размера на полигона, друг фактор е хомогенността на изследваните хабитати. В хетерогенни хабитати данните, получени от полигони с по-голяма дължина от ширина често са с по-малка статистическа разлика помежду им, в сравнение с компактни плотове със същата площ.

Освен полигоните, които обикновено са оградена площ за преброяване на организми в даден район се използват и други техники. **Линейните и точковите трансекти** са специализирани полигони, при които търсенето на изследваните организми става по тясна ивица с позната площ.

Обилието на изследваните популации може да се установи също и с извадъчни единици, които не са полигони. При тези случаи се използват оценки, с които се описва разстоянието между индивидите в дадена площ. Тези методи са базирани на допускането, че броят на индивидите в една популация може да се определи чрез измерване на средното разстояние между индивидите в популацията или между индивидите и случайно избрани точки в местообитанието. Това са т. нар. **Distance методи,** които се използват в ботанически изследвания и са адаптирани за инвентаризация на редки растения или други неподвижни организми. Подходът може да се използва при популационни изследвания и на по-подвижни видове животни, чрез отчитане на плътността на техните гнезда, дупки, места за спане или изпражнения.

Предимства на тези *дистанционни* методи пред полигоните и трансектите при използването им за мониторинг са: 1. Не са чувствителни към грешки на отчитане, които се случват при преброяване близо до границите на полигоните. Това води до по-точна оценка на обилието. И 2. Времето и усилията за получаване на адекватна извадка при дистанционните оценки в дадено място са по-малко, отколкото когато се търси всеки индивид от даден вид в полигон. Това води до увеличаване на ефективността на мониторинговата програма.

Всички дистанционни методи включват методи за случаен избор на точките и ориентация по компас.

- *Определяне на минималния необходим размер на индивидуалните наблюдения*.

Пример на преценка за подходящ размер на полигони: Целта на изследването е да се определи плътността на популацията на охлюви от род *Patella* по скалистия литорал (скали, постоянно заливани от водата). Решено е да се използват квадрати, в които да се преброяват охлювите. Броят на отчетените индивиди ще зависи от големината на квадратите и плътността на популацията. Ако се използват малки

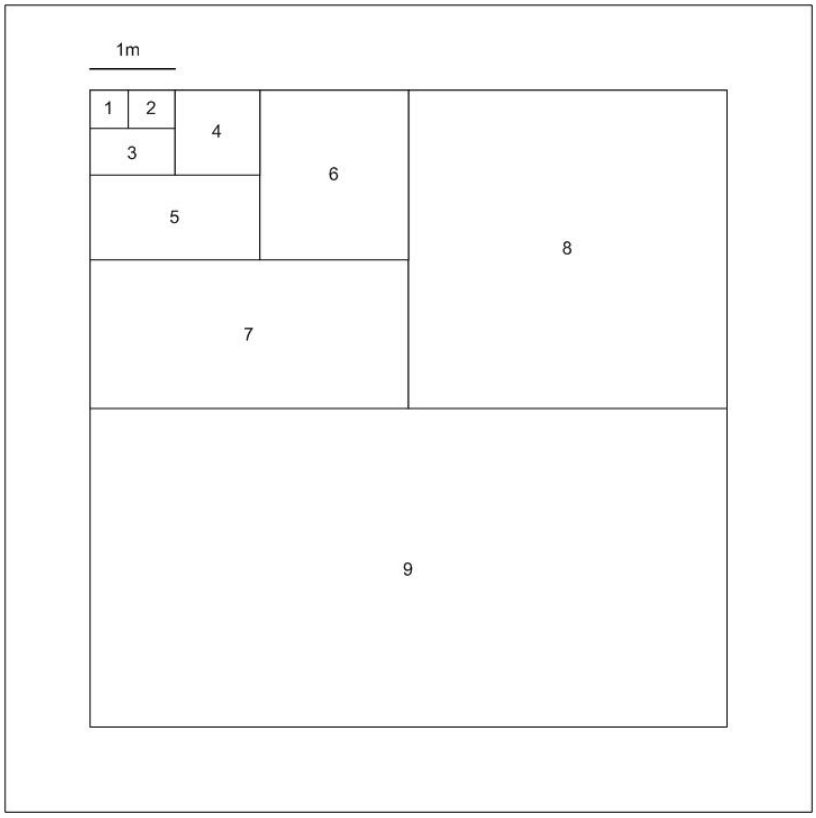
квадрати и охлювите имат ниска плътност, тогава ще има много квадрати с ниски и нулеви стойности. В този случай извадката ще е обективна само ако се вземат голям брой квадрати. С по-голяма репрезентативност би била извадка с по-големи извадъчни единици – така ще се увеличи общата пробна площ и ще се редуцира броят на нулевите стойности. Ситуацията би могла да се усложни още повече, ако охлювите са с групово или равномерно разпределение (Виж по долу Схеми за съставяне на извадка от квадрати - определяне големината и разполагането им в зависимост от пространственото разпределение).

Когато видовете или други изследвани белези се записват като присъствие отсъствие, резултатът може да се представи с честота (срещаемост) (Напр. ако видът *A* се среща в 35 от 100 квадрата, неговата честота е 35%). Големината на пробната площ е достатъчна, когато **стойностите на честотата варират между 20% и 70%.** Стойности под 20% предполагат, че извадъчната единица (напр. квадрат) е твърде малка и би трябвало да се увеличи.

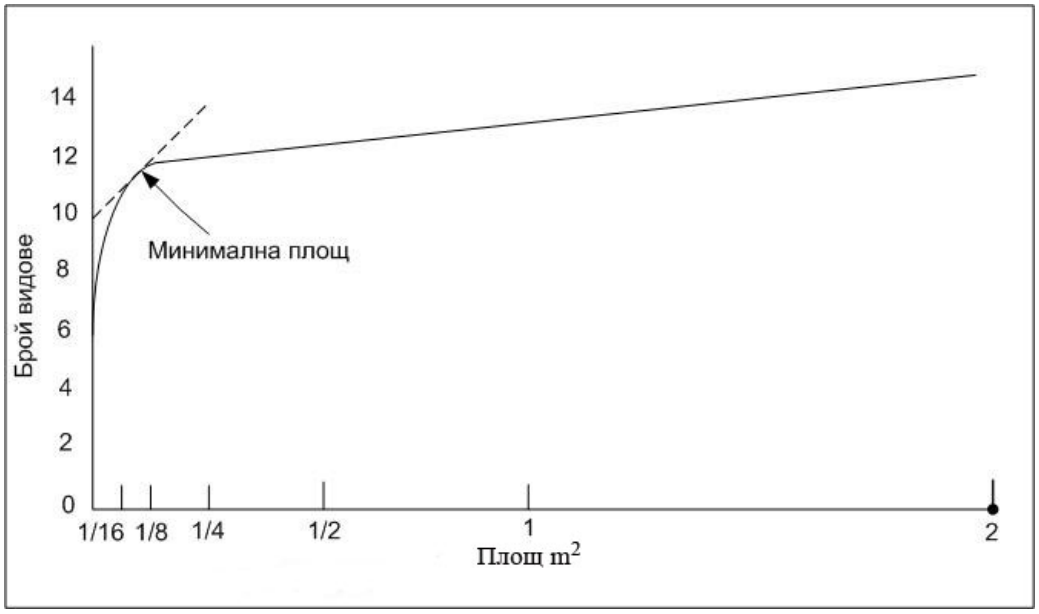
При инвентаризация на видовия състав може да се направи оценка на изискваната **минимална необходима пробна площ**, при която се получава представителна извадка. Това става чрез съставяне на графика, която съпоставя кумулативният брой видове и площта на извадъчната единица. Това е най-малката площ достатъчна за да се получи адекватна и репрезентативна извадка, при която кривата преминава в права (достига плато) или площта, в която могат да се открият 95% от установените видове.

Пример за определяне на минимална необходима пробна площ при растения е показан на Фиг. 4.

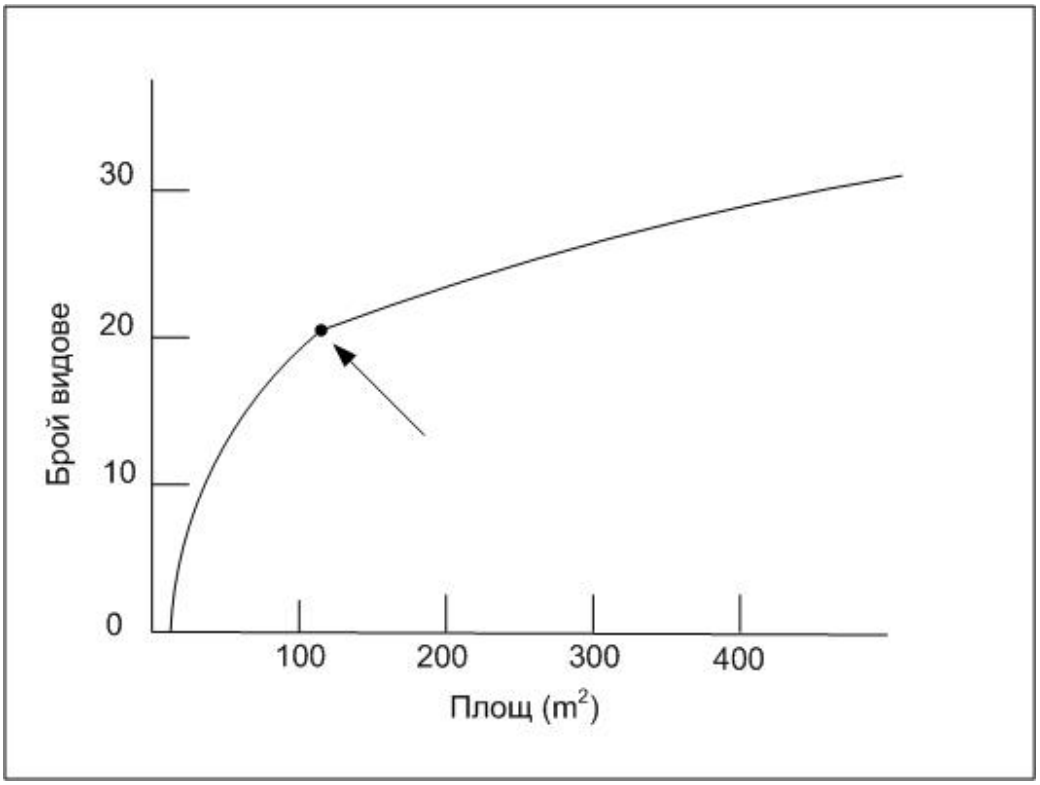
а)


1m
1
2
3
4
5
6
7
8
9

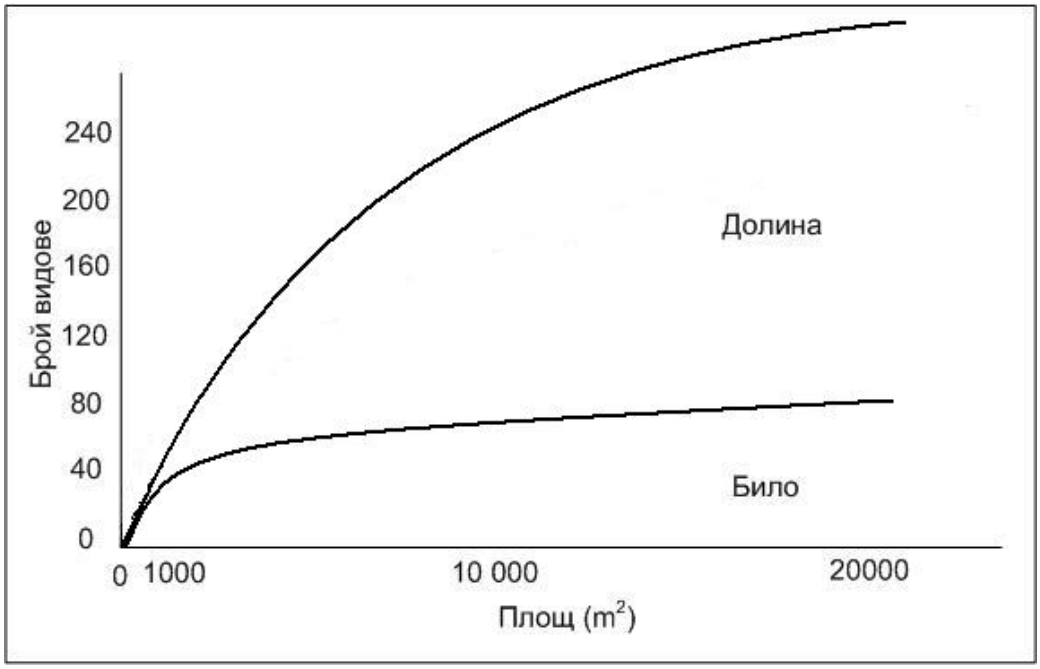
б)


Брой видове
14
12
10
8
6
4
2
0
Минимална площ
1/16
1/8
1/4
1/2
1
2
Площ m2

в)



г)



Фиг. 4. Определяне на минималната необходима площ при различни растителни съобщества. а) Система от вместени един в друг полигони за определяне на

минимална площ на квадрати; б) Графиката е построена за тревиста растителност върху дюни и показва, че минималната площ е 0.13 m^2; в) Графиката е построена за горски съобщества и показва, че минималната площ е около 100 m^2; г) Графиката е построена за две места – долина и било с тропически дъждовни гори и показва, че минималната площ по билото е 1 000 m^2, а в долината 20 000 m^2.

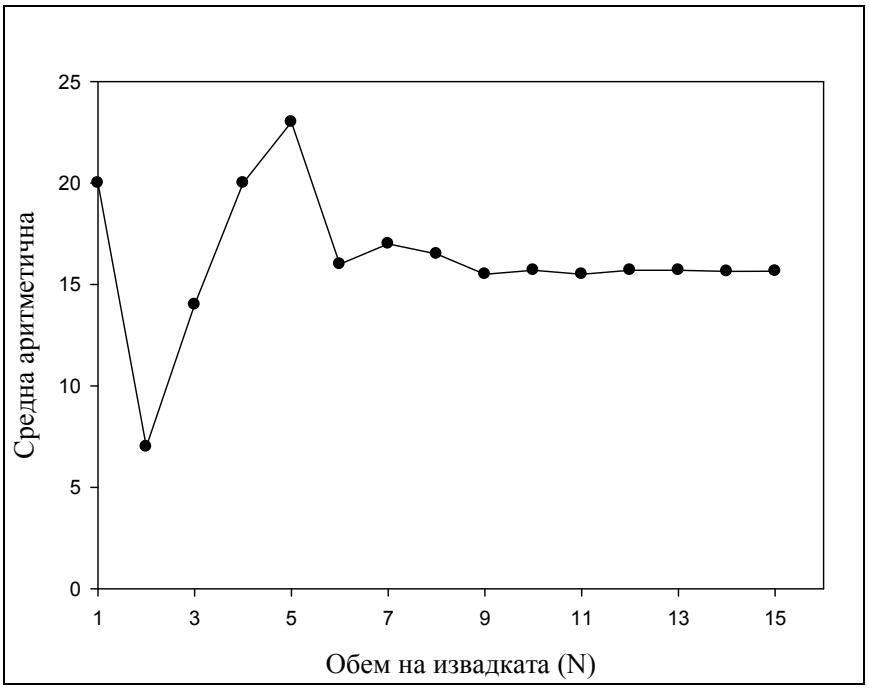
Този подход може да се прилага и в други ситуации, при които се оценява качествения състав на дадено съобщество. Напр. минимален брой капани или обем на извадката се съпоставят по същия начин с кумулативната крива на броя видове (видово богатство).

- *Оценка на необходимия минимален брой индивидуални наблюдения (обем на извадката).*

Данните от пилотните изследвания са най-надеждните средства за установяване на минималната необходима големина на извадката, за да се достигне необходимата прецизност и сила на статистическите тестове.

Използват се графики на съпоставяне на средната аритметична на изследваната променлива с обема на извадката. Могат да се използват и други извадъчни показатели като дисперсия, стандартно отклонение, стандартна грешка на средните, доверителни интервали.

Средните аритметични се калкулират като кумулативни средни аритметични – с увеличаване на обема на извадката и броят на средните се увеличава. Първата стойност от графиката е средната аритметична на първото индивидуално наблюдение, втората точка е стойностите на средната аритметична на индивидуалните наблюдения 1 и 2, третата е за средната аритметична на 1, 2 и 3 и т.н. докато не се използват всички налични стойности. Първоначално стойностите на средната аритметична ще варират значително, но с увеличаване на обема на извадката вариацията ще намалее и стойностите ще се стабилизират. Точката, в която средната аритметична варира по-малко от 10 % се приема, че е подходящият обем на извадката. (Фиг. 5)



Фиг. 5. Крива на средните аритметични за определяне на минималния необходим обем на извадка.

- *Брой на необходимите повторения и избягване на псевдоповторенията*

С оглед да се оцени големината на стандартната грешка, големината на пространствената или времева вариация или пък въздействието на ефекти при провеждане на експерименти, извадъчните единици и съответните измервания трябва да се повторят. Повторението е събиране на информация от повече от една пробна единица.

Идеята за псевдоповторенията е свързана с това, че не винаги считаните за повторения на индивидуални наблюдения са такива в действителност. Съществуват два типа на псевдоповторения:

1. Разширение на изследването извън границите на дадена генерална съвкупност към друга неизследвана генерална съвкупност. Пример: третиране като случайна извадка на квадрати от 1 m^2, разположени на полигон от 1 ha, който от своя страна е случайно разположен в много по-голяма площ на пожарище.

2. Анализ на зависими променливи с методи за независими. Напр. повторни наблюдения на маркирани с предаватели животни, които се третират като случайна извадка от използвани от животното точки на хабитата, въпреки че всъщност наблюденията са по-близко в пространството.

Когато зависимите данни се анализират като независими, обемът на извадката е по-голяма от ефективния брой на независимите наблюдения. Това често дава твърде много достоверни резултати, получени от тестове за достоверност, а доверителните интервали са по-тесни, отколкото би трябвало да са в действителност.

За да се избегне псевдоповторение, добро правило, което да се следва е статистическите изводи да се основават само на 1 стойност от всяко независимо събрано индивидуално наблюдение освен, ако в някои случаи не се обработват с подходящи тестове. Напр. ако 5 квадрата са случайно разположени на изследваната площ, тогава статистическите тестове за това място трябва да се базират на 5 стойности независимо от броя растения, животни, почвени проби и т.н., които са преброени или измерени във всеки квадрант. Подобно и ако се използват 5 радиомаркирани животни, заключенията за популацията би трябвало да се базира на извадка с обем 5, независимо от броя пъти, в които животното е било локализирано.

1.2.2. Схеми за съставяне на извадка.

Изискванията за съставяне на извадка са: разполагането на индивидуалните наблюдения да е случайно и нетенденциозно; индивидуалните наблюдения да са разположени добре сред цялата генерална съвкупност; индивидуалните наблюдения да са независими помежду си.

Таблица. 2. Видове схеми (дизайн) за събиране на извадка:

Схема	Употреба	Предимства	Недостатъци
1.Обикновена случайна	Използва се при относително малки географски райони с хомогенен хабитат, когато не е необходим голям брой извадъчни единици.	Тестовете, необходими за анализиране на данните са най-прости в сравнение с останалите схеми.	По случайност някой места в рамките на генералната съвкупност може да останат неизследвани. Времето за пътуване между извадъчните единици е значително, когато пробната площ и/ или обема на извадката са големи. Ограничената случайна схема за събиране на извадка и систематичната случайна са по-удачни, когато популацията е с групово разпределение.
2.Подредена в слоеве случайна (Stratified)	Използва се, когато изследваните белези реагират много различно на някои ясно дефинирани черти на хабитата. Тъй като включва взимане на случайна извадка в рамките на всеки слой, всеки слой трябва да е съставен от относително малък географски район с хомогенен хабитат и не много голям брой на извадъчните единици във всеки слой.	Дава по-ефективна оценка на популационните параметри, отколкото обикновената случайна, когато изследваните белези варират силно при определени характеристики на хабитата.	Изискват по-сложни анализи в сравнение с обикновената случайна. Когато географският район във всеки слой и/или броят на извадъчните единици са големи, тогава другите видове (по-надолу) схеми за събиране на извадка са по-ефективни. По случайност някои площи в рамките на всеки слой може да останат неизследвани.
Систематична	Използва се при всякакви ситуации, тъй като първата извадъчна	Най-добрият тип схема на събиране на извадка. Осигурява	В случай че броят на възможните индивидуални наблюдения е

Схема	Употреба	Предимства	Недостатъци
	единица е избрана случайно и извадъчните единици са на достатъчно голямо разстояние една от друга, за да се считат за независими. Може да се използва като част от клъстерните и двустепенните схеми.	по-добро разпръсване на извадъчните единици, отколкото обикновената случайна. Данните могат да се съберат значително по-ефективно от колкото при обикновената случайна схема и в същото време да бъдат анализирани със същите формули, както при нея.	ограничен до по-малко от 25-30 може да даде съмнителни резултати. В този случай по-добре да се използва ограничената случайна извадка.
Ограничена случайна	Би трябвало да се използва само когато обемът на извадката ще е по-малко от 25-30, в противен случай систематичната схема е по-добър избор.	Осигурява по-добро разпръсване на извадъчните единици, отколкото обикновената. Данните могат да бъдат анализирани със същите формули, както при случайната.	Не е толкова ефективна, колкото систематичната, когато обемът е над 25-30.
Клъстер	Използва се, когато е трудно или невъзможно да се вземе случайна извадка от отделни елементи. Определя се клъстер от елементи и след това се взима случайна извадка от клъстерите (обикновено използвайки систематичната схема). Така се оценява всеки елемент във всеки клъстер. В мониторинговите програми най-често се използва за оценка на белег относно индивиди (напр. средна височина, брой цветове на растение) При тази ситуация	Често струва по-малко да се вземе извадка от сбор от елементи в клъстер, отколкото извадка от еднакъв брой елементи избрани случайно от генералната съвкупност. В повечето случаи не е практично да се взима случайна извадка от индивиди, освен в редки ситуации. Вместо това белезите се измерват върху всеки индивид в извадка от квадрати (които функционират като клъстери).	Всички елементи във всеки клъстер трябва да се измерят. Ако клъстера се състои от голям брой елементи, двуетапният дизайн е по-ефективен. Трудно е да се определи съотношението между броя и големината на клъстерите. Изисква по-сложни тестове при анализиране.

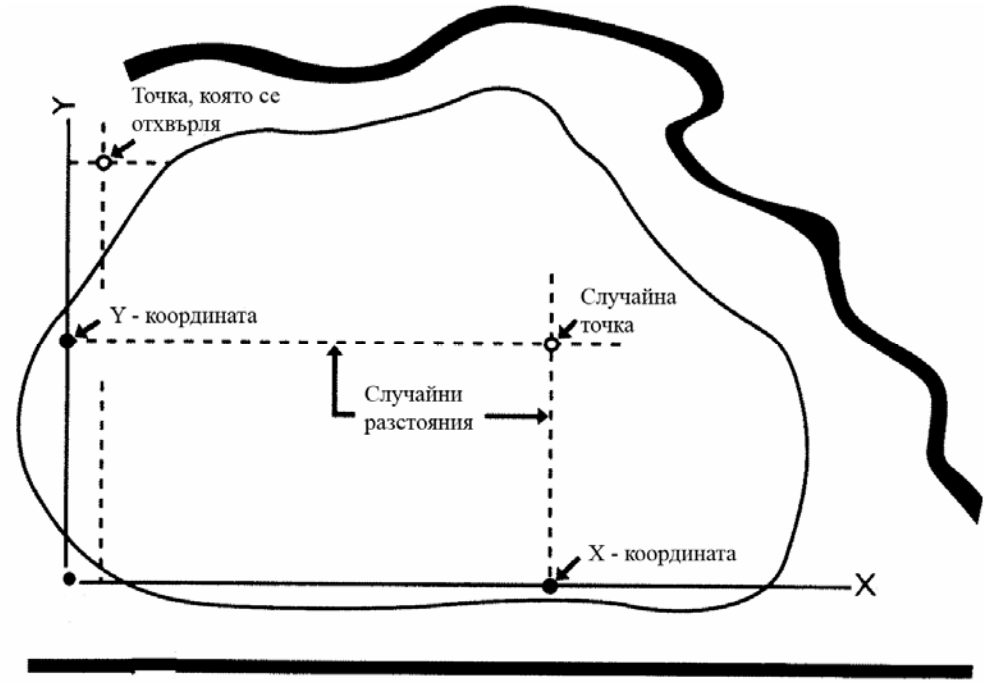
Схема	Употреба	Предимства	Недостатъци
	квадратите са клъстери.		
Двуетапнна	Подобен на клъстерния дизайн при определянето на групи от елементи (напр. растения) и взимане на случайна извадка (обикновено систематична) от тези групи. Тук обаче втората извадка от елементи е взета в рамките на всяка група. Както и при клъстерния дизайн основната употреба е да се оценят дадени белези, свързани с индивиди.	Същите предимства, както при клъстерния дизайн. Двата типа са единствените ефективни средства за оценка на белези, свързани с индивиди. Когато броят на индивидите във всяка група (квадрат) е голям е по-ефективна схема в сравнение с клъстерния.	Има стандартни отклонения, свързани с двата етапа на взимане на проби (за разлика от клъстерния дизайн, където няма стандартно отклонение свързано със стойностите измерени на второто ниво). Това води до по-сложни формули за анализ на стойностите и стандартните грешки (стандартното отклонение на вторичната извадка може да бъде игнорирано при условие, че не се прилага корекционен фактор за ограничена – популация към стандартната грешка на първичната извадка.
Латински квадрати +1	Използва се, когато се цели да се контролира дисперсията в две различни направления.	Дава по-прецизни оценки в сравнение с обикновената случайна и систематизирана схема, когато се определя разпределение на животни, показващи пространствена автокорелация. За разлика от систематичната схема, тук може да се изчисли обективно дисперсията, тъй като има 2 полигона които са случайно избрани.	Сложни формули за оценка на дисперсията.

Схема	Употреба	Предимства	Недостатъци
Двойна	Използва се, когато даден белег (напр. действителна стойност на биомаса) е трудно да се измери Но корелира с допълнителна променлива (напр. окомерна оценка на биомасата), която е по лесна за измерване. Втората променлива се мери в голям брой извадъчни единици, докато първата се мери само в подмножество от извадъчни единици Извадките обикновено се взимат по систематична схема.	Ако допълнителната променлива е достатъчно лесна за измерване и силно свързана с истинската променлива, е много по-ефективна схема от директното измерване на променливата, която ни интересува.	Сложни формули за анализи на данните и определяне на необходимия обем за извадка.
Случайна извадка от индивиди	Използва се в редки случаи. Когато целта е да се измери (преброи) върху индивидуални растения е по-добре да се използва клъстерния или двуетапния дизайн.	В няколкото ситуации, когато може да се вземе такава извадка изчисленията при анализите са по-лесни от тези при клъстерния и двуетапния дизайн.	За целите на мониторинг не е практично да се взима такава извадка.

1.2.2.1. За видове със средно и голямо обилие

- **Обикновена случайна**: 1. всяка комбинация от определен брой индивидуални наблюдения има еднаква вероятност да бъде избрана 2. изборът на което и да е индивидуално наблюдение не е свързан с което и да е друго.

а) *метод на случайни координати* – избират се случайни координати за всяка от двете оси, точката в която се пресичат е мястото на извадъчната единица. Координати, които попадат по границата на изследваната генерална съвкупност се отхвърлят (Фиг. 6).



Фиг. 6. Разполагане на точки по схемата за случайни координати.

Методът е ефективен за малки индивидуални наблюдения като полигони, използвани за отчитане на срещаемост, но не работи добре при индивидуални наблюдения като лентови, линейни и точкови трансекти.

С този метод са свързани два проблема: 1. когато случайната точка е разположена така, че ако индивидуалното наблюдение е с по-голям размер излиза извън границата на целевата генерална съвкупност. Ако такива точки се отхвърлят то резултатите ще са тенденциозни, защото извадъчните единици ще са събрани около

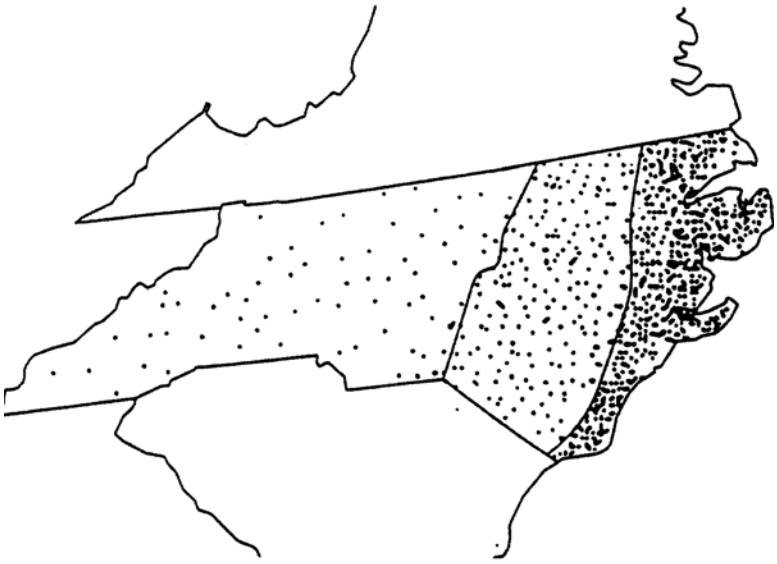
центъра на изследваната площ. 2. има възможност от припокриване на индивидуални наблюдения.

Не са ефективни и при точки, квадрати и полигони използвани за отчитане на биомаса и покритие поради времето, изискващо се да се разположат около 100-200 такива малки индивидуални наблюдения.

б) *метод на клетките с грид* – най-често използван, площта се разделя на концептуални гридове, където 1 клетка отговаря на индивидуално наблюдение.

- **Случайна, организирана в слоеве.** Включва разделяне на генералната съвкупност на две или повече подгрупи (*strata*) преди взимане на извадка. Слоевете се разделят по такъв начин, че извадъчните единици във всеки един да са много сходни, докато между слоевете да са много различни. Във всеки един слой се взима случайна извадка.

Слоевете се дефинират базирайки се на реакцията (на белега, който се следи) към характеристиките на местообитанието, които не се очаква да се променят с времето. Примерни характеристики, използвани за очертаване на слоевете са тип почва, изложение, основен тип растителност (напр. горска, тревиста) и почвена влажност.

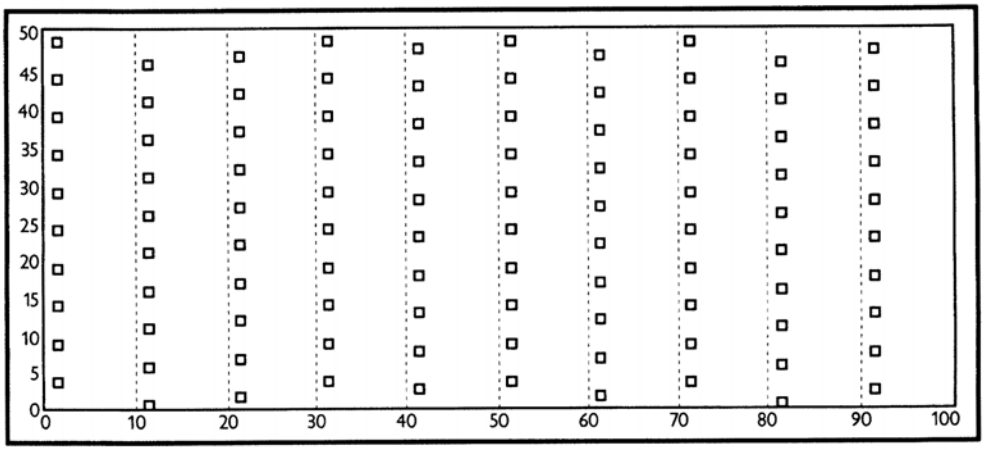


Фиг. 7. Схема на случайна, организирана в слоеве извадка. Различен брой полигони всеки от по 4 ml^2, са разположени в три слоя.

Едно от предимствата на този дизайн е, че броят на извадъчните единици може да варира между слоевете, в случай че изследвания белег реагира различно на различните характеристики на местообитанието (Фиг. 7). Индивидуалните наблюдения може да са разположени равномерно във всеки слой; спрямо големината на всеки слой; спрямо броя на индивидите, които изследваме във всеки слой или спрямо степента на вариране на белега във всеки слой.

- **Систематична със случаен старт.**

Началната точка на симетрично разположение на индивидуалните наблюдения е случайна. Пример. Равномерно разположени индивидуални наблюдения - квадрати по трансекти (Фиг.8).



Фиг. 8. В макроплот 50m x 100m, се взима извадка от 100 квадрата 1m x 1m за определяне на относителна численост. Квадратите са подредени по продължение на трансекти. Както трансектите, така и квадратите са систематично разположени със случаен старт. Стартовите точки на трансектите са по периферията, а на квадратите стартовите точки са по дължина на трансектите.

Тази схема е често използвана в зоологичните изследвания, позволява лесно разпознаване на пробните точки и в повечето случаи, но не винаги осигурява данни със сравнима точност и прецизност като тези на чисто случайната извадка.

Използва се при изследване на растения, за да улесни позиционирането на полигоните за изследване на срещаемост и точките за оценка на покритие. При този подход се поставя основна линия през изследваната съвкупност минаваща или през центъра или по едната й страна. Трансектите, започващи от случайна точка

вървят перпендикулярно на тази линия в двете посоки като посоката, за да е избрана случайно се определя чрез хвърляне на монета.

Ако са добре съставени, систематичните схеми могат да се анализират като обикновена случайна извадка.

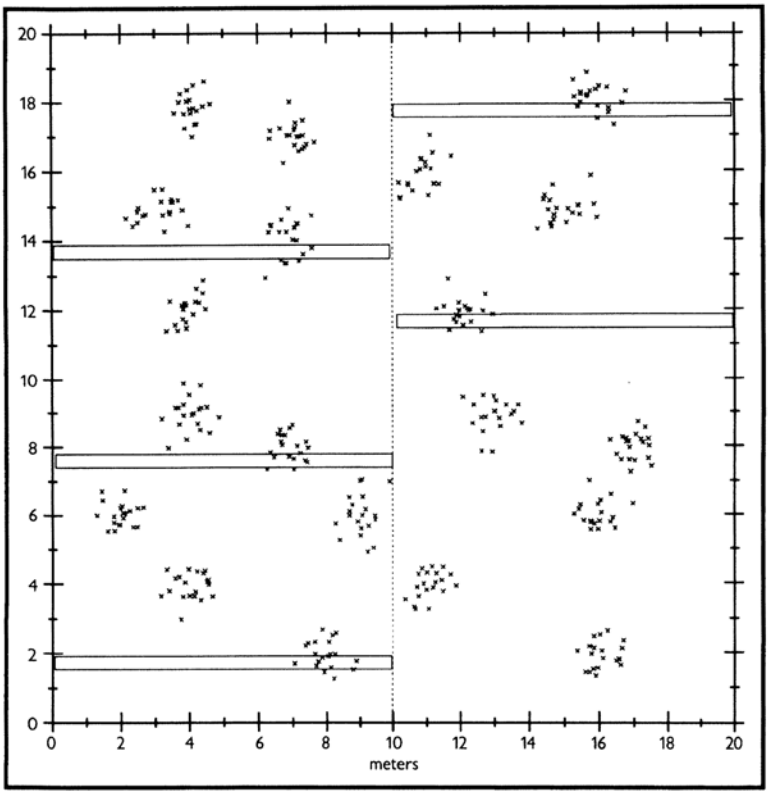
Трябва да се има предвид, че при оценяване на плътност систематичният дизайн на извадка може да доведе до спорни резултати, ако се създаде ситуация, при която има малък брой потенциални извадки. Пример. В един макрополигон се разполагат систематично 10 квадрата 1m x 50m със случайна точка на начало на 2 m от оста X, след което квадратите са разположени на 10 метрови интервали. В този случай тъй като позицията на квадратите вече е фиксирана с поставянето на първия квадрат е възможно да се вземат само 10 извадки, в зависимост коя от 10-те възможни стартови точки е избрана в първия 10 метров сегмент на генералната съвкупност (от 0 до 9). Разпределението на извадките (*sampling distribution*) за този дизайн трябва да е еднакво (плоско) разпределение, а не звънцевидната форма на нормално разпределение, тъй като има само 10 възможни различни извадки и съответно 10 възможни средни аритметични. Отнасянето на такава извадка към случайните ще доведе до лоша оценка на стандартната грешка. Ограничената случайна извадка решава този проблем.

Минималният брой извадки, които трябва да са възможни при систематичната схема на вземане на извадка е 30.

- **Ограничената случайна извадка.** Тук се определя обема на извадката *n*, от които ще има нужда за постигане на целите на мониторинга, след което генералната съвкупност се разделя на *n*-еднакви по размер сегмента. Във всеки от тези сегменти се поставя индивидуално наблюдение (полигон или трансекта) на случаен принцип.

- **Клъстерен дизайн за взимане на извадка**. Използва се, когато е невъзможно да се вземе случайна извадка Определя се клъстер от елементи и след това се взима случайна извадка от клъстерите (обикновено използвайки систематичната схема). След това се измерва всеки елемент на случайно избраните клъстери. Използва се в мониторинга, когато се налага да се установяват показатели относно индивиди като ниво на опаразитяване при животни, среден брой цветове на растение и т.н.

Пример: искаме да проследим височината на растенията X в популацията Y. В популацията има прекалено много растения, за да бъдат измерени лесно всички. На случаен принцип се поставят квадрати (полигони). Измерват се всички индивиди в квадрат (Фиг. 9).



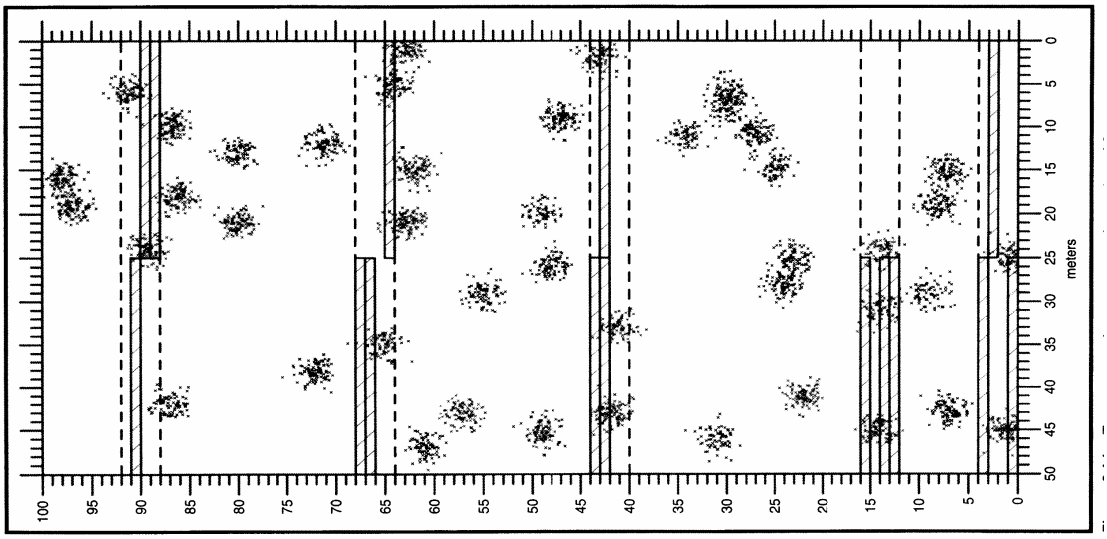
Фиг. 9. Пример на клъстерно събиране на извадка за определяне на средната височина на растенията в дадена популация. 5 квадрата (тесните правоъгълници) са разположени на случаен принцип в генералната съвкупност и във всеки един се измерва дължината на всички растения.

Този и двуетапния дизайн са единствените схеми на пробовземане, които са ефективни при следните примери: установяване на броя продуцирани семена за растение, биомаса за растение, средна височина или размер за растение. Тук квадратът е клъстер, а всяко растение – елемент. В зоологични изследвания примери са – размери на яйцата при птици (гнездата са клъстер, а яйцата – елемент), брой яйца на гнездо (дърветата с гнезда или квадрати са клъстер, а гнездата – елементи), хранителни предпочитания на риби (уловът в 1 гриб

(рибарска мрежа) е клъстер и всеки стомах на риба е елемент), размери на бръмбари (капанът е клъстер, а всеки бръмбар – елемент). Понякога елементите се третират грешно като независими извадъчни единици.

- Двуетапен дизайн за събиране на извадка. Определят се групи от елементи, за които ще се правят анализи. Взима се случайна извадка от групите. В рамките на всяка група се взима случайна или систематична извадка от елементи. Групата се нарича първична извадъчна единица, а елементите са вторични извадъчни единици (Фиг. 10). Тук също се взимат белези на индивидуални организми. Използва се и за увеличаване на прецизността при преброяване на големи бозайници (напр. елени, кози).

Нарича се двуетапен метод защото на първи етап се взима случайна извадка от първични извадъчни единици, а на втория обикновено систематична извадка от вторични извадъчни единици.

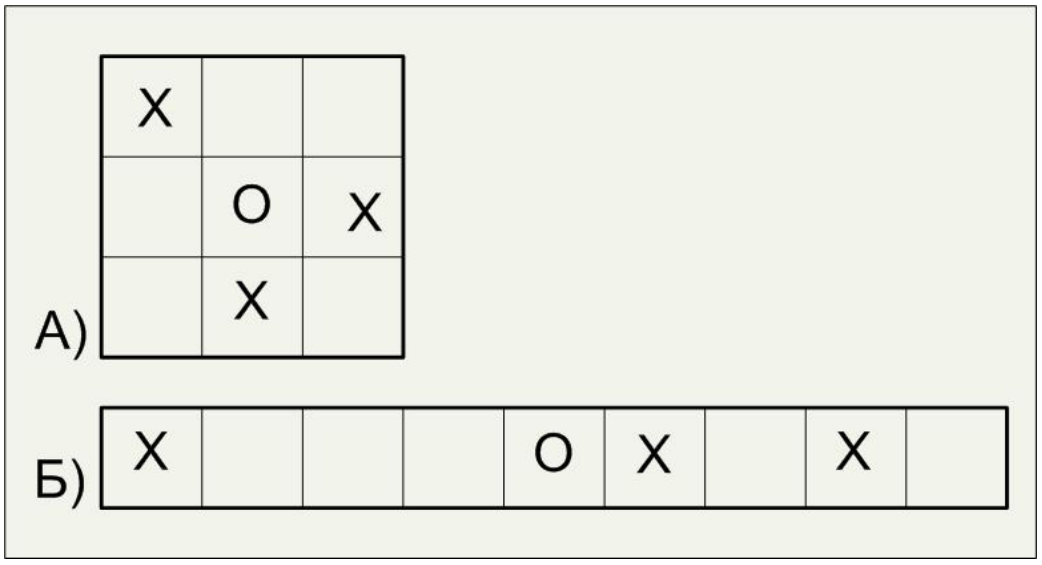


Фиг. 10. Двуетапно взимане на извадка, за определяне броя на цветовете за 1 растение на определен вид растение. Пет полигона 4m x 50m (първични извадъчни единици, правоъгълниците отбелязани с пунктирани линии) са разположени случайно в генералната съвкупност с групово пространствено разпределение и три полигона 1m x 25m (вторични извадъчни единици, защрихованите тесни правоъгълници) са разположени случайно във всеки един от петте големи

полигона. Броят цветове за растение се определя във всеки от по-малките полигони.

- **Латински квадрати +1**. Въведен 90-те години метод за съставяне на извадка, при който всеки един полигон първоначално е избран случайно от всяка уникална комбинация на редове и колони в квадратна рамка, после се избира друг плот от останалите възможности независимо от разположението. Редовете и колоните винаги са еднакъв брой.

Пример: Рамка съдържаща 3x3 квадрата, т.е. 9 клетки, от които се взима извадка с обем 4 (равна на n+1, където n е броя на редовете или колоните). За почва се със случаен избор на първия полигон (отбелязва се с X) от първи ред с три възможности. После втори полигон се избира от 2-те колони, които не съдържат първия полигон. Третия полигон е останалата клетка от трети ред в свободната колона, която не е свързана с колоните заети от първите два полигона. Накрая се избира допълнителния +1 полигон, който се избира на случаен принцип от останалите 6 свободни клетки и се отбелязва с O.(Фиг. 11 А)



Фиг. 11. А) Извадка по схемата Латински квадрати +1 с обем от 4 извадъчни единици взети от квадратна рамка разделена на 9 клетки. Клетките, маркирани с X са първоначално избраните, а с O е елементът „+1”, избран от останалите свободни 6 клетки. Б) Разтеглена форма на първата схемата, за да може да се приложи към линейни хабитати.

За да се прилага този метод не е задължително площите да са под формата на квадрати. Фиг. 11 Б) показва разтеглена форма на Фиг. 11 А), която може да се използва при линейни хабитати, каквито са потоците и реките.

При неправилни форми може да се направи мрежа от по-малки квадрати, които да се напаснат към формата.

- **Двойна схема за извадка**. Нарича се още двуфазна. Включва оценка на две променливи. Използва се, когато белегът, който ни интересува се измерва много трудно. Тогава се взима малка извадка, в която се отчита основния белег, а допълнителния, който се отчита лесно е в много по-голяма извадка. Малките извадки са като подизвадки във всяка извадъчна единица от голямата извадка на допълнителната променлива. След което се прави калибровка.

Пример: Този метод често се използва при определяне на наземна биомаса в тревисти съобщества. Поради това, че е бавно и скъпо да се откъсва, изсушава и да се претегля биомасата на тревистите растения в многобройни извадъчни единици, биомасата се оценява визуално след съответно обучение на наблюдаващите. Определят се напр. 100 случайни квадрати в генералната съвкупност и във всички тях се определя визуално биомасата. В последствие в подизвадка от 10 от тези 100 квадрата биомасата се определя по традиционния начин чрез откъсване, изсушаване и претегляне. Следва калибровка и ако допълнителната променлива корелира силно с основната, тогава прецизността е висока.

Друг пример за употреба е при изследване в гори на обема на дървесината на дадено място. Допълнителната променлива се взима на око, а в подизвадките се повалят дървета, за да се определи точно.

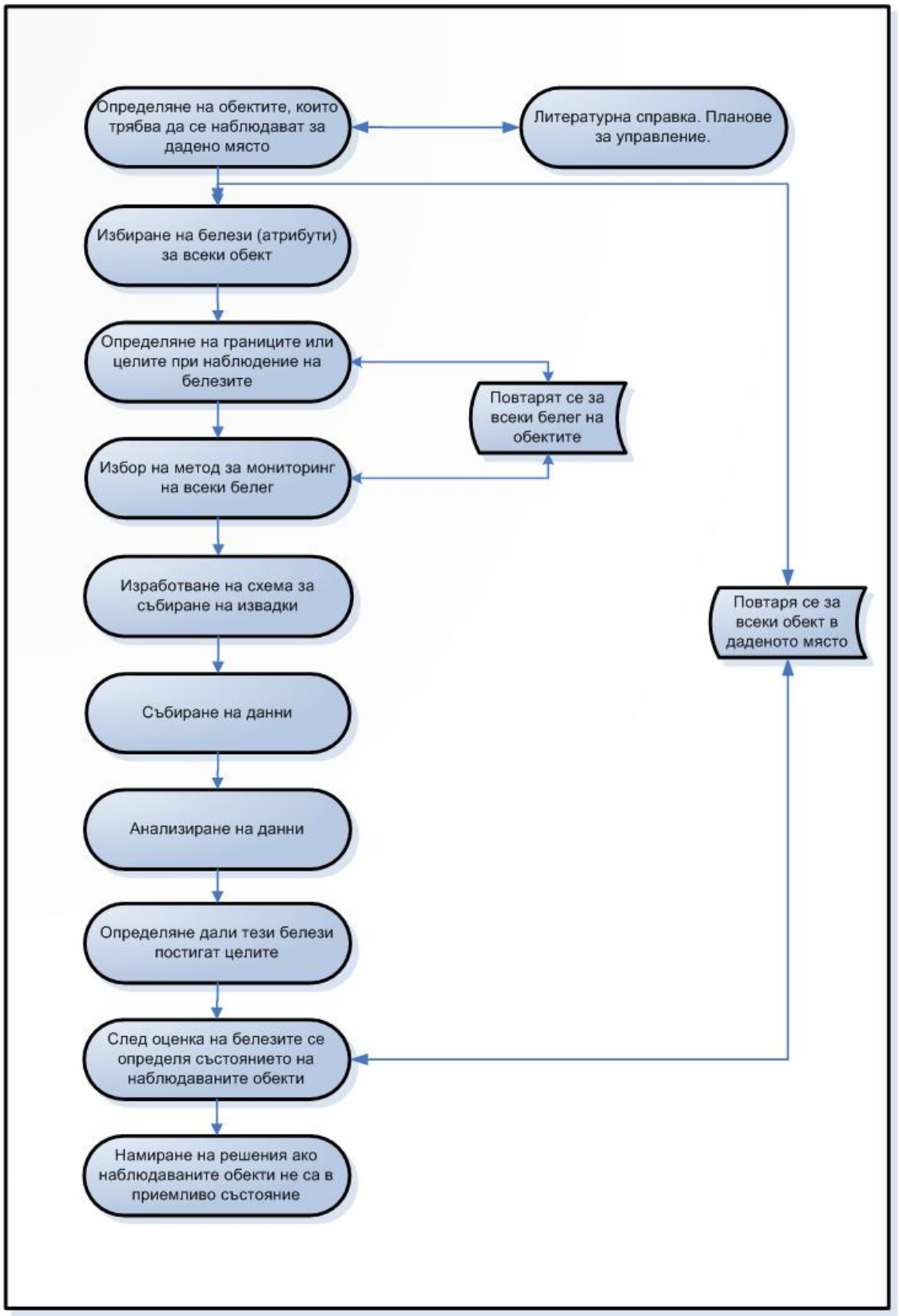
- **Случайна извадка от индивиди** – не се препоръчва да се използва.

1.2.2.2. За редки видове и видове, живеещи на гъсти групи.

Прилага се **адаптивен клъстерен дизайн**: Прави се първоначална случайна схема на пробовземане (обикновена, подредена в слоеве или систематична) , ако в един от полигоните бъде открит индивид от генералната съвкупност, която ни интересува съседните полигони от горе, от долу, от ляво и от дясно също се включват в извадката. Ако някой от околните полигони съдържа индивиди от същия вид, също

се включват полигоните отляво, от дясно, отгоре и отдолу на него и т.н. (в един или повече от тях вече е извършено обследване). Този процес продължава докато в нито един от допълнителните полигони не се намери търсения вид. Сборът от първичните и вторични полигони се нарича *мрежа*. Първичните полигони в които не е открит вида се считат като мрежа от един полигон.

1.2.3. Изработване на мониторингови схеми



Фиг. 12. Алгоритъм при изработване на мониторингови схеми.

Честотата на наблюденията (годишно, на всеки три години и т.н.) зависи от това какви параметри се следят, очакваната степен на промяна (напр. дълголетните видове растения и животни може да се наблюдават по-рядко), дали видът е рядък, от тенденциите в числеността (рискът от загуба на много редките или много застрашени видове е по-висок) и от наличните ресурси за мониторинг.

Мониторингът се състои основно от събиране на качествени и количествени данни за оценка на популациите на видовете.

Събирането на качествени данни обикновено е с по-ниска интензивност от количествените, но при много случаи може да е също така ефективно.

Мониторингът с ниска интензивност (събиране на качествени данни) може да се планира като предупредителна система при възникване на проблем, която да провокира по-интензивен мониторинг или научно изследване. Промените в популациите трябва да са много големи или очевидни, за да бъдат установени при такъв мониторинг. Поради тази причина в случай че бъдат установени промени то е подходящо веднага да се предприемат действия по управление на територията, за защита на популацията на дадения вид.

Примери за мониторинг със събиране на качествени данни:

- *Присъствие – отсъствие на вида*.

- *Оценка на условията в местообитанието*: Неколкократна оценка на качеството на хабитата. Може да се установят видими и големи промени, които могат да се документират със снимки, видео или чрез описание в стандартизиран формуляр.

- *Установяване на размера на популациите*: Визуална оценка на популационния размер, често давана в класове (такива като напр. 0, 1-10, 11-100, 101-1000, 1001+), осигурява повече информация от колкото просто присъствие и отсъствие на вида.

- *Установяване на демографското разпределение*: Това е процент от популацията или брой индивиди, обединени във възрастови класове като ювенилни, възрастни неполовозрели, половозрели възрастни и синилни.

- *Оценка на състоянието на популацията*: Наблюдаващият оценява състоянието на популацията чрез увеличаване или не на територията, болести, хищници, и др. фактори.

- *Фототочки*: Снимки, които се правят от една и съща позиция в една и същ кадър при всяко наблюдение.

- *Фотополигони*: Те са на границата между качествения и количествения мониторинг. Обикновено са подробни снимки от птичи поглед на полигон в рамките на един кадър. Размерът на полигона варира в зависимост от височината на фотоапарата и типа леща, но обикновено варира от 30cm x 30cm до 1m x 1m. Фотополигоните могат да дадат и количествени данни за малка част от популацията, или могат да се използват при измерване на покритие и/или плътност.

- *Картиране на граници*: Картирането на периметъра на дадена популация дава промените в площта, заемана от популацията.

Количественият мониторинг е с висока интензивност и изисква измерване или броене на определен белег. Съществуват основно три типа количествени подхода: 1. Преброяване на цялата генерална съвкупност, 2. Вземане на извадка и 3. Демографски.

При *пълното преброяване*, се преброява или измерва всеки един индивид в целевата популация. Основно предимство на този подход е, че мерната единица е брой, а не показател, базиран на извадка. В този случай не се изисква статистическа обработка за анализиране на текущия статус или промените във времето на наблюдаваната популация.

При *вземането на извадка* се оценява само част от популацията. С извадката е свързана и определена грешка на репрезентативността при оценка на параметрите (това е разликата между извадъчните показатели и истинските параметри на генералната съвкупност). Като средство за оценка на тази грешка се използват статистически анализи (виж по-горе т.1.1). Извадка с количествени данни би трябвало да се взима само ако резултатите ще се анализират статистически, тъй като грешката свързана със събирането на данните може да бъде достатъчно голяма, за което са необходими статистически анализи за интерпретиране на наблюдаваните промени.

Демографският мониторинг включва маркиране и проследяване на съдбата на индивидите през определено време.

При наблюдения на сезонни и годишни цикли е необходимо да се разграничат сезонните промени в популацията от общите тенденции на промяна на изследваните параметри. При дългосрочните програми това може да стане като се взима извадка всяка година в точно определен сезон или при определен стадии от развитието на организмите. Тъй като климатът в тези периоди ще се различава през годините е необходимо да се снемат и тези показатели. При достатъчно събрани данни, обаче с помощта на статистически анализи може да се разграничат сезонните промени от тенденциите. Това става с подходящ обем на извадката и достатъчно време (над 5 години) за диференциране на тенденциите.

Необходимо е схемите да се напасват към историческите промени във времето особено, когато се наблюдава поддържана и управлявана по даден начин защитена територия. Напр. ако се наблюдава някакъв оздравителен процес в дадена гора винаги трябва да има и контрола за да се разграничат промените в резултат от предприетите действия и естествения процес на промяна в местообитанието.

Необходимо е да се контролират и имат предвид случайните вариации в популациите в зависимост от промените в условията на околната среда, които варират непрекъснато. Ефектът на тези вариации би довел до увеличаване на общата грешка. Съществуват статистически анализи за определяне на тези случайни вариации.

1.3. Преглед и оценка на данни от проведен досега мониторинг, с оглед на пригодност за статистическа обработка

1.3.1. Общи положения, засягащи събираната информация от формуляри:

- При извършване на наблюдения е необходимо да се избягва субективна преценка при даването на категории, тъй като различните хора няма да я дадат еднакво и това води до систематична грешка. Примери: в повечето формуляри се определя облачност по степени от 0 до 11 и вятър - силен, умерен, слаб. Препоръчва се точно измерване на вятъра като скорост и посока, след което при обработка на данните може да се обединяват в класове по стойности.

- Повечето видове се наблюдават с няколко различни методики (напр. прилепи в горите се отчитат с комбинирани методи – мрежа, батдетектор и къщички; при мечки – следи и екскременти). Препоръчително е, когато не е направена калибровка на методите, количествените оценки да се правят за всеки отделен метод.
- При използване на извадъчни единици от различен тип, напр. площадки и трансекти, както е при отделните видове земноводни и влечуги, данните за един вид, получени от различни извадъчни единици трябва да се анализират отделно.
- Данните за температура и валежи от формулярите могат да се използват само при статистическа обработка за преценка на климатични условия подходящи за откриване на видовете. Още повече, в повечето случай (с изключение на средно зимно преброяване на птиците) при полева работа се избират добри климатични условия за наблюдения. В случай, че за целите на мониторинга е необходимо да се търси връзка с независими променливи температура и влажност се препоръчва поставяне на влаготермометри за отчитане на средна месечна или годишна температура и относителна влажност на въздуха на наблюдаваните места.

1.3.2 Преглед на формулярите по избрани групи за мониторинг

Висши растения

При спазване на методиката данните са подходящи за статистическа обработка. При малки находища, където е извършено пълно преброяване не е необходима статистическа обработка.

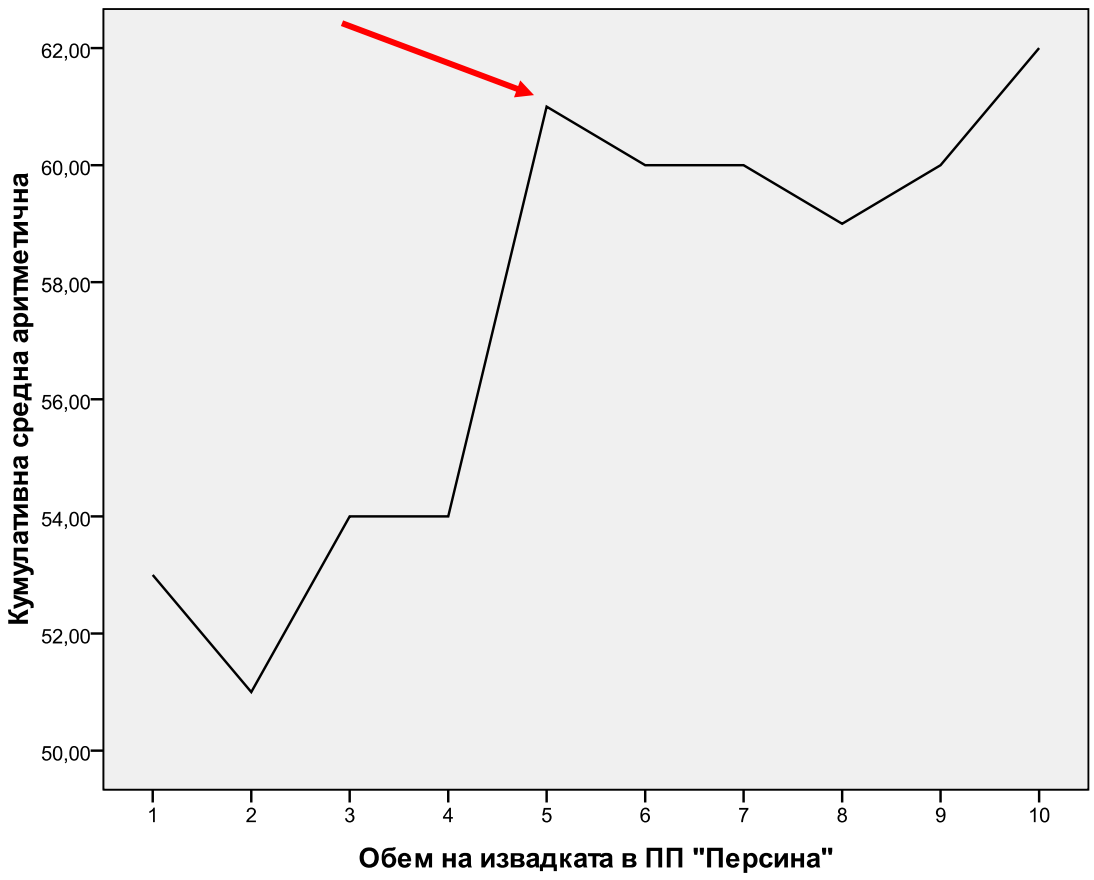
Риби

Прегледани са формулярите за мониторинг на *Neogobius fluviatilis* в ПП „Персина". Различните методи на сбор при отделните точки (индивидуални наблюдения) водят до различна систематична грешка. Едно от правилата при статистическа обработка е методиката на сбор да е еднаква за различните точки. В случая при отделните точки са използвани различни методи – гриб с различна площ и електрориболов. Сравними са само тези данни, които са получени по еднаква методика. В случая трябва да се внимава и за псевдоповторения в точките взимани срещу сградата на ПП.

Ако приемем, че повторенията на пробите са по някоя от стандартните схеми за съставяне на извадка и методите на сбор са еднакви, данните ще са подходящи за статистическа обработка. Тогава е възможно да се направи крива на опита, с която да се определи минималният необходим обем на извадката. От получените данни (Табл. 3) се вижда, че кумулативната средна аритметична на абсолютната дължина на рибите варира под 10% след достигане на обем от 5 индивидуални наблюдения. Приема се, че това е минималният необходим обем (Фиг. 13).

Таблица 3. Данни от мониторинг на *Neogobius fluviatilis* в ПП „Персина", проведен на 18.10.2008 год.

Индивидуални наблюдения	Брой индивиди	Средна дължина (mm)	Кумулативна средна аритметична
1	3	53.67	53.67
2	4	48.50	51.08
3	11	60.36	54.18
4	2	53.50	54.01
5	13	90.54	61.31
6	8	54.50	60.18
7	2	65.00	60.87
8	16	46.19	59.03
9	5	71.00	60.36
10	6	80.67	62.39



Фиг. 13. Крива на ефективността - вариращата с увеличаване на обема средна аритметична на абсолютната дължина на уловените екземпляри *Neogobius fluviatilis* в ПП „Персина". В случая поради ниската вариация на показателя е достатъчен обем от 5 индивидуални наблюдения (точки от водоема).

Подобна крива може да се направи за всеки друг параметър на популацията, който искаме да проследим.

Влечуги и земноводни:

Освен общите препоръки от резултатите да момента се вижда, че голям брой трансекти и площадки са с нулеви стойности. Извадки от 3 трансекти или 2 площадки на по-големите места не са достатъчни за количествена оценка на популацията. Този брой е подходящ само за мониторинг на присъствие и отсъствие на видовете. Препоръчва се да се направи оценка на ефективността с графика за

минимален обем, както и извадъчните единици да са еднакви по форма и размери за място.

Птици:

СЗП, Обикновени, Гнездящи, Мигриращи

Освен общите препоръки при даване на субективни оценки при определяне на категории, всички събрани данни са подходящи за статистическа обработка.

Прилепи:

Освен общите препоръки, от гледна точка на статистическата обработка на данните за получаване на относителна численост се препоръчва къщичките, поставяни в горските хабитати да са разположени в равен брой трансекти в различните местообитания по някои от подходящите схеми (напр. систематична със случаен старт). При налични данни от първите години на мониторинга може да се направи оценка с графика за минимален необходим обем.

Коза:

От данните се вижда, че където има най-голям брой трансекти там е установена и най-голяма численост. Възможно е некоректно анализиране на данните – от статистическа гледна точка броят маршрути може да е различен спрямо изследвана площ (виж извадка организирана в слоеве), но винаги трябва да се прецени дали данните са сравними – оценка на опита. При провеждане на мониторинга 2009-2010 не е спазена методиката: Напр. маршрутите по Централен Балкан са минавани на различни дати в много различни периоди през пролетта и есента и освен това, броят им не е еднакъв в различните периоди на отчет. Анализ на възрастова, полова структура и преживяемост може да се направи при извадките с повече от 50 обекта.

Благороден елен:

При спазване на методиката е възможна обработка на възрастова, полова структура, преживяемост и оценка на риска (при отчитане на цялата популация – демографски жизнени таблици, а при извадка с достатъчен обем – статистически жизнени таблици)

Мечка:

Освен общите препоръки, данните от екскременти на този етап могат да служат само като показател за присъствие на вида. Определяне на брой индивиди по тях не

е коректно. Препоръчва се допълнителна оценка за обема на извадките – брой необходими трансекти. Голяма част от тях са с нулеви стойности.

Вълк:

Освен общите препоръки, за статистическа обработка на данните с цел количествена оценка на популациите се препоръчват данните от видени глутници, следи и бърлоги. Белезите се анализират отделно. Останалите белези могат да се използват само за анализи – присъствие/отсъствие на вида, анализи свързани с биологията на вида (напр. хранителен спектър при екскременти или убити животни). Методиката за мониторинг не предполага целево съставяне на извадки и обработката и анализирането на данните трябва да са съобразени с това.

Лалугер:

Освен общите препоръки за възможност от систематична грешка при определяне на категории за определени белези, всички събрани данни са подходящи за статистическа обработка.

1.3. Литература:

Buckland S. T., D. R. Anderson, K. P. Burnham, J. L. Laakf, D. L. Borchers, L. Thomas (2004) Advanced distance sampling. Oxford University Press. 416 pp.

Elzinga C., D. Salzer, J.W. Willoughby, J.P. Gibbs (2001) Monitoring Plant and Animal Populations. Blackwell Science, Inc. 360 pp.

Gelfand A., P. Diggle, M. Fuentes, P. Guttorp (2010) Handbook of spatial statistics. Chapman & Hall, CRC Press. p. 131-145

Hill D., M. Fasham, G. Tucker, M. Shewry, P. Shaw (2005) Handbook of Biodiversity Methods Survey, Evaluation and Monitoring. Cambridge University Press. p. 3-104; 253-270

Henry P., S. Lengyel, P. Nowicki, R. Julliard, J. Clobert, T. Celik, B. Gruber, D. S. Schmeller,V. Babij, K. Henle (2008) Integrating ongoing biodiversity monitoring: potential benefits and methods. Biodiversity Conservation, 17: 3357–3382

Plattner M., S. Birrer, D. Weber (2004) Data quality in monitoring plant species richness in Swetzerland. Community Ecology 5(1): 135-143

Stohlgren T. (2007) Measuring plant diversity: lessons from the field. Oxford University Press. pp 74-117

Thompson W., G. White, Ch. Gowan (1998) Monitoring Vertebrate Populations. Academic Press Inc. 365pp.

Waite S. (2000) Statistical ecology in practice. Pearson Education Limited. 414 pp.

2. Възможности за статистически анализ с помощта на софтуерния продукт SPSS, закупен в рамките на проект "Разработване на Информационна система към НСМБР в България".

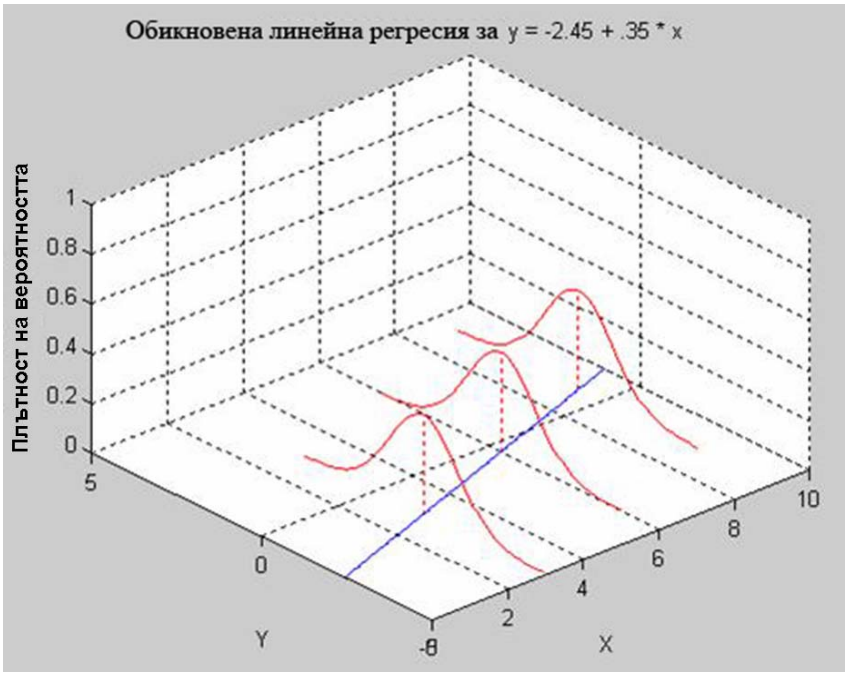
2.1. Цел и задачи на основните статистически методи, включени в следните модули:

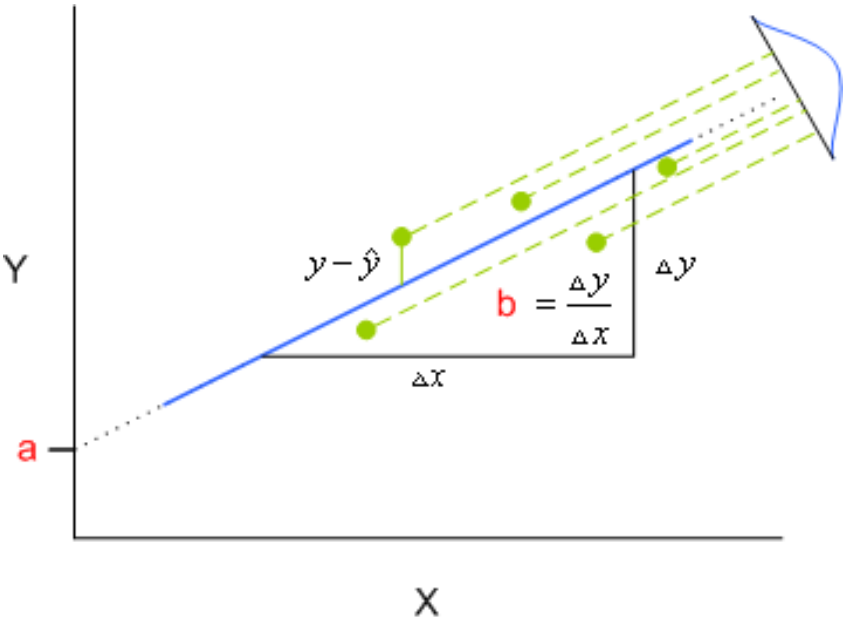
2.1.1 SPSS® Advanced Statistics 18

Модулът **Advanced Statistics** дава продължение на *GLM Univariate* анализа (общ линеен модел с една променлива) в **Statistics Base**. Включва редица процедури, имащи за основа регресионен и дисперсионен анализ. Може да се използва, за да се предвиди дадена стойност на белег в зависимост от стойностите на други „независими" променливи. При тези анализи се изследва връзката на зависимата променлива (белег) (напр. обилие или присъствие отсъствие на вида) с независимите променливи (напр. фактори на средата), както и взаимодействията на самите независими променливи. В зависимост от типа белези, с които се борави се използват различни модели на регресия с различни допускания за вида на данните и техните разпределения.

2.1.1.1. Обикновена и множествена линейна регресия.

Най-елементарният случай на регресия е, когато наличните данни се състоят от една зависима променлива *Y*, и една независима променлива, за която се приема, че се измерва без грешка *X* – Модел I на *Univariate linear regression*.





Фиг. 14. Графика на обикновена линейна регресия, показваща разпределенията на стойностите Y за всяка стойност на X.

Когато X са няколко променливи X_1, X_2…Xp, то тогава се използва модела на *multivariate linear regression* (множествена линейна регресия). Предполагаемият модел за връзката между стойностите на променливите е:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \varepsilon,$$

където ε са отклоненията от линията на модела на стойностите на Y за всяка стойност на X със средна аритметична = нула и константно стандартно отклонение σ. Моделът се оценява чрез изчисляване на коефициентите на стойностите на *X,* които правят отклоненията да са минимални.

Полученото уравнение, на базата, на което се построява правата линия на регресия е:

$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2 + \ldots + b_p X_p$$

където стойностите на *b* са избрани така, че да минимизират сумата на квадратите като грешка:

$$SSE = \Sigma(Y_i - \hat{Y}_i)^2$$

където $\hat{Y}_i$ е стойност дадена от уравнението на регресия, която съответства на стойностите на зависимата променлива Y_i за всяка стойност на X., и сумата е за всички *n* на брой стойности на Y.

Има различни начини за оценка на пасването на уравнението на регресия. Един от тях включва раздробяване на наблюдаваната вариация в стойностите на Y на част, която може да бъде обяснена от стойностите на X, и част която не може да се обясни с X (това е SSE). Общата вариация на стойностите на *Y* се определя от общата сума на квадратите.

$$SST = \Sigma(Y_i - Y)^2$$

Т.е. общата сума на квадратите е сума от сумата на квадратите на отклоненията *residual* (SSE) и сумата на квадратите обяснена с регресията (SSR), така че:

$$SST = SSR + SSE$$

Частта от вариацията на *Y,* обяснена от уравнението на регресия се нарича коефициент на детерминация

$$R^2 = SSR/SST = 1 - SSE/SST$$

което е добър индикатор за ефективността на регресията. При изчисленията се дава и коефициента на корелация R. Стойности на R близки до 1 показват, че правата линия на регресия е добро описание на връзката между X и Y. R = 0 – означава, че

въобще не може да се предвиди стойността на Y от X. При R = 1 идеално може да се предвиди стойността на Y от X

За оценка на пасването на регресионното уравнение се използват няколко подхода. При множествената, както и при обикновената регресия има много процедури за оценка, които могат да се приложат, когато грешката на регресия ε се допуска, че е независима случайна с нормално разпределение, средна аритметична = 0 и постоянна дисперсия σ^2. Тест дали уравнението на регресия обяснява значителна (достоверна) част от общата вариация на *Y* се базира на дисперсионен анализ (ANOVA) и обикновено се дава в таблица. F-тестът сравнява наблюдаваната вариация в Y, обяснена от уравнението на регресия с вариацията поради случайни отклонения (*residual*). От тази таблица, може да се изчисли стойността F,

$$\mathbf{F = MSR/MSE = (SSR/}p\mathbf{)/[SSE/(}n - p - 1\mathbf{)]}$$

и да се сравни с табличното F-разпределение при определено *p* и степени на свобода (df) $n - p - 1$. Ако получената стойност на F е по-голяма от табличната „критична" стойност, тогава има достоверно доказателство, че *Y* е свързана с поне една от променливите *X*. Програмата обикновено направо изчислява P, когато е <0.05 уравнението на регресия пасва на данните достоверно.

Таблица 4. Таблица на Дисперсионен анализ (ANOVA) за Множествена регресия

Източник на вариация	**Сума на квадратите**	**Степени на свобода (df)**	**Mean Square**	**F**
Регресия	SSR	p	MSR	MSR/MSE
Грешка	SSE	$n - p - 1$	MSE	
Общо	SST	$n - 1$		

Изчислените регресионни коефициенти също могат да се тестват поединично, за да се види дали са достоверно различни от нула. Ако дори един от коефициентите не е достоверно различен от нула, тогава няма доказателство, че Y стойностите са свързани с X. Използва се t-тест, с който се проверява дали βj е статистически достоверно различна стойност от 0. Нулевата хипотеза гласи, че коефициентът b пред независимата променлива е 0 и той не води до предвиждане на стойностите на

зависимата променлива. t е отношението на регресионния коефициент b към неговата стандартна грешка:

$$t = \frac{b_j}{SK(b_j)}$$

където SK(*bj*) е стандартната грешка на *bj*. Тази стойност се сравнява с табличната стойност на t-разпределението при $n - p - 1$ степени на свобода (програмата изчислява стойностите на вероятността да сгрешим, отхвърляйки нулевата хипотеза – *p*). Нулевата хипотеза се отхвърля, когато получената стойност на t е > от табличната, както и когато изчисленото от програмата *p* е < α (<0.05; 0.01 или 0.001). Ако *bj*/SK(*bj*) е статистически достоверно различна от нула, то тогава има доказателство че β*j* е различна от нула.

В допълнение може да се изчисли точността на получените регресионни коефициенти *bj* чрез изчисляване на доверителните интервали при 95% ниво на достоверност. Доверителните интервали за β*j* са *bj* ± t5%,*n*−*p*−1*bj*/SK(*bj*), където t5%,*n*−*p*−1 е абсолютната стойност, която е > от вероятността 0.05 за t-разпределение със степени на свобода $n - p - 1$.

Ако независимите променливи при множествена регресия са подредени по важност X_1 до *Xp*, тогава е полезно да се напасват регресионните уравнения, отнасяйки *Y* към X_1; *Y* към X_1 и X_2, и т.н. докато *Y* бъде свързан с всички *X* променливи. Вариацията в *Y,* свързана с *Xj* като се имат предвид ефектите на променливите от X_1 до X_{j-1} се дава от т.нар. *extra сума на квадратите* при добавяне на X_j към модела.

$$SSR(X_j, X_1, X_2, \ldots, X_{j-1}) = SSR(X_1, X_2, \ldots, X_j) - SSR(X_1, X_2, \ldots, X_{j-1})$$

Това позволява отново да се направи F-статистика за това дали X_j е статистически достоверно свързана с Y. Ако променливите *X* не корелират помежду си, стойностите на F ще останат същите независимо от подредбата на независимите променливи, включени в регресията. Обикновено обаче независимите променливи корелират по между си и подредбата им в модела може да е от съществено значение. Това означава, че при множествената регресия с корелиращи помежду си независими променливи, когато се анализират резултатите за взаимоотношенията

между *Y* и *Xj* задължително трябва да е в контекста на това, че в уравнението присъстват и другите независими променливи в същото време.

В уравнението на линейна регресия обикновено се включва и т. нар. intercept - коефициент *a*. Това е точката, в която линията на регресия пресича оста Y (фиг. 13).

2.1.1.2. Общ линеен модел - *General linear model (GLM)*. Представлява линеен статистически модел с изходна променлива с количествени непрекъснати величини (*quantitative, scaled*) и две или повече независими променливи (т. нар. *predictor*) най-малкото едната, от които е качествен номинален белег (*nominal, non-scaled*). В модела се включват следните анализи: ANOVA (Дисперсионен анализ), ANCOVA (анализ на промяната в съвместната дисперсия (вариацията) на две променливи в един и същи времеви период) или MANOVA, MANCOVA, обикновена линейна регресия, t-тест и F-тест.

Ако има само една колона в Y (т.е. една зависима променлива), тогава моделът може да бъде отнесен също и към множествената линейна регресия *multiple regression model* (*multiple linear regression*).

Общият случай на *GLM* е, когато ***Y*** и ***ε*** са няколко колони, т.е. *GLM* представлява разширение на линейната множествена регресия за една зависима променлива.

Условия за прилагане на този модел са извадката да е случайна; дисперсиите да са еднакви; и ако не нормално, то поне симетрично разпределение на стойностите на белезите. (Проверяват се предварително с тестове и графики).

Общото уравнение на модела е следното:

$$\mathbf{Y = x\beta + \varepsilon}$$

където ***Y*** е матрица с редици от много на брой наблюдения (измервания), ***X*** е т.нар. *design matrix* (матрица с независимите променливи – напр. индекси с 0 и 1, показващи принадлежност към дадена група, време и т.н.), ***β*** е матрица, съдържаща параметри, които не са известни и ще се оценяват с модела (дефинират приноса на всеки компонент от матрицата, съдържаща стойностите на **X** към стойностите на **Y**). Изчисляват се така, че да минимизират грешката - ***ε***, която представлява матрица, съдържаща т.нар. *residuals* (отклонения - разликата между наблюдаваните стойности на Y за всяка стойност на X и тези предсказани от модела). Стойностите

на отклоненията в модела трябва да имат нормално разпределение. Ако не са с нормално разпределение, се използват други линейни модели, отнасяни към обобщения линеен модел (*generalized linear models*).

$$\begin{pmatrix} Y_1 \\ Y_j \\ Y_J \end{pmatrix} = \begin{pmatrix} X_{11} \cdots X_{1i} \cdots X_{1I} \\ X_{j1} \cdots X_{ji} \cdots X_{jI} \\ X_{J1} \cdots X_{Ji} \cdots X_{JI} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_j \\ \beta_J \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_j \\ \varepsilon_J \end{pmatrix}$$

Тестването на хипотези с *GLM* може да се прави по два начина: множествен (*multivariate*) *GLM* или под формата на няколко отделни теста с по една зависима променлива (*univariate*).

Вход на данните в SPSS: Зависимата променлива представлява количествен белег. Променливите, които се посочват като фактори са качествени белези (categorical). Те могат да имат числени стойности или качествени състояния. Ковариращите променливи (Covariates, predictors) са количествени белези свързани със зависимата променлива.

GLM Univariate най-общо представлява регресионен и дисперсионен анализ на една зависима променлива (белег) (напр. обилие на вида) спрямо един или повече фактори (напр. фактори на средата). ***GLM Multivariate*** е разширение на *GLM Univariate* и позволява изследване на няколко зависими променливи. При множествения *GLM* всички колони на ***Y*** се тестват заедно, докато при *univariate* тестове, колоните на ***Y*** се тестват независимо, т.е. като многократни *univariate* тестове с един и същи дизайн на матрицата. ***GLM Repeated Measures*** от своя страна представлява разширение на *GLM Multivariate* и позволява изследване на няколко зависими променливи с определен брой повторения на едни и същи измервания (наблюдения).

Факторните променливи разделят стойностите на зависимата променлива на групи. С *GLM* се тества следната нулева хипотеза: дали има ефект на независимите променливи върху средните аритметични на различните групи на зависимата променлива. Може да се изследват и взаимоотношенията между различните

фактори, както и ефекта на всеки един от тях върху зависимата променлива (тъй като някои фактори могат да са случайни). Също така може да се видят и ефектите на ковариране на факторите.

В резултат могат да се предвиждат стойности на зависимата променлива при определени стойности на независимите променливи.

С *GLM* могат да се тестват т. нар. балансирани и небалансирани модели. Дизайнът на данните е балансиран, ако всяка матрица в модела има един и същ брой случаи.

След като дисперсионният анализ (*F* test) покаже достоверни разлики, може да се приложи и т. нар. съпоставяне (*contrasts*), както и да се използват допълнителни тестове (*post hoc tests*), за да се установят разликите между средните аритметични на конкретни променливи. Периферните (*marginal*) средни дават оценка на предвижданите средни стойности за всяка от матриците в модела, и графиките на тези средни (*profile plots = interaction plots*) позволяват да се визуализират някои от връзките между променливите.

Типове съпоставяне (*contrasts*):

Deviation – Сравнява средните аритметични за всяка стойност (качествено състояние) на фактора (освен reference category) със средната за всички стойности. Стойностите на факторите може да са подредени всякак. В програмата под *factor levels* се имат предвид стойностите на факторите.

Simple – Сравнява средните аритметични за всяко стойност на фактора със средната за определена стойност на същия. Този тип се използва обикновено, когато има контролна група.

Difference – Сравнява средната за всяка стойност на фактора (освен първата) със средната за предишните стойности.

Helmert – Сравнява средните за всяка стойност на фактора (освен последната) със средната за следващите стойности.

Repeated – Сравнява средните за всяка стойност на фактора (освен последната) със средната за следващите стойности.

Polynomial – Сравнява линейния ефект, квадратичния ефект, кубичния ефект и т. н.. Първата степен на свобода дава линейния ефект измежду всички категории,

втората квадратичния ефект и т.н. Често се използва за оценка на полиномни тенденции.

WLS Weight позволява да се уточни променлива, която дава различна тежест на отчитаните стойности в анализа *weighted least-squares* (*WLS*), когато се налага да се компенсира различна прецизност на събиране на данните. Ако в даващата тежест променлива има стойности, които са нула, отрицателни или липсват, се изключват от анализа. Тази опция не е задължителна.

Когато се избира типа модел, трябва да е в зависимост от типа на данните. Пълният факторен модел съдържа всички основни ефекти на факторите, всички основни ефекти на ковариращите променливи и всички взаимодействия между факторите. Съществува опция *Custom*, с която се уточняват интересуващите ни взаимодействия и всички условия, които да се включат в модела.

За оценка на различните хипотези се използват съответно Тип I, Тип II, Тип III и Тип IV сума на квадратите. Тип III е по подразбиране и се използва при балансирани или при небалансирани модели, когато няма липсващи стойности.

В модела обикновено се включва и т. нар. *intercept*. Тогава уравнението има следния вид:

$$\mathbf{Y = \beta x + c + \varepsilon}$$

Intercept е стойността, в която правата линия на модела пресича у-ординатата на графиката. Може и да не се включва, когато графиката започва от 0.

Допълнителните *(post hoc)* тестове се правят с цел да се види по двойки разликата между различните средни стойности. Използват се само за фиксирани т. нар. *between-subjects* фактори. При *GLM Repeated Measures*, тези тестове не са активни, ако няма такива фактори и се използват за средните за различните нива на т. нар. *within-subjects* фактори. При *GLM Multivariate*, тези тестове се правят за всяка зависима променлива поотделно.

Изборът на конкретен тест зависи от типа на данните и информацията, която ни интересува. Налични са следните тестове: най-малка достоверна разлика (LSD), Bonferroni, Sidak, Scheffé, Ryan-Einot-Gabriel-Welsch множествен *F*, Ryan-Einot-Gabriel-Welsch multiple range (R-E-G-W), Student-Newman-Keuls, Tukey's honestly significant difference, Tukey's *b*, Duncan, Hochberg's GT2, Gabriel, Waller-Duncan *t*

test, Dunnett (one-sided and two-sided), Tamhane's T2, Dunnett's T3, Games-Howell, и Dunnett's *C*. Тест на Levene за хомогенност на дисперсията. Изчисляват се и показатели от дескриптивната статистика като средни аритметични, стандартни отклонения, обем (брой индивидуални наблюдения) на всяка от зависимите променливи.

Тестовете на Bonferroni и Tukey са често използвани параметрични тестове (изискват еднаквост на дисперсиите и нормалност на разпределението на стойностите). Те са базирани на t-теста на Student и правят сравнение по двойки на средните на всички променливи. При по-голям брой променливи, тестът на Tukey е с по-голяма сила от Bonferroni, а при малко на брой променливи Bonferroni е с по-голяма сила. Тестът на Dunnett е вид t-тест, който сравнява няколко средни с една контролна средна. Последната категория по подразбиране е контролната категория, може да избере и първата категория за контролна.

LSD - тестът е еквивалент на множество индивидуални t-тестове между всички двойки групи. Тестът R-E-G-W, тестът на Duncan и тестът на Student-Newman-Keuls са множествени процедури. Първият е с по-голяма сила, но при неравен обем на матриците се препоръчват другите два (рангови, непараметрични тестове – подреждат средните по ранг и изчисляват най-малка и най-голяма разлика). Когато не е спазено изискването за еднаквост на дисперсиите се използват Tamhane's T2, Dunnett's T3, Dunnett's C и Games-Howell тестовете. Те не са валидни и не се правят, ако в модела има много фактори. Рангов тест е и Waller-Duncan t-тест, който използва т.нар. Bayesian подход, по името на Thomas Bayes. Използва средната хармонична на обема на извадките, когато не са еднакви. Тестът на Scheffé е с по-малка сила, в сравнение с останалите.

Обратна връзка - *GLM multivariate*:

Резултатите (OUTPUT) се представят в няколко таблици и графики:

В първата таблица – *Box's test* – тества нулевата хипотеза, че съвместната дисперсия на матрицата от стойности на зависимите променливи е еднаква измежду всички групи. $P<0.05$ означава, че условието за еднакви дисперсии не е спазено, и че е възможно резултатите от модела да са подвеждащи. $P>0.05$ означава, че нулевата хипотеза остава в сила. Точен е при големи обеми на

извадката и при нормално разпределение на стойностите. За това се прави допълнителен тест:

Тест на *Levene* за хомогенност на дисперсията на отклоненията *ε*. Тъй като този тест също е чувствителен само при големи обеми на извадката, условието може да се провери и визуално с графиките *spread vs. level plot*.

При неспазване на тези условия от променливите се налага трансформация на данните или друг модел.

В таблицата *Multivariate tests* се дават резултатите от 4 теста за достоверност на всеки ефект в модела.

- Pillai's trace – колкото е по-висока стойността, толкова по-голям ефект в модела оказва дадена независима.
- Wilks' Lambda. Стойностите му варират от 0 до 1. Колкото е по-близо до 1 стойността, толкова по-малък ефект има независимата.
- Hotelling's trace – сумата на eigenvalues в матрицата. Колкото е по-висока стойността, толкова по-голям ефект в модела оказва дадена независима.

Когато стойностите на Hotelling's trace и Pillai's trace, са приблизително еднакви това означава, че ефектът на независимата променлива по всяка вероятност не допринася много за модела.

- Roy's largest root е най-голямата eigenvalue стойност. Колкото е по-висока стойността, толкова по-голям ефект в модела оказва дадена независима. Когато тази стойност е равна на Hotelling's trace, ефектът е свързан предимно само с една от зависимите променливи, има силна корелация между зависимите променливи или ефектът не допринася много за модела.

В таблицата е включена и F- статистика, показваща достоверността на ефекта на независимите променливи, както и на ефекта на тяхното взаимодействие.

- Partial eta squared – по-големи стойности показват по-голяма част от вариацията обяснена с модела. (макс. стойност е 1)

В таблица *Between subject effects* се дават сумите на квадратите и F-статистиката за всички независими променливи.

В таблица *Between subjects SSCP* се дават матрици на хипотезите (за факторите по отделно и тяхното взаимодействие), които се тестват и грешката на сумата на

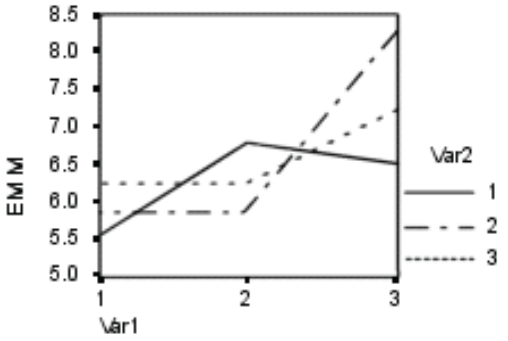
квадратите за тестване на ефектите в модела. В зависимост от това колко зависими променливи се изследват, всяка матрица съдържа толкова редове и толкова колони.

Таблица за съпоставяне нивата на факторите - т.нар. contrasts. От тук може да се види приноса на всяко ниво. При P< 0.05 може да се заключи че разликата не е случайна (е достоверна).

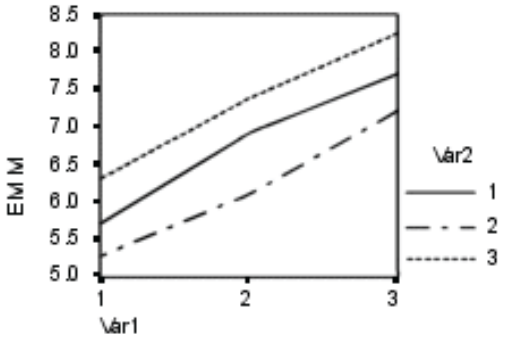
Profile plots (interaction plots) представлява графика с линии, всяка една, показваща получената т.нар. *marginal* средна на зависимата променлива за всяка независима променлива (covariate) при всички стойности на даден фактор (качествена незвасима). При анализите с много зависими променливи се правят отделни профилни графики за всяка зависима променлива. Тези графики за даден фактор показват дали т.нар. маргинални средни нарастват или намаляват при различните стойности на фактора (т.е. нагледно взаимоотношенията на всяка зависима с всяка независима променлива). При два и повече фактора, успоредните линии на графиката показват липса на взаимодействие между факторите. Когато линиите не са успоредни означава, че има взаимодействие.

Фиг. 15. Профилни графики.

Има взаимодействие между факторите



Няма взаимодействие между факторите



2.1.1.3. Компонентен анализ на дисперсията - *Variance Components Analysis* е вид дисперсионен анализ, който се използва, когато факторите (независимите променливи) са подредени йерархично. При такова подреждане един фактор е „вместен" в друг фактор (т.е. е част от друг фактор). При този анализ дисперсията на зависимата променлива се разлага на фиксирани и случайни компоненти.

Използва се за оценка на приноса на всеки случаен ефект към дисперсията на зависимата променлива в смесените модели (които изследват независими променливи даващи случайни или фиксирани ефекти върху независимата). Когато се определят компонентите на дисперсията, може да се прецени къде да се обърне внимание с цел да се редуцира дисперсията. Програмата предлага четири метода за такъв анализ: *minimum norm quadratic unbiased estimator* (MINQUE), дисперсионен анализ (ANOVA), максимална правдоподобност (ML), ограничена максимална правдоподобност (REML).

Пример 1. за йерархично подреждане на фактори е при извадка организирана в слоеве. Напр. един блок от даден терен е разделен на три площадки и във всяка от тях се отчита даден параметър от случайни места. Всяка площадка се разделя на субплощадки и в тях отново на случаен принцип се отчита втори параметър. За да се получат повторения всичко се повтаря в няколко блока.

Пример 2. отчетено е натрупването на телесна маса на прасета в 6 различни прасила за един месец. Променливата – прасило е случаен фактор с 6 нива. (6-те прасила са случайна извадка от голяма съвкупност от прасила). В резултат на такъв тест се доказва че варирането в теглото (= дисперсията на белега тегло) може да се обясни повече с различията в прасилата, отколкото с различията на прасенцата в едно прасило.

Вход на данните в SPSS: Зависимата променлива е количествен белег. Факторите са качествени белези. Те могат да имат числени стойности или буквени. Най-малко един от факторите трябва да е случаен. *Covariates* – са независими количествени променливи, които са свързани със зависимата променлива.

Изисквания за тези анализи са моделните параметри на случайния ефект да имат средни аритметични = 0 и ограничени константни дисперсии, и да няма корелация между тях. ANOVA и MINQUE не изискват нормалност на разпределението. ML и REML изискват нормално разпределение на параметрите и случайния ефект.

2.1.1.4. Смесени линейни модели - *Linear Mixed Models* разширява GLM, така че да допуска данните, които се анализират да показват ковариране (съвместна дисперсия) и непостоянно вариране. Така се моделират не само средните

аритметични на данните, но и дисперсията (варирането) и съвместното вариране (*covariance*).

Mixed models се разглеждат като линеен модел с фиксиран ефект на независимите променливи **β**, случаен ефект **U**, и ефект на случайните отклонения **ε**.

Уравнението на модела е:

$$\mathbf{Y = X\beta + Zu + \varepsilon}$$

Където: **Y** = зависимата променлива (данни за обилие, плътност и т.н.); **β** = регресионни коефициенти пред независими променливи; **X** = стойности на променливата с фиксирани ефекти, **Z** = *design* матрица, която свързва **Y** с **U**; **U** = стойности на променливите със случайни ефекти със средна аритметична = 0 и матрица **G** на *variance-covariance* (дисперсия-съвместна дисперсия), **ε** = стойности на отклоненията *residuals* със средна аритметична = 0 и *variance-covariance* матрица **R.**

2.1.1.5. Обобщени линейни модели - *Generalized Linear Models (GENLIN)*. Използват се за предвиждане на стойности на зависими променливи с дискретни разпределения на стойностите – меристични белези - брой (с Поасоново разпределение на отклоненията), бинарни данни – 0 и 1 (с Биномно разпределение на отклоненията), пропорции, съотношения и индекси (с Биномно разпределение на отклоненията), данни с постоянен коефициент на вариация (с Гама разпределение на отклоненията), анализи на преживяемост (с експоненциално разпределение на отклоненията) и на такива зависими променливи, които имат нелинейна връзка с независимите променливи. На практика повечето данни от екологични изследвания не са с нормално разпределение на отклоненията (грешките), което прави GLM по-малко приложим и на негово място се използват *GENLIN* моделите, които включват и самия GLM като опция.

GENLIN моделите включват различни регресионни модели, използвани за анализиране на екологични данни.

При обобщените линейни модели зависимата променлива Y, е свързана с независимите променливи X_1, X_2, …, X_p под формата на следното уравнение:

$$\mathbf{Y = f\,(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p) + \varepsilon}$$

където f(x) е една от няколкото позволени функции, а ε е отклонения *residual* със средна аритметична = 0 и едно от няколкото позволени разпределения на стойностите.

Например ако f(x) = x и разпределението на стойностите на ε е нормално, то формулата придобива вида на уравнение на обикновената множествена регресия.

$$\boldsymbol{Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \varepsilon},$$

Ако f(x) = exp(x) и ако отклоненията на Y са с Поасоново разпределение на стойностите, тогава това уравнение отговаря на т.нар. *log-линеен* модел, което се използва при анализиране на меристични (брой индивиди, брой гнезда и т.н.) белези. Уравнението за очакваната стойност на Y получава вид на:

$$\mathbf{E(Y) = exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta p X p) + \varepsilon}$$

Наименованието log-линейно идва от това, че логаритъм от очакваната стойност Y е линейна комбинация от променливите X.

Ако f (x) = exp(x)/[1 + exp(x)] това прави очакваната стойност на Y равна на:

$$\mathbf{E(Y) = exp(\beta_0 + \beta_1 X_1 + \ldots + \beta p X p)/[1 + exp(\beta_0 + \beta_1 X_1 + \ldots + \beta p X p)] + \varepsilon}$$

Това е т.нар. логит модел (*logistic* модел) за Y, която има стойности = 0 (показваща липса на събитие) или = 1 (показваща наличие на събитие).

Има много други възможности за моделиране в рамките на общото уравнение (Табл. 5).

Таблица 5. Модели на регресия, включващи се в *GENLIN*.

Име на модела	Уравнение на модела	Разпределение на стойностите на Y				
		Нормално	Поасоново	Биномно	Гама	Обратно Гаусово
Линейна регресия	$Y = \sum \beta_i X_i + \varepsilon$	по подразбиране	позволено	спорно	позволено	позволено
Log-линеен	$Y = \exp\left(\sum \beta_i X_i\right) + \varepsilon$	позволено	по подразбиране	спорно	позволено	позволено
Логит регресия	$\frac{Y}{n} = \frac{\exp\left(\sum \beta_i X_i\right)}{1 + \exp\left(\sum \beta_i X_i\right) + \varepsilon}$	спорно	не възможно	по подразбиране	спорно	спорно
Реципрочен	$Y = \frac{1}{\left(\sum \beta_i X_i\right) + \varepsilon}$	позволено	позволено	спорно	по подразбиране	позволено
Пробит	$\frac{Y}{n} = \phi\left(\sum \beta_i X_i\right) + \varepsilon$	спорно	не възможно	позволено	спорно	спорно
Двойно експоненциален	$\frac{Y}{n} = 1 - \exp\left(\left(\exp \sum \beta_i X_i\right)\right) + \varepsilon$	спорно	не възможно	позволено	спорно	спорно
Квадратен	$Y = \left(\sum \beta_i X_i\right)^2 + \varepsilon$	позволено	позволено	спорно	позволено	позволено
Експоненциален	$Y = \left(\sum \beta_i X_i\right)^{\frac{1}{a}} + b + \varepsilon$	позволено	позволено	спорно	позволено	позволено

Обратен корен квадратен	$Y = \frac{1}{\left(\sum \beta_i X_i\right)^{\frac{1}{2}} + \varepsilon}$	позволено	позволено	спорно	позволено	по подразбиране

Таблица 6. Подходящи модели на регресия в зависимост от вида на изследваните белези.

Зависима променлива	Независима променлива	Подходящ тип регресионен модел
Качествени номинални белези с две състояния	Всякакви	Двоична логистична регресия (*Binary logistic regression*)
Качествени номинални белези с две състояния	Качествени номинални или в ординална скала	Логит-логаритмични линейни модели (*Logit Loglinear Analysis*)
Качествени номинални белези с повече от две състояния	Качествени номинални или в ординална скала	*Multinomial logistic regression*
Качествени белези в ординална скала	Качествени номинални или в ординална скала	Ординална регресия
Количествени непрекъснати белези	Качествени номинални или в ординална скала	Дисперсионен анализ (ANOVA)– частен случай на линейната регресия (метод на най-малките квадрати)
Количествени меристични белези	Всякакви	log-линейна регресия
Количествени непрекъснати белези	Всякакви	Множествена регрессия

Обобщените линейни модели обикновено се напасват към данните използвайки метода на максималната правдоподобност (*Maximum likelihood*), т.е. стойностите на неизвестния параметър се изчисляват като стойности, които правят вероятността на получените стойности на модела, колкото се може по-голяма.

Пригодността на модела (т.е. пасването му към данните) се измерва от отклонението - D, което е минус два пъти максимализирната log- *likelihood*, със степени на свобода (df) равни на броя на наблюденията (обема на извадката) минус броя на оценяваните параметри.

Вход на данните при *GENLIN*: Зависимата променлива може да бъде количествен метричен белег, мeристичен (брой), двоични (т.е. 0 и 1) или събития при провеждане на лабораторни и полеви опити. Факторите са качествени белези. Ковариращите независими променливи *offset* и *scale weight* (само ако е необходимо да се дава тежест) са количествени белези.

Условие за използването на тези модели е променливите да са независими наблюдения.

Пример при мeристични белези. Тъй като моделът е логаритмичен данните на независимата променлива, описваща зависимата (*offset* – начална независима променлива) се трансформира в логаритми.

Обратна връзка:

В таблици се дава обща информация за модела и променливите, включени в него, брой липсващи стойности, основна дескриптивна статистика.

Таблица *Goodness of fit*: Стойностите на отклонението (*Scaled Deviance*) и тестът $\chi 2$ със степените на свобода би трябвало да са близо до 1.0 при регресия на данни с Поасоново разпределение. Когато са по-големи от 1 показва, че пасването на модела е резонно.

Таблица *Omnibus test* показва резултата от сравняване на нашия модел с нулев модел (*likelihood-ratio chi-square test*). Стойност на P (*significance*) <0.05 показва, че текущия модел пасва по-добре от нулевия модел.

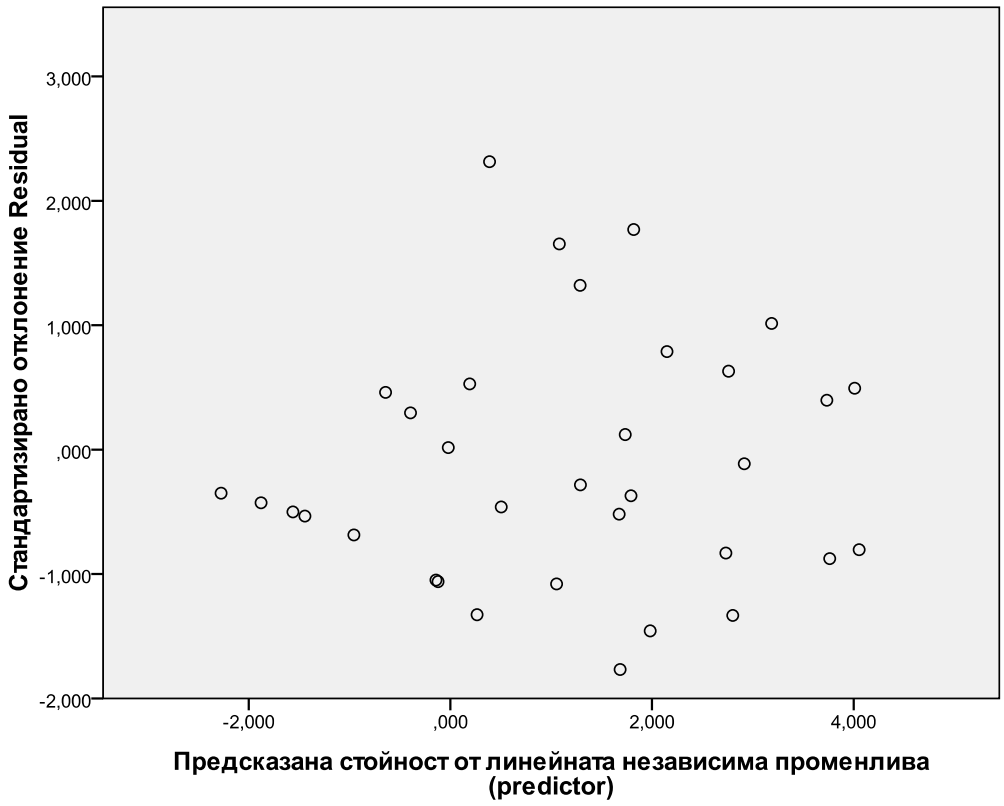
Таблица *Test of Model Effects* включва тестове за всеки фактор дали имат ефект в модела. Тези, при които P (*significance*) <0.05 имат изявен ефект (*main-effects*).

Таблица *Perimeter Estimates* обобщава ефекта на всяка независима променлива. Положителните (или отрицателните) коефициенти **B** за независимите променливи (за различните състояния на качествените променливи, и стойностите на количествените променливи) показват положителна (или обратна) връзка между независимата и зависима променливи. Дава се и статистика на достоверността на коефициентите.

За всеки един от факторите се в отделни таблици са изчислени т.нар. *marginal means*, стандартни грешки, доверителни интервали за количествената независима променлива (predictor) при различните стойности на факторите (качествени белези). От таблицата може да се изследват разликите между стойностите на даден фактор. Резултатите от тестовете за достоверност по двойки показват дали разликите са случайни или достоверни ($P<0.05$).

Таблицата *Overall test* показва резултатите от всички съпоставяния *contrasts* на сравненията по двойки в по-горната таблица.

Графиката на стандартизираните отклонения residual (Фиг. 16) трябва да се направи допълнително. Целта е да се провери дали варирането на отклоненията се променя със стойностите на линейния predictor (количествената независима променлива).



Фиг. 16. Графика на стандартизираните отклонения *residual* от модела.

Тъй като няма как да се тества модела срещу стандартизиран модел, както е при GLM за такъв „стандарт" се приема регресия за отрицателно биномно разпределение и се тества степента на *likelihood* при еднакви всички други настройки. Разликата показва дали поасоновия модел обяснява по-добре данните.

2.1.1.6. Общ log-линеен модел - *General Loglinear Analysis* - анализират се меристични белези (брой).

2.1.1.7. Logit log-линеен модел - *Logit Loglinear Analysis* - анализират се връзките между зависима променлива - качествен белег и една или повече независими променливи също качествени белези.

Опцията *Model Selection Loglinear Analysis* помага да се избере кой модел да се използва – *General Loglinear Analysis* или *Logit Loglinear Analysis*.

2.1.1. 8. Анализи на преживяемост - *Survival analysis*.

Данните за една популация често се обединяват под формата на жизнени таблици ***Life Tables*** с цел оценка на преживяемостта/смъртността на индивидите и оценка на риска. Обикновено се правят отделни таблици по пол поради различната преживяемост на индивидите от двата пола.

Жизнените таблици изследват разпределението на събития, свързани с определен период от време, зависещи от стойностите на независимата променлива. Често не може да се наблюдават всички събития (начални или крайни) в даден период такива данни се наричат цензурирани. Поради тази причина се използват жизнени таблици, а не обикновена множествена регресия. Основната идея на жизнените таблици е да подразделят периода на изследване на по-малки времеви периоди. За всеки по-малък интервал всички индивиди, които се наблюдават най-малкото в този интервал се използват за изчисление на вероятността на крайното събитие (напр. смърт) проявяващо се в този интервал. Вероятностите за всеки един интервал се използват за оценка на общата вероятност на събитието в различни времеви отрязъци.

С помощта на жизнените таблици в SPSS се изчислява брой на емигриращи индивиди, брой на имигриращи индивиди, брой на индивиди изложени на риск, дял на умрели индивиди, дял на оцелели индивиди, кумулативен дял на преживяли индивиди (и стандартна грешка), плътност на вероятността *probability density* (вероятност на една случайна променлива да попадне в даден интервал = интеграл от плътността на променливата около този интервал) (и стандартна грешка), степен на риск (и стандартна грешка) за всеки времеви интервал за всяка група, средно време на преживяемост за всяка група и тест на Wilcoxon (Gehan) (непараметричен тест за достоверност на разлики между медиани на две групи свързани по двойки данни) сравняващ разпределението на преживяемостта между различни групи. Може да се ползват и други тестове за свързани по двойки извадки.

Правят се графики – криви на преживяемост, log преживяемост, плътност, степен на риск, смъртност.

- Статистически жизнени таблици

Вероятност за смърт на индивидите през даден период:

$$q_i = \frac{D_i}{R_i}$$

Където q_i – вероятност за смърт през периода *i*, D_i – брой умрели индивиди за периода i, R_i – брой индивиди в риск от смърт в началото на периода i (това са живите индивиди в началото на периода)

$p_i = 1 - q_i$ е обратната вероятност - вероятност за преживяване, оцеляване през този период.

Това уравнение се нарича още риск (*hazard*) и се различава от т. нар. *hazard function*:

$$\lambda(t_{mi}) = \frac{2q_i}{h_i(1+p_i)}$$

Където t_{mi} е големината на интервала (често данните са за периоди от 1 година и тогава интервала е = 1), степента на риск се оценява в средата на интервала (поради това е изписан и с инициали t_{mi}), q_i е рискът в края на интервала.

В повечето случай обаче в периода на изследване някои индивиди поради различни причини излизат извън обсега на наблюдение - т.нар. загубени и цензурирани случаи (напр. изгубени радионашийници, убити при незаконен лов и т.н.) и не се знае точно кога това се е случило. Приема се, че разпределението на такива случаи през времевия интервал е равномерно и за това на такива индивиди им се дава половината период от време, така половин индивид преминава целия период от време и формулата за риска придобива следния вид:

$$q_i = \frac{D_i}{R_i - \frac{L_i}{2}}$$

L_i е броят на загубени или цензурирани индивиди.

Обратната вероятност се нарича още вероятност за преживяемост, изчислява се и т.нар. кумулативна вероятност за преживяемост (*cumulative probability of survival* или още *survival function*)

Напр. ако в изследването са заложени периоди от 1 година кумулативната вероятност за преживяемост за втората година ще е вероятността за преживяемост през първата година по вероятността за преживяемост втората. Това е пример за условна вероятност. Вероятността за преживяемост от началото на изследването до края на

втората година е условие от преживелите през първата година. Кумулативната вероятност за края на третата година ще е вероятността за преживяемост третата година по вероятността за преживяемост през втората и т.н.

Разлика между вероятност за преживяемост и кумулативна вероятност за преживяемост: Първата вероятност дава вероятността от оцеляване през втората година само за индивидите, които са били налични в началото на годината. Не всички индивиди обаче оцеляват до началото на годината и за това кумулативната вероятност дава вероятността за оцеляване през втората година за всички индивиди от началото на периода на изследване.

Кумулативната вероятност се изчислява за всички периоди и се представя графично с т.нар. *survival curve*.

Вход на данните в SPSS – променливата, съдържаща времевите периоди трябва да е с количествени стойности. Променливата *status* трябва да е бинарна или качествена, кодирана с цифри. Събитията се кодират като единична стойност или интервал от последователни стойности. Факторните променливи трябва да са качествени, кодирани с цифри.

- Kaplan-Meier Survival Analysis. Препоръчва се да се използва при обем на извадката по-малък от 50. При този анализ се изчислява преживяемост в дялове. Дава непараметрична оценка на разпределението на вероятността за крайното събитие (напр. смърт) за една извадка, която съдържа точното време на крайното събитие или цензурирани данни за времето на събитието (т.нар. *right censored data* - точка над която е дадена стойност, но не се знае колко е точната стойност). *Kaplan-Meier Survival* анализът изчислява относителна преживяемост и време и дава крива на преживяемост нар. *Kaplan-Meier survival curve*. Т.е. с този анализ може да се изчисли делът от индивиди от дадена популация, които биха оцелели в определен период от време при същите условия на средата.

Различава се от статистическите жизнени таблици по:

1. Вместо да постави случаите на смърт в условен период, се използва точното време на събитието.

2. Вместо да изчислява функцията на преживяемост за фиксирани периоди от време, това става само за времето, в което е настъпила смърт. Т.е. някои точки ще са близки, а други по-отдалечени във времето.
3. При статистическите жизнени таблици графиката на преживяемостта се променя само в края на всеки времеви интервал, докато при *Kaplan-Meier* се променя всеки път, когато настъпи крайното събитие. Винаги може да се направи разлика и да се разпознаят двете графики, тъй като при статистическите жизнени таблици еднаквите интервали са по оста X – време, докато при *Kaplan-Meier* еднаквите интервали са по оста Y – вероятност.
4. Индивидите, за които се губят данните или са цензурирани случаи са „в риск" до времето, в което отпадат от изследването. Това означава, че те се изваждат от таблицата във времето между две събития (напр. смърт на други два индивида), техните данни се използват при изчисляване на *survival rate* (процента на преживели индивиди) за първото събитие, но не и за второто.

За функцията на преживяемост се изчислява стандартна грешка:

$$SE(P_i) = P_i\sqrt{\frac{1-P_i}{R_i}}$$

Обикновено с увеличаване на времевия интервал и стандартната грешка се увеличава, тъй като изчисленията за преживяемостта се правят върху все по-малко и по-малко индивиди.

Условия за използване на анализите за преживяемост:

1. Определима точка на начало
2. Крайна точка
3. Загубите на информация за индивидите не трябва да е свързана със събитието, което се изследва (напр. смърт)
4. Да няма тенденция за бавна промяна в субектите, тъй като допускането за хомогенност на групата няма да е в сила. Продължителността на едно такова изследване в такъв случай трябва да е до 5 години

Вход на данните в SPSS. Променливата време трябва да е количествена непрекъсната величина, променливата статус може да бъде качествен или непрекъснат количествен белег, факторите и променливата *strata* трябва да са качествени.

Сравняване на преживяемостта на две и повече групи:

- сравняване на двете криви в определена точка чрез z-тест. С този тест може да се сравняват функциите на преживяемост на две групи в определен момент от времето. Тъй като е параметричен тест, допускането е кумулативните стойности на преживяемост да са с нормално разпределение.

$$Z = \frac{Pi_1 - Pi_2}{\sqrt{\left[SE\left(Pi_1\right)\right]^2 + \left[SE\left(Pi_2\right)\right]^2}}$$

Където Pi_1 и Pi_2 са стойностите на кумулативната вероятност за преживяемост P за групите 1 и 2 в случайно избран интервал i (или време t, ако се използва анализа на *Kaplan-Meier*), SE е стандартната грешка на P в този времеви интервал.
В тази точка може да се изчисли и относителния риск (*Relative Risk*). Стойността на RR е отношение на вероятността от сбъдването на някакво събитие в група 1 към вероятността от сбъдването му в група 2.

$$RR = \frac{1 - Pi_1}{1 - Pi_2}$$

За тестване на достоверност на RR отново се използва z-тест.

За сравняване на групите през всички периоди се използва тестът на *Mantel-Cox log-rank*, който е модификация на χ^2 - теста. Той е непараметричен тест. Сравнява броя наблюдавани събития с броя на очакваните. Нулевата хипотеза гласи, че няма разлика между групите. Ако не съществуват разлики между групите, тогава при всеки интервал (или време) общият брой на събитията ще е разделен между групите според броя на индивидите в риск. Напр. ако група А и група Б имат еднакъв брой индивиди, то събитията във всяка група ще са еднакъв брой, ако А е с два пъти по-голям обем, тогава и броят на събитията ще е два пъти по-голям при А в сравнение с Б.
Ei_k - очакваната честота за групата *к (к=1 или 2)* в интервала i е:

$$Ei_k = Di \times \frac{Ri_k}{Ri_1 + Ri_2}$$

Където *Di* е общият брой на случай на смърт. Тогава:

$$\chi^2 = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2}$$

Общият относителен риск в този случай ще бъде:

$$RR = \frac{O_1 / E_1}{O_2 / E_2}$$

- Регресия на Кокс - *Cox Regression* (*Cox proportional hazard model*) Регресията на Кокс се използва за моделиране на събития в даден период от време, базирано на стойностите на дадени ковариращи независими променливи, т.е изследва се връзката между преживяемостта на индивидите и няколко фактора (независими променливи) и така може да се предвиди риска от смърт.

1. Работи с различен брой на независими променливи (covariates). Те могат да са дискретни или непрекъснати величини
2. Третира непрекъснатите величини като такива
3. Дава оценка на големината на разликата между групите, т.е с други думи еквивалент на анализите на съвместна дисперсия (ANCOVA) за данни за преживяемост.

Принадлежи към групата на полу-параметричните методи. Регресията на Кокс е предиктивен модел и се базира на множествената линейна регресия, в която времето на преживяемост е зависимата променлива **Y**. Ефектът на независимите променливи **X** се дава от коефициента **β** в уравнението на множествената регресия и така се оценява степента на риск от смърт при влиянието на тези независими променливи.

Регресията на Кокс разширява уравнението за риск при статистическите жизнени таблици по следния начин: Пропорционалният риск за време t е вероятността за дадено събитие във времето t, при дадена преживяемост до времето t и за специфична стойност на независимата (прогнозираща) променлива X.

Допускане, което се прави е, че ефектът на независимите променливи зависи от стойностите на тези променливи, но не зависи от времето. За да се провери това

допускане се правят криви на преживяемост, ако те се кръстосват, това означава, че не изпълняват условието за независимост от времето. Друг начин е чрез изчисляване на относителния риск, който при дадена стойност на X не трябва да се променя с времето. Пропорционалният риск за точно определено време t за специфична стойност на X = (някаква константа, зависеща от t) по (някаква функция, зависеща от X), т.е.:

$$H(t/X) = c(t) \text{ x } f(X),$$

където *C* е константа зависеща от времето и *f* е функция от X. Константата *C* дава информация за това колко бързо кривата върви надолу. (това е *slope* наклон – отговаря на b в другите регресии)

Вход на данните в SPSS. Данните трябва да са количествени, но променливата *status* може да бъде качествена или непрекъсната величина. Независимите променливи (*covariates*) могат да са непрекъснати или качествени величини, ако са качествени трябва да бъдат кодирани. Променливите *Strata* трябва да са качествени, кодирани с цифри или с букви.

Необходимата големина на извадката за определена сила на тестовете при анализи за преживяемост се определя по стандартни формули.

2.1.2. SPSS® Regression 18

Модулът **Regression** в SPSS дава продължение на обикновената линейна регресия в Statistics Base и логистичните регресионни модели, включени в групата *Generalized Linear Models*, но с някой допълнителни възможности за анализ в изходния файл.

При мониторинг на дадена популация най-често се отчита обилие или присъствие/отсъствие на вида. Стойностите на обилието на даден вид дори и в местообитания със сходни условия на средата обикновено се подчиняват на асиметрично разпределение (напр. лог-нормално), с голям брой ниски и средни стойности и няколко много високи стойности. Разпределението на логаритъм от стойностите на обилието е близко до нормалното. Следователно за линейната регресия (метода на най-малките квадрати) е най-добре да се използват трансформирани данни чрез логаритмуване (в този случай обаче възниква проблем с нулевите стойности – видът отсъства – в този случай логаритъм от 0 е неопределена величина).

За анализ на присъствие и отсъствие се използва метода на логит-регресия.

Двоичната логистична регресия - ***Binomial (binary) logistic regression*** е форма на множествената регресия, приложена за бинарни данни като зависима променлива, т.е. променливата Y има само две възможни стойности. (напр. зависимата променлива може да има стойност 1 или стойност 0 съответно при сбъдване или несбъдване на дадено събитие(напр. присъствие и отсъствие на даден вид), ако вероятностте за стойност на променливата1 = q, то вероятността за стойност 0 е = 1 – q. Този тип променливи се наричат променливи на Bernoulli (или бинарни).

Multinomial logistic regression е продължение на бинарната логистична регресия, където качествената зависима променлива има повече от две състояния. Напр. освен „присъствие" и „отсъствие", може да има и трета група – „невъзможно да се проследи" (Табл. 6).

В логистичната регресия се прилага метода на максимална правдоподобност *maximum likelihood* след като трансформира зависимата променлива в *logit* променлива. *logit* е натурален логаритъм от вероятността на зависимата променлива да има определена стойност или не (обикновено да има стойност 1 в *Binomial* логистичен модел, или да има най-високата стойност в *multinomial* модела). Така логистичната регресия оценява вероятността на дадено събитие, което се е случило (стойност).

Нелинейна регресия (*nonlinear regression*)

В зависимост от данните, понякога линейните видове регресия не могат да обяснят взаимоотношенията на зависимата и независимите променливи. Линейните модели изискват линейни параметри, което може да се постигне и чрез трансформация на данните. Моделът на нелинейна регресия е за параметри с нелинейно разпределение и има следното уравнение:

$$\mathbf{Y = f(X, \theta) + \varepsilon}$$

където: **Y** е зависимата променлива; **f(X, θ) -** нелинейна функция на зависимата променлива от независимата променлива **X** с параметри **θ; ε** - отклонения residuals

Нелинейни функции, които често се използват в анализите на биологични явления са групите на екпоненциалната, квадратична, кубична, сигмоидална, хиперболична и т.н. (Табл. 5)

В SPSS е необходимо да се дефинира уравнението, на което най-добре пасват данните.

Weighted least squares: Дава тежест на определени данни в редицата.

Two-stage least squares: Помага да се вземат предвид корелациите между независимите променливи и грешките.

Probit analysis: Използва се при анализа на данни за т.нар. *dose-response* тестове. В общия случай дава оценка на стойностите на дадени стимули (влияние на фактор) използвайки *logit* или *probit* трансформация на дяловете от извадката, подложена на това влияние.

Обикновено се използват в токсикологията. С помощта на този тест се оценяват летални дози от даден препарат или вещество и времето за преживяване на индивиди от даден вид при определени дози. За оценка на LT-50 и LC-50 се използва кумулативно нормално разпределение (пробит) или *logit*. LT-50 (средно време за оцеляване) е фиксирано време на реакция на даден организъм към даден токсин. LC-50 (средна летална доза) е фиксирана концентрация на даден токсин, която води до смъртта на половината от организмите, подложени на токсина за определен период от време.

2.1.3. SPSS® Missing Values 18

Методите за анализ на частично липсващи данни може да се групират в следните категории:

1. Методи, базирани на записите, в които не липсват стойности. Когато някои променливи имат липсващи стойности най-лесният начин е тези записи да се изключат напълно и данните да се анализират само с пълни променливи. Това може да върши работа при малки липси на данни, но при по-големи, ще доведе до голямо отклонение от истината.

2. Методи, базирани на заместване на липсващите стойности (*imputation*). Липсващите стойности се попълват и получените вече пълни данни се анализират.

Често използван метод на заместване е т.нар. *hot deck imputation* (липсващите стойности се попълват от случайно избрани сходни стойности на същата

променлива). Този метод е доразвит и включва метода на най-близкия съсед (*nearest neighbour hot deck imputation*) и *Bayesian bootstrap*. Друг метод е *mean imputation*, където се замества със средните аритметични на стойностите, които са налични. Трети метод е *regression imputation*, където липсващите стойности за дадена част от променливата се изчисляват чрез регресия с други известни променливи с пълни стойности. За да е валиден резултата при използване на такова заместване е необходимо да се ползват модификации на стандартните тестове, за да позволи да се различи статуса на истинските стойности и запълнените липсващи данни.

3. Методи, при които се придава тежест на стойностите.

Рандомизирани оценки за извадъчни данни, които включват тип липсващи данни „без отговор" (напр. при анкетиране, когато данните не са взети поради нежелание на анкетирания да отговори), които често се базират на даване на тежест (*design weight*), които са обратно пропорционални на вероятността да бъдат взети от генералната съвкупност.

Напр., нека y_i да е стойност на променливата Y за ***i***-тия дял в генералната съвкупност. Тогава средната аритметична често се изчислява чрез:

$$\Sigma\, \pi_i^{-1}\, y_i \,/\, \Sigma\, \pi_i^{-1}$$

Където сумите са за индивидуалните наблюдения в извадката, където π_i е вероятността делът ***i*** да попадне в извадката, а π_i^{-1} е *design weight* за дела ***i***. При даването на тежест, за да се приспособят към nonresponse тип липсващи данни е необходимо уравнението да се модифицира. Уравнението в такъв случай добива вида:

$$\Sigma\, (\pi_i\, \hat{p}_i)^{-1}\, y_i \,/\, \Sigma\, (\pi_i\, \hat{p}_i)^{-1}$$

Където сумите сега са за индивидуалните наблюдения, които отговарят, където $\hat{p}_i$ е оценка на вероятността от of response за дела i, обикновено дела на съответните дялове в дадена подгрупа в извадката.

Даването на тежест съответства на *mean imputation*, ако напр. даваните тежести са константни в подкласовете на извадката, тогава и двата подхода заместват липсващите стойности със средната аритметична на подгрупата и съответната единица, на която е дадена тежест със съответния дял във всеки подклас. Така те

дават една и съща средна аритметична на генералната съвкупност, но не и една и съща дисперсия на извадката.

4. Модели. Широк клас от методи, описващи модела на липсващите данни и базиращи се на заключенията от *likelihood* метода, оценени чрез процедури като *maximum likelihood*. Предимство на този метод са приспособимостта; избягването на *ad hoc* методите; допусканията в модела могат да се оценяват; възможност за оценка на дисперсията като вторични резултати на *log-likelihood* подхода, който взима предвид непълнотата на данните.

Първата опция в SPSS – ***Missing value analysis***. Позволява да се направи пълна статистика на липсващите стойности за различни видове белези – дескриптивна статистика на липсващите стойности, регресионен анализ, E-M анализ и т.н.

Част от методите за анализ могат да се използват само при определени модели на липсващи стойности. Това налага първо да се подредят данните в някакъв ред с цел да се види моделът на липсване. Познаването на механизмите, водещи до липса на определени данни са от ключово значение за избора на анализите и интерпретирането на резултатите.

Когато липсват данни в повече от една променлива е най-подходящо да се използват базирани на максималната правдоподобност анализи – *likelihood analysis*.

- Оценка на средната аритметична и съвместната дисперсия за данни с монотонен модел на липсващи данни.

Ако се предположи, че данните са подредени в монотонен модел, лесен подход за оценка на средната аритметична, съвместната дисперсия и др. е да се ограничат анализите до пълните части на всички наблюдавани променливи. Което означава да се изключат редовете с липсващи стойности, което би довело до загуба на голяма част от данните.

Също така и в повечето случаи данните, които са пълни не са случайна извадка от оригиналната извадка (т.е данните не са MCAR - *missing completely at random* - липсващи напълно случайно), и съответните анализи ще са подвеждащи.

Ако се допусне, че данните са с многомерно нормално разпределение, средните аритметични и съвместните дисперсии (*covariance matrix*) и др. параметри може да се

изчислят чрез *maximum likelihood*. За монотонни данни това е улеснено, правейки възможно оценките за *maximum likelihood* да се изчислят чрез поредица от регресии.

- Оценка на средната аритметична и съвместната дисперсия за данни с различни модели на липсващи стойности.

Много данни не показват удобния монотонен модел на липсващи стойности. Методите за оценка на средната аритметична и матрицата на дисперсиите за няколко променливи, които могат да се приложат на всеки модел на липсващи стойности са базирани на *maximum likelihood*. Условие за използване на метода е променливите да са с многомерно нормално разпределение.

Алгоритъмът *Estimation-maximization* (EM) е важна обща техника за намиране на оценката на *maximum likelihood* от непълни данни. Този EM алгоритъм се използва и в по-особени случаи на липсващи стойности при модели на компонентите на дисперсия (*variance components* модели), факторен анализ и времеви редици, които могат да се разглеждат като ситуация с липсващи стойности с многомерно нормално разпределение със специфична средна аритметична и дисперсия.

- Оценка на средната аритметична и съвместната дисперсия за качествени номинални данни.

В този случай данните се подреждат в таблица 2x2, където крайните колони са частично класифицирани.

При комбинация от няколко количествени и няколко номинални променливи.

- Оценка на средната аритметична и съвместната дисперсия при данни, които вероятно не липсват случайно.

Multiple imputation в SPSS дава статистика на моделите на липсващите стойности (т.е. в една матрица от данни по какъв начин са разпределени липсващите данни.) – *Analyze patterns*

Анализира се модела на липсващите стойности по колони, по редове и по стойности.

1. Дават се кръгови диаграми. По този начин може да се избере начинът на заместване на липсващите данни, т.е. да се прецени дали при изтриване на редовете (или колоните) с липсващи стойности (методът е стандартен при

2. Дава се таблица със статистика за всяка колона - процент на липсващи стойности, и средна стойност и стандартно отклонение на валидните стойности в колоната.
3. Прави се графика с моделите на липсващите стойности.

Всеки ред съответства на група от редове с един и същ модел на пълни и непълни данни.

Графиката подрежда променливите и моделите така, че да открие монотонност (липса на промяна на стойностите) там, където съществува. Колоните се подреждат от ляво на дясно по белези с най-малко липсващи към тези с най-много липсващи стойности. След това моделите се подреждат по последната колона (първо наличните стойности и после липсващите), после предпоследната и т.н. работейки от дясно наляво. Това показва дали може да си използва метода за монотонно попълване на липсващите стойности, или ако не, то до колко нашите данни съответстват на монотонния модел. Ако данните са монотонни, тогава всички празни и пълни клетки в графиката ще бъдат равномерно разпределени и няма да има „островчета" (струпвания) от пълни клетки в долната дясна част на графиката и струпвания на празни клетки в горната лява част на графиката.

В примерната графика данните не са монотонни и трябва да се въведат много стойности, за да се постигне монотонност.

Когато моделите го изискват, се дава и допълнителна графика, която показва % от случаите с даден модел.

Multiple imputation представлява запълване на клетките с липсващи стойности и се прилага, след като е анализиран модела на липсващите данни (монотонен или друг тип). Прави няколко варианта на заместване на стойностите.

2.2. Примери за приложението на функциите в модулите при разрешаване на биологични проблеми

Методите на регресионен анализ, включени в модулите *Advanced Statistics* и *Regression* биха могли да се използват за отговор на следните екологични въпроси:

1. Оценка на екологични параметри за даден вид, като напр. оптимум и екологична амплитуда на вида.
2. Определяне на този фактор на средата, който има най-голям дял във варирането на определени популационни параметри на вида (посредством проверка на статистическата достоверност), както и на тези фактори, които по всяка вероятност нямат голямо значение.
3. Прогнозиране на изменението в параметрите на вида (обилие или присъствие-отсъствие) в дадено местообитание в резултат на анализите за влиянието на един или повече фактори на средата.
4. Прогнозиране на значението на дадени фактори на средата в местообитанията от анализите за влиянието им върху един или повече видове (калибровка).

Предварително трябва да се направи анализ на типа данни и типа разпределение на стойностите на зависимата и независими променливи, за да се прецени кой тип регресия да се приложи.

За да се постигне линейност във взаимоотношенията на независимите и зависимите променливи понякога може да се наложи трансформация на данните, в противен случай се използват нелинейните регресионни модели.

Тъй като *Generalized linear models* обобщават повечето видове регресионни модели, се използват най-често при анализ на резултати от мониторингови програми и прогнозиране на тенденции в обилието, видовото богатство, наличие на вида в даден хабитат.

1. Hiltrud Brose и Michael Nobis (2009) правят изследване в Швейцария на промените във видовото богатство на висшите растения като прогноза за влиянието на климатичните промени в определени местообитания. Потенциалното влияние на промените в околната среда върху бъдещото видово богатство е изследвано чрез моделиране. Използват обобщени линейни модели *Generalized linear models,* за да свържат видовото богатство, установено по трансектен метод с параметрите на околната среда за всяка група. Видовото богатство на съответните групи е моделирано

в комбинация използвайки принадлежността към дадена група като фактор. В резултат са получени два модела – един за групите базирани на статуса на видовете – местни и неспецифични за мястото видове и втори модел - за групи разделени по биогеографски произход на видовете. Видовото богатство е проектирано използвайки прогнозни данни за затопляне на климата за 2050 год. и сравнено с видовото богатство в днешно време.

В резултат моделите дават следните прогнози:

- Общо увеличаване на видовото богатство на висши растения.
- Броят на местните видове средно ще намалее, докато ще се увеличи броя на видове, които не са били характерни за страната.
- Видовото богатство на Алпийските и Планински видове ще намалее, докато на Европейските, Медитеранските и неевропейските видове ще се увеличава.
- Предвижданите промени са сравнени и по надморска височина.

2. Pokluda *et al.* (2011) изследват ефектите на различни биотични и абиотични фактори върху улавянето на вида *Carabus hungaricus*, включен в директивите по Natura 2000. Целта е да се знае къде и при каква комбинация от условия в хабитатите трябва да се поставят капани, за да може да се осъществява най-ефективно бъдещ мониторинг на популациите на вида. Методът на улавяне е чрез живоловни почвени капани с примамка. Индивидите са улавяни, маркирани и пускани. Събирани са в продължение на 1 година. Изчислен е среден брой улавяния на капан за дадено местообитание.

- Ефектът на хабитата върху броя улавяния на *C. hungaricus* (трансформирани с ln) е изследван чрез еднофакторен дисперсионен анализ *one-way ANOVA*, следвана от тест на Tukey's HSD за извадки с различен обем N.
- Ефектът на растителността върху броя на улавянията на *C. hungaricus* е изследван чрез *Generalized Linear Models*. За да се установи пространственото ниво на данните за растителността с най-голям ефект върху улавянето на *C. hungaricus,* са тествани отделно пълни модели (quasi-поасоново разпределение на случайната вариация, log линейна регресия, ln-трансформирани независими променливи) за данните за растителността на три пространствени нива (в кръг около капаните с радиус 0.5 m, 2.5 m, 5 m).

За независими променливи са използвани процентно покритие около капаните на следните видове растителност за трите пространствени нива:

Обща растителност

тревиста растителност

широколистна тревиста растителност

къси треви (<20 cm)

високи треви(>20 cm)

листна постилка

затворена гора

единични дървесни растения (дървета и храсти над 50 cm)

къси храсти (<50 cm)

обработваема земя

степни треви доминирани от *Stipa* spp.

наличие на инвазивния вид *Glycyrrhiza glabra*

наличие на *Calamagrostis epigejos*

Зависимата променлива е брой улавяния на бръмбара за изследвания период.

Моделите се оценяват и сравняват чрез стойностите на D *model deviance* (= обяснената вариация от модела. Който модел е с най-голяма стойност на D, той има най-висока обяснителна стойност за зависимата променлива) и теста на Mallows Cp. След като се определи пространственото ниво с най-голяма обяснителна сила, всички променливи за растителността на това ниво се подлагат на допълнителни тестове. Техните независими (marginal) ефекти върху улавянето на *C. hungaricus* се оценяват чрез F-тест.

- Чрез GENLIN е изследван и ефектът на абиотични фактори върху броя на улавянията на *C. hungaricus* (поасоново разпределение на случайните отклонения ε, log линейна регресия). Изследвани са следните абиотични променливи:

температура

влажност

pH на почвата

съдържание на азот в почвата

светлина

соленост

наклон на терена (в градуси)

- Чрез GENLIN са изследвани различията между половете в използването на местообитанието, които касаят абиотичните фактори (биномно разпределение на случайните отклонения ε, *logit* линейна регресия).

Делът на женските в общия брой улавяния е зависимата променлива, абиотичните фактори са независимите променливи. Нулевата хипотеза, която се тества е, че предпочитанията на женските към дадено местообитание не се различава по време и извън размножителния период. Анализите са правени отделно с данните от целия активен сезон с изключение на размножителния период (March 26th–August 8th, 2006) и отделно с данните от размножителния период (August 9th–November 6th, 2006).

Тъй като в използваните GENLIN модели независимите фактори са голям брой, нивото на достоверност се коригира с *Bonferroni correction*. Стойности на P между 0.05 и коригираната стойност на нивото на достоверност се считат за достоверни.

Благодарение на тези анализи може да се дадат препоръки за управление на дадена територия с цел благоприятстване на популацията на вида.

3. При изследване на влиянието на фрагментираните ландшафти върху разнообразието на прилепите е използван GENLIN модели с Поасоново разпределение на грешката за сравняване на α – разнообразието в седем типа местообитания. Поасоновото разпределение се използва при описание на меристични (брой) белези, когато се знае колко пъти се е случило нещо (в този случай – брой видове за една нощ), но не знаем колко пъти не се е случило. Зависимата променлива е брой видове прилепи отчетени за нощ във всеки хабитат, което дава общо 120 записа.

Тъй като данните са от извадки взимани на всеки 2 месеца за 2 години първоначално времето било включено като независима променлива. Резултатите обаче показали че променливата време не оказва достоверен ефект върху видовото богатство (D=16.49, df 23) нито взаимодействието и с променливата тип хабитат не оказват достоверно влияние (D = 87.93, df 90) и в последствие е изключена от модела.

Така единствено типа хабитат остава като независима променлива за обяснението на вариацията във видовото богатство.

α – разнообразието е анализирано, използвайки два израза за видово богатство: 1. кумулативно α – разнообразие – общ брой видове установени през общия брой нощи за хабитат и 2. средно α – разнообразие – сумирани всички видове прилепи, хванати през всички нощи разделено на броя нощи на улов за хабитат.

4. Reynolds S. (2006) прави изследване на влиянието на ветрогенераторите върху постоянни и мигриращи популации на прилепи с цел да се изработят протоколи за мониторинг за надеждна оценка на риска от полетата с ветрогенератори. Използва два вида методики за улавяне – мрежа и акустичен.

- Изследвана е лятната активност на прилепите чрез *General linear model Multivariate* и *post hoc* тест – на *Tukey multiple comparisons*

Параметри:

За да се изследва денонощната активност на прилепите през летния период, всяка нощ се разделя на три периода с еднаква продължителност:

ранен (19.00–22.59 h)

среден (23.00–02.59 h)

късен (03.00–07.00 h).

Изследваните местообитания са категоризирани в 5 типа:

Пътеки и пътища

Реки и ручеи

Езера

Полета

Влажни зони

Блата

Броят уловени прилепи в даден период от дадено място е зависимата променлива, а факторите (независими променливи) са 2 – период на улавяне и тип хабитат.

- Чрез *General linear model Multivariate* и *post hoc* тест – на *Tukey multiple comparisons* е изследвана и лятната активност на мигриращите прилепи.

Параметри:

За да се изследва денонощната активност на прилепите през летния период на миграция, всяка нощ се разделя на три периода с еднаква продължителност:

Ранен (19.00–22.59 h)

Среден (23.00–02.59 h)

Късен (03.00–07.00 h).

За изследване на сезонната вариация в активността на прилепите периода на изследване е разделен на 3 еднакви интервала:

Ранен (10 април – 4 май)

Среден (5 май - 29 май)

Късен (30 май - 22 юни)

За да се определи на каква височина летят прилепите са поставени микрофони на кула ориентирани спрямо преобладаващия вятър (посоката на турбините) на три различни височини:

приземно ниво (грубо 7 m над земната повърхност),

над дървесната растителност (грубо 25 m над земната повърхност)

ниво на турбината (50 m над земната повърхност).

Броят на установените прилепи в даден период от дадено място и на дадена височина е зависимата променлива, а факторите (независими променливи) са 3 – период на улавяне през нощта, период на улавяне през лятото и височина, на която са установени. Ако активността на височината на турбините е голяма, то тогава има висок риск за прилепите. Тъй като активността на прилепите не е постоянна величина от тези анализи може да се определят точните периоди, когато рискът е висок.

- Анализите на миграторната активност в зависимост от метеорологичните условия са ограничени в периода 10 април до 22 юни, когато са събирани тези данни. За установяване влиянието на вятъра и температурата отново е използван *GLM multivariate* с последващ тест на *Tukey multiple comparisons*. Фактори на средата са 2 – посока на вятъра за всеки ден и средна температура. Посоката на вятъра е категоризирана в 8 категории по следния начин: средни дневни стойности азимут на 8,45° сегменти - N-NE, ENE, E-SE, S-SE, S-SW, W-SW, W-NW и N-NW. Поради големия брой дни без установена активност на прилепите (19 нощи, 26% от дните), данните за активността са категоризирани:

липса на активност (0 прилепа/нощ)

ниска активност (1–2 прилепа/нощ)

средна активност (3–6 прилепа/нощ)
висока активност (>6 прилепа/нощ)
Тези групи са установени *post hoc*, за да се минимизира вариацията във всяка група. Променливите с метеорологични данни също са анализирани чрез корелационен анализ (Pearson correlation), за да се определи степента на независимост на факторите.

5. Brand *et al*. (2010) изследват влиянието на растителността и хидроложкия режим на реката San Pedro върху популационната плътност на различни видове птици.
Отчитана е плътност на птици по точкови трансекти, започващи от корито на река и вървящи перпендикулярно на реката. Изследвани са няколко локалитета по протежение на реката, като във всеки локалитет са изминавани по 11-18 точкови трансекти.
Използвана е регресия *Mixed model*, където се включват и независими променливи със случаен ефект. Зависимата променлива е плътност на птиците за всяка точка.
Независими променливи (*predictors*) - хидроложки режим (целогодишен, пресъхващ, краткотраен), тип растителност (преобладаващи видове растения)
независима променлива със случаен ефект са локалитетите, което спомага да се премахне случайната вариация свързана с локалитета.
Дизайнът на този експеримент е подобен на split-plot тъй като растителният тип е на ниво точка, а хидроложкият режим е на ниво локалитет.

В модула *Advanced statistics* се включват и анализите свързани с т.нар. *Life tables***,** *Kaplan-Meier, Cox regression* анализи за преживяемост често използвани при анализ на смъртност и риск на различни популации обект на мониторинг.

6. Koehler G. & Pierce D. (2005) изследват три популации на маркирани мечки с радиопредаватели. Извадката е от 136 мечки от три местообитания за период от 5 години. Първите две години ловът е бил разрешен, а следващите години е забранен и ловът е незаконен. Сравняват степента на преживяемост през ловния сезон и през другите сезони, както и периода преди забраната за лов на мечки с този след. Полът и годините на отстреляните мечки се отразява на демографската структура на

популациите. Използват *Kaplan-Meier* анализа за оценка на средна възраст (*median age*) и преживяемост (*survival*). Данните са отчитани седмично.
За цензуриран случай (censored case) се счита, когато в дадена седмица мечката е изгубила нашийника или се е развалил или е излязла извън обхват на телеметрията.
Вкарват се данни за всички мечки за всяка година, всяко местообитание и се сравняват параметрите на преживяемост между различните години, между различните местообитания, между ловния период и останалия период, преди забраната за лов на мечки и след това, между половете чрез log-rank тест. Използван е еднофакторен дисперсионен анализ (ANOVA) за анализ на разликите в седмичната степен на преживяемост (стойностите са трансформирани корен квадратен аркуссинус с цел да се изпълни условието за нормалност на разпределението и еднаквост на дисперсиите). За да се провери дали цензурираните случай не са резултат от нелегален лов е използван t-тест, като данните отново са трансформирани корен квадратен аркуссинус. Сравнени са степента на цензуриране на данните за мъжките и женски през ловния сезон и останалото време. За да се установи дали мечките може да са убити, но да не са съобщени, се сравнява седмичната степен на цензуриране за мечките през ловния сезон и останалото време.

7. Person D.& Russel A. (2008) използват *Cox regression* за оценка на влиянието на различни местообитания (в част от които има човешка намеса) върху риска от смърт на 55 вълка, следени с радионашийници. Рисковете са от лов и капани в зависимост от достъпността и близоста на хабитата до човешка дейност. Сравняват степента на смъртност за постоянните и преминаващите вълци, използвайки *Cox proportional hazards regression*, за да свържат структурата на хабитата в рамките на 100-m радиус около засечените вълци с риска от смърт на постоянните и преминаващи вълци. Включват и независими променливи (*covariates*) като разстояние до пътища, сечища, езера и потоци.

8. Millar J. & Zammuto R. (1983) на базата на литературни данни за няколко вида бозайници използват следните параметри, за да анализират преживяемостта при различните видове. Изходни данни за построяването на жизнените таблици са:

1. Брой на женски, които оцеляват във всеки възрастов клас
2. Размер на котилото (брой родени индивиди)
3. Възраст, при която настъпва полова зрялост (години) - това е първата възраст, на която над 50% от женските дават потомство

На базата на тези данни се изчисляват:

1. Продължителност на 1 поколение (години) - средната възраст на ♀, които дават потомство/ броя на новородените
2. Очаквана продължителност на живот при раждане (години)
3. Очаквана продължителност на живот при полова зрялост (години)
4. Репродуктивност при полова зрелост
5. Остатъчна репродуктивност при полова зрелост

9. Inchausti P. & Halley J. (2003) изследват устойчивостта на дадена популация свързана с варирането на големината на популацията във времето и спектралния цвят на популационната динамика (получен в *time series analysis*), използвайки анализа на *Cox proportional hazards*. Изследвали са ефектите на независимите променливи върху риска, който показва потенциалът популационното обилие да намалее 90% за даден малък времеви интервал. Ефектът на независимата променлива на всяка времева стъпка действа мултипликационно на неизвестната, обща основна степен на риск, която се приема, че е еднаква за всички индивиди. За цензурирани случаи се считали тези популации, които не показват 90% намаляване на първоначалното обилие по време на изследването. Анализите са правени общо за всички данни, както и за индивидуални таксони (Mammalia, Aves, Osteichtyes, Insecta), за трофични нива (herbivores carnivores, secondary carnivores), и за тип хабитат (сухоземен или воден). Резултати от *Cox proportional hazards* анализа са "наклона" (slope), даващ ефекта на независимите променливи (вариране на популацията и спектрален цвят), и кривите на преживяемост, които показват разликите между изследваните групи животни (напр. водни срещу сухоземни). Положителен наклон показва, че независимите променливи увеличават степента на риск, която авторите в случая наричат степен на quasi-extinction (измиране на популацията), и следователно намалява времето за крайното събитие в случая времето за измиране. Статистическата достоверност на Cox

анализите на риска е определена чрез *log-likelihood* тест, а на наклоните чрез тест на *Wald*.

10. Woodal et al. (2005) използват анализите за преживяемост *survival analysis* като нова методика за определяне, мониторинг и подобряване на здравето на гори, с влошаващо се състояние. Изследвани са няколко вида дъб за 17 годишен период от време. Целите им са: да сравнят функциите на преживяемост на различните групи (видове дъбове, тип гора, място, физиография, собственост и стопанисване на горите) чрез log-rank тест; да се сравнят ефектите на независимите променливи (регион, изложение, позиция на склона, наклон, етажно покритие на дъба (каква част от земната повърхност покрива проекцията на короните на дъбовете), същинско покритие (каква част от земната повърхност заемат основите на дърветата), склоп (като се погледне нагоре каква част от небето не се вижда, това покритие не включва клоните и стволовете на дърветата), корона, повреда на дървета, бонитет (индекс за условията, които предлага местообитанието като комбинация от влага и състав на почвата) върху смъртността и функцията на преживяемост на дъбовете на ниво дървета и на ниво място чрез *log-rank* тест; да сравнят функцията на риск (*hazard functions*) измежду три категории инвентаризирани дъбови дървета (всички, здрави дървета, болни дървета).

Модулът ***Missing values*** дава възможност да се направи анализ на липсващите данни от мониторинга на национално ниво. След получаване на срези от информация за изследваните параметри на различните групи организми от националната база данни може да се анализира типа на липсване на данните, да се прецени дали може да се изключат случаите с липсващи данни от анализите, да се прецени дали да се заместят липсващите стойности или да се прецени, че анализите са невъзможни поради много липсващи данни, които не могат да се заместят по различните методи.

11. Jackson S. *et al.* (2004) оценявайки ефективността на избраните места за мониторинг на птици във Великобритания, за да могат да изчислят популационните индекси за всяка година, използват заместване на липсващи данни от различните изследвания чрез линейна интерполация.

2.3. Препоръки относно експорта на данните от базата данни към съответните модули на SPSS с цел статистически анализ на резултатите от мониторинга на регионално и национално ниво.

В анализите от гореописаните модули на SPSS се изисква изследваните белези да са подредени в колони. Препоръчително е в справките всяка променлива да има колонка с метаданни (етикети с пълното название на белега и типа белег). Необходимо е да се внимава при записите с десетична запетая дали програмата ще ги разчете, когато са с точка или запетая. По принцип, за да ги приеме програмата изисква десетична запетая. В такъв формат трябва да се експортират от базата данни.

Много от анализите изискват данните от различни групи да се поставят в една колона, а до тях да е съответната групираща променлива (която кодира принадрлежността на стойностите към определена група). Пример – за сравняване на разпределението на белези е необходимо да се направят боксплотове в една обща графика – данните за всички белези са в една колона, а в съседната е групиращата променлива, която чрез цифри указва коя стойност към кой белег се отнася. Т.е. при възможност, когато ще се правят такива анализи да има опция срезовете на информация директно да се подреждат в една колона и във втора колона кодове за принадлежност към даден белег. Кодиращата променлива също трябва да има обяснителни етикети (коя стойност на кой белег отговаря).

Файлът в текстови формат (tab delimited) от справките за връзка с SPSS не може да бъде импортиран, тъй като не се разчита от програмата. Обикновен текстови файл се приема без проблем. Да се погледне дали проблемът не е в кодирането на текстовия формат.

При една част от анализите, където се използват качествени номинални или в ординална скала белези се изисква кодиране на състоянията или заместване с т.нар *dummy variables*. При кодиране е необходимо дефиниране на стойностите с определен код. Това става ръчно или автоматично в самата SPSS.

2.4. Литература:

Brand L., Stromberg J., Noon B. (2010) Avian Density and Nest Survival on the River: Importance of Vegetation Type and Hydrologic Regime. *Journal of Wildlife Management* 74(4):739–754

Brose H. & Nobis M. (2009) Partitioning species richness of vascular plants to analyse the impact of climate change at the landscape scale. *Poster GFO*.

http://faculty.chass.ncsu.edu/garson/PA765/missing.htm

http://www.ruwpa.st-and.ac.uk/distancesamplingreferences/

Inchausti P. & Halley J. (2003) On the relation between temporal variability and persistence time in animal populations. *Journal of Animal Ecology* 72, 899–908

Jackson S., M. Kershaw, K. J. Gaston 2004. Size matters: the value of small populations for wintering waterbirds. *Animal conservation* 7, 229-239

Jongman R.H.G., C.J.F. ter Braak, O.F.R. van Torgeren. 1996. Data analysis in community and landscape ecology. *Cambridge University Press*. pp. 29-79

Koehler G. & Pierce D. (2005) Survival, cause-specific mortality, sex, and ages of American black bears in Washington state, USA. *Ursus* 16(2):157-166

Larose, D. 2006 . Data mining methods and models John Wiley & Sons, Inc – от33 стр.

Little, R. J. A. & Rubin, D. B. 1987. Statistical analysis with missing data. *New York: John Wiley & Sons*. 278 pp.

Manly B. 2009. Statistics for environmental science and management. Second edition. *Chapman & Hall/CRC*. 295 pp.

Merson D., I. McHale. 2010. Data Cleaning and Missing Data Analysis. *Lecture presentation.*

Millar J. & Zammuto R. (1983) Life histories of mammals: an analysis of life tables. *Ecology* 6 4(4): 631-635

Moreno C. & Halffter G. (2001) Spatial and temporal analysis of α, β and γ diversities of bats in a fragmented landscape. *Biodiversity and Conservation* 10: 367–382

Person D.& Russel A. (2008) Correlates of Mortality in an Exploited Wolf Population. *Journal of wildlife management* 72(7):1540–1549

Pokluda P., Hauck D., Cizek L. (in press) Importance of marginal habitats for grassland diversity: fallows and overgrown tall-grass steppe as key habitats of endangered ground-beetle Carabus hungaricus. *Insect Conservation and Diversity* (2011).

Reynolds S. (2006) Monitoring the Potential Impact of A Wind Development Site on Bats in the Northeast. *Journal of wildlife management* 70(5):1219–1227

SPSS Advanced statistics 17. 2007. *Chicago, IL: SPSS Inc*. 189 pp.

SPSS Missing values 17. 2007. *Chicago, IL: SPSS Inc*. 123 pp.

SPSS Regression Models 16. 2007. *Chicago, IL: SPSS Inc*. 45 pp.

Vincent D., L. Godet, R. Julliard, D. Couvet, F. Jiguet. 2007. Can common species benefit from protected areas?. *Biological conservation* 139: 29-36

Woodal C., Garmbsch P., Thomas W., Moser W. (2005) Survival analysis for a large-scale forest health issue: Missouri oak decline. *Environmental Monitoring and Assessment* 108: 295–307

3. Биостатистически методи за обработка на информацията от Националната база данни, за биологичните групи: висши растения, бозайници, птици.

3.1.Обобщаване на резултатите относно актуалното състояние на обектите на мониторинг по заложените параметри

3.1.1. Дескриптивна статистика

Дава информация за извадъчните показатели, тяхната стандартна грешка и вариация.

Пример: Данни от мониторинг на *Lilium jankae* на Витоша (2009). Извършва се на 5 места: резерват Бистришко бранище, Голи връх, м. Коминте, м. Конярника, м. Матница, м. Меча поляна. Искаме да представим параметрите на популацията в планината.

Таблица 7. Изходни данни от мониторинг на *Lilium jankae* на Витоша 2009.

Площадка	Площ на популацията ha	Плътност %	Проективно покрие на вида %
Бистришко бранище 1	0.0674	0.3	5
Бистришко бранище 2	0.0432	0.5	1
Бистришко бранище 3	0.0493	0.5	1
Бистришко бранище 4	0.028	0.1	1
Голи връх	0.179	0.1	1
Коминте	0.072	0.2	3
Конярника 1	0.0875	0.2	1
Конярника 2	0.002	1.9	3
Матница 1	0.2	0.025	1
Матница 2	0.04	0.1	1
Меча поляна	0.1448	0.04	1

Таблица 8. Доклад от SPSS за стойностите на извадъчните показатели.

Statistics

		Площ	Плътност	Покритие
N	Valid	11	11	11
	Missing	0	0	0
Mean		,083018	,360455	1,727273
Std. Error of Mean		,0193681	,1618929	,4065578
Median		,067400	,200000	1,000000
Mode		,0020[a]	,1000	1,0000
Std. Deviation		,0642367	,5369379	1,3483997
Variance		,004	,288	1,818
Skewness		,847	2,781	1,800
Std. Error of Skewness		,661	,661	,661
Kurtosis		-,438	8,311	2,611
Std. Error of Kurtosis		1,279	1,279	1,279
Range		,1980	1,8750	4,0000
Minimum		,0020	,0250	1,0000
Maximum		,2000	1,9000	5,0000

a. Multiple modes exist. The smallest value is shown

Обратна връзка:

N дава обема на извадката. За всички белези той е 11.

Размахът (*Range*) е разликата между максималната и минималната стойност на белега. В случая за площта на популацията е 0,198 ha, на плътността 1,875%, на покритието е 4 %

Минимум и максимум са границите, в които варира белега. В случая площта варира от 0,002 до 0,2 ha; плътността варира от 0,025 до 1,9%, а покритието варира от 1 до 5%.

Средната аритметична представлява центъра на разпределението на стойностите на белега. Стандартната грешка на средната е мярка за прецизността на данните в извадката. Стандартно отклонение е в същите мерни единици като средната и е мярка за вариацията в извадката.

В случая популацията на *Lilium jankae* на Витоша е със средна площ 0,083 ha и стандартно отклонение ± 0,064 ha – показва силна вариация на белега, което се вижда и от размаха. Стандартната грешка на средната е 0,019.

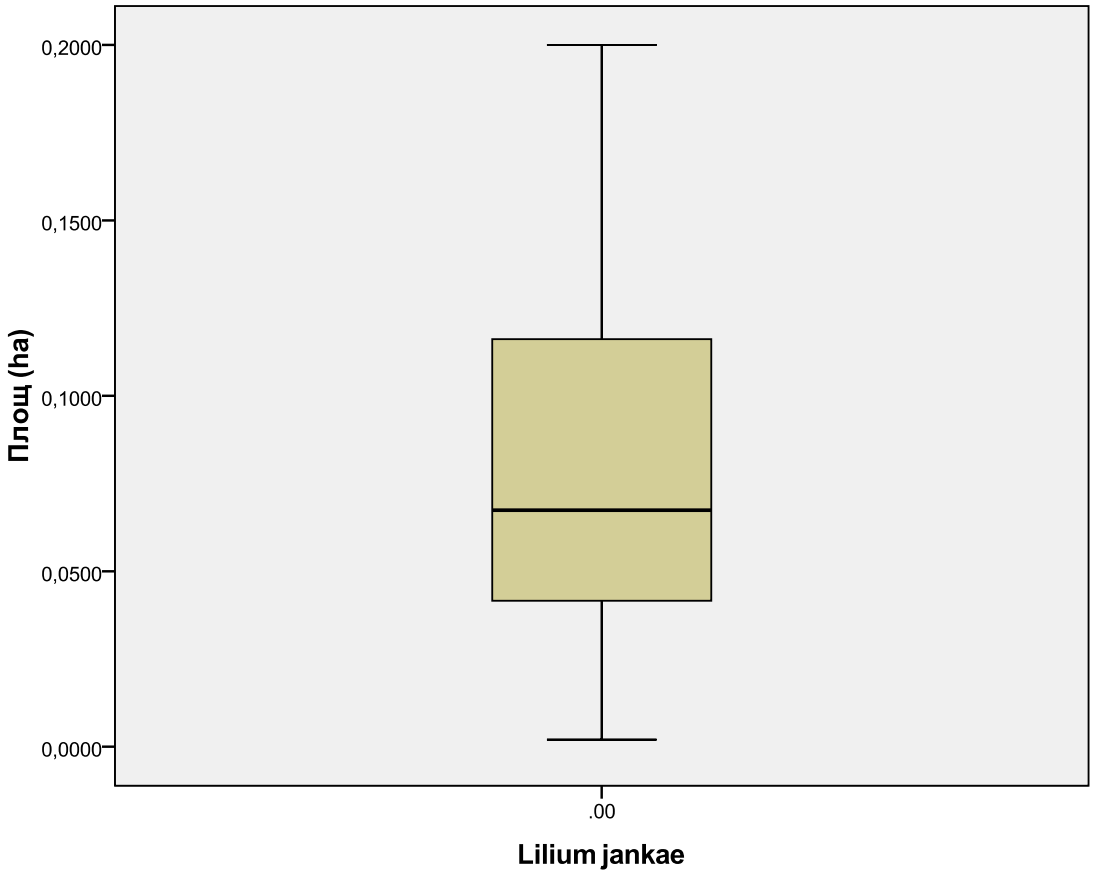
Средната плътност на популацията е 0,36 % ± 0,53% - силно вариращ белег. Стандартна грешка на средната 0,16. Средното покритие е 1,72% ± 1,34 %– също силно вариращ белег

Медианата (стойност на белега, която разделя извадката на две равни части) също е важен показател, когато съвпада със средната аритметична разпределението на белега е симетрично.

Коефициента на асиметрия (*Skewness*) показва отклонението на честотите от симетричното разпределение. В случая белегът площ е със слаба положителна асиметрия 0,84, плътността е с много силно изразена положителна асиметрия – 2,781, със силна положителна асиметрия е и разпределението на стойностите на покритието. Положителната асиметрия означава натрупване на стойности в дясната част на вариационната крива.

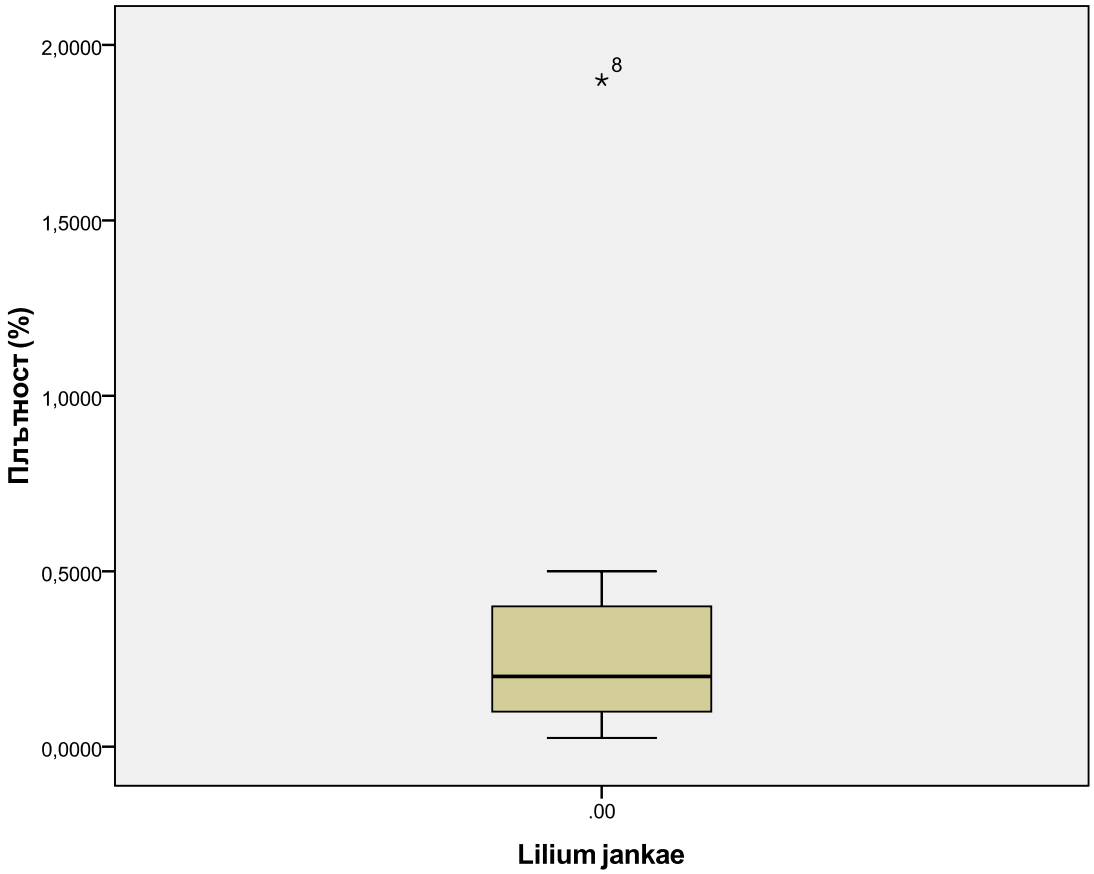
Коефициентът на ексцес (*Kurtosis*) показва степента на отклонение на върха на емпиричната крива от този на кривата на нормалното разпределение. Площта е с отрицателен коефициент на ексцес, което показва по-нисък връх, а на плътността и покритието е със силно отклоняващ се нагоре остър връх.

Особеностите в разпределението на белега може да се визуализират графично чрез хистограма, полигон на честотите и боксплот. В случай, че искаме да сравним белези боксплотовете може да се наредят в една графика, може и да са поотделно, когато ни интересува разпределението на един белег сам за себе си. В случая поради различните скали на белезите е по-подходящо всеки да е на отделна графика (Фиг. 17, 18 и 19). Но ако искаме да сравним разпределението на площта или някои от другите показатели през различни периоди на мониторинг или в различни флористични райони е желателно да са на една графика.



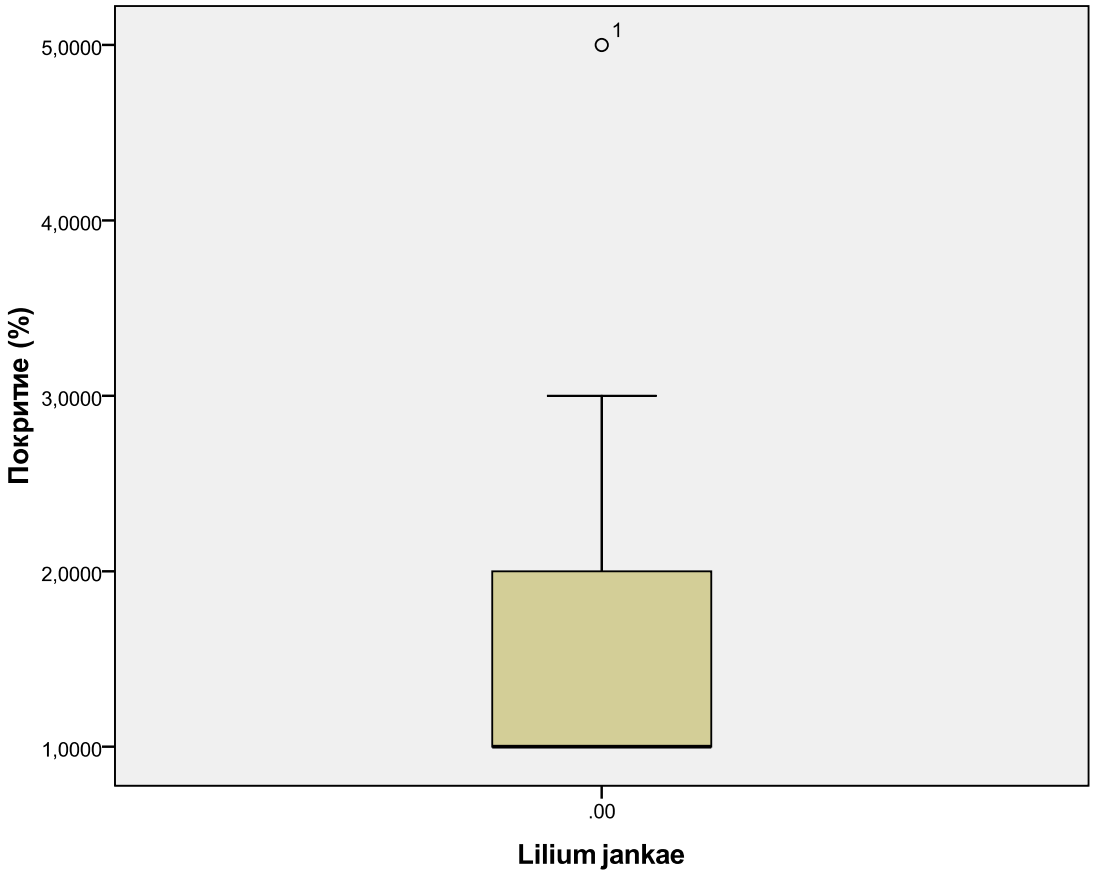
Фиг. 17. Box-plot графика на разпределението на стойностите на площта (в ha) на популацията на *Lilium jankae* в ПП Витоша.

От графиката се вижда асиметрия в разпределението на стойностите като има натрупване на честоти (повтарящи се стойности) в по-ниските стойности от медианата.



Фиг. 18. Box-plot графика на разпределението на стойностите на плътността (в %) на популацията на *Lilium jankae* в ПП Витоша.

От графиката се вижда асиметрия в разпределението на стойностите като има натрупване на честоти в по-ниските стойности от медианата. Силната асиметрия, големият коефициент на ексцес се вижда, че се дължат на една силно отклоняваща се от останалите стойност.



Фиг. 19. Box-plot графика на разпределението на стойностите на плътността (в %) на популацията на *Lilium jankae* в ПП Витоша.

От графиката се вижда силна асиметрия в разпределението на стойностите. Над 50% от тях са 1 (съвпада с медианата). Има и една силно отклоняваща се стойност (над 90-та процентила).

3.1.2. Тестове за сравняване на две и повече извадки.

Наблюдаваните белези за различните периоди на мониторинг може да се сравняват с тестове за достоверност на различията на свързани по двойки извадки или чрез дисперсионни анализи. Ако има достоверни разлики, следователно има промяна във времето. Тъй като ни интересува и посоката на промяната (т.е. знака на разликата) е необходимо да се извършват т.нар. one-tailed тестове.

Параметричните тестове изискват нормалност на разпределението и еднаквост на дисперсиите на изследваните белези. Ако това условие не е спазено, е необходимо данните да се трансформират (Табл. 9) или да се проведат непараметрични тестове, които обаче са с по-малка сила от параметричните.

Таблица 9. Най-често използваните трансформации на данни в биологични изследвания.

Тип трансформация	**Кога е подходящо да се използва**
Логаритъм	Използва се за меристични белези, когато средните аритметични корелират положително с дисперсиите. Правило е да се използват, когато най-голямата стойност на променливата е най-малко 10 пъти по-голяма от най-малката.
Корен квадратен	Използва се за меристични белези с Поасоново разпределение.
Реципрочна	Използва се, когато отклоненията residuals показват строго фуниевиден модел, често срещано при данни с много стойности близки до 0.
Аркусинус корен квадратен	Добър за пропорции или бинарни данни.
Box-Cox трансформация	Използва се при регресия.

Пример: Искаме да сравним числеността на популацията от диви кози в НП Централен Балкан през 2009 и 2010 год. (Може да се направи само при спазени условия за съставяне на сравними извадки или пълно преброяване – виж препоръките в т.1.). На цялата територия на Централен Балкан през 2009 и 2010 год съответно са наблюдавани 226 и 247 индивида. За да сравним числеността през двете години дали се различава достоверно трябва да използваме чифтен тест (Before and After) тъй като сравняваме по двойки маршрутите за двете години. (Използваните данни в примера за 2009 година са хипотетични, данните за 2010 година също са приети условно, тъй като не е спазена методиката)

Условие за използването на параметричен тест е данните да са с нормално разпределение. В случая това условие не е спазено и използваме непараметричния еквивалент – тест на Wilcoxon.

Таблица 10. Доклад на SPSS за проведен тест на Wilcoxon

Descriptive Statistics

	N	Mean	Std. Deviation	Minimum	Maximum
Брой 2009	42	5,3810	8,17776	,00	50,00
Брой 2010	42	5,8810	10,85874	,00	69,00

Ranks

		N	Mean Rank	Sum of Ranks
Брой 2010 - Брой 2009	Negative Ranks	11[a]	11,50	126,50
	Positive Ranks	13[b]	13,35	173,50
	Ties	18[c]		
	Total	42		

a. Брой 2010 < Брой 2009

b. Брой 2010 > Брой 2009

c. Брой 2010 = Брой 2009

Test Statistics[b]

	Брой 2010 - Брой 2009
Z	-,742[a]
Asymp. Sig. (2-tailed)	,458

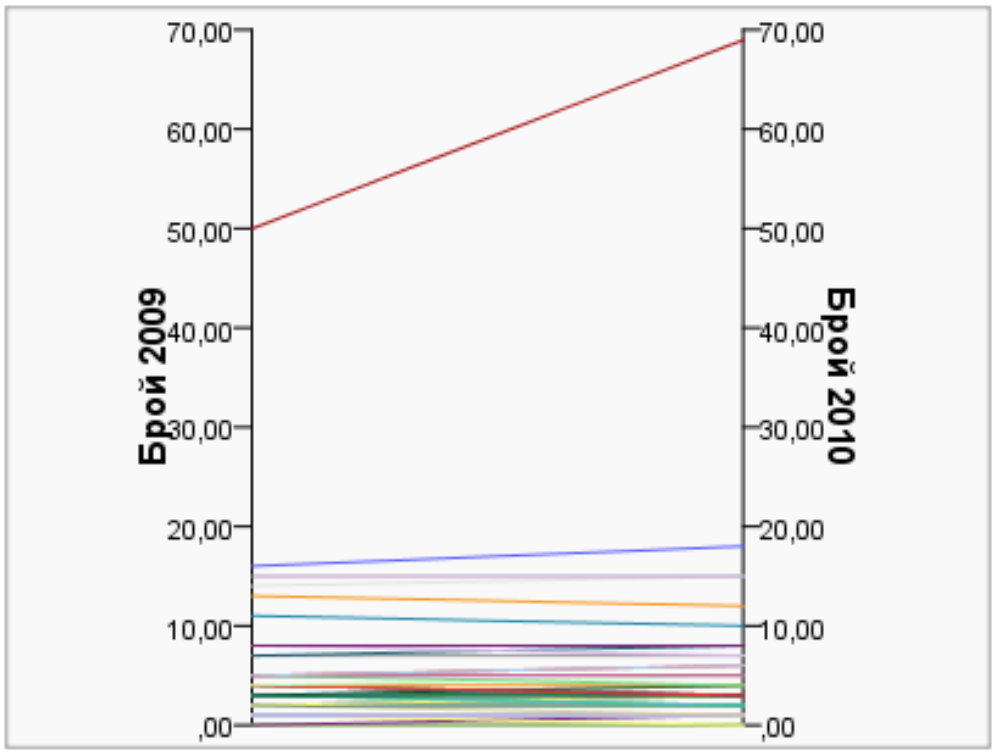
a. Based on negative ranks.

b. Wilcoxon Signed Ranks Test

Обратна връзка:

Винаги, когато се провежда тест е полезно да се даде и дескриптивна статистика за данните (Табл. 10) . В случая резултатът от теста, който ни интересува е, че P = 0.458 т.е. >0.05. Това означава, че няма достоверни различия в числеността за двете години и ако има разлики, то те са случайни.

Графично, промените в числеността изглеждат по следния начин (Фиг. 20).



Фиг. 20 Графика на промените в числеността. на дивите кози по наблюдаваните маршрути в Централен Балкан от 2009 до 2010 год.

Ако искаме да представим данни за всички изследвани места на ниво популация в България може да се направи дисперсионен анализ ANOVA еднофакторен, когато е за един период от време и многофакторен, когато е за повече периоди от време.

Пример:

Таблица 11. Обобщени резултати от мониторинг на дивата коза – есен 2010

Район	Брой установени индивиди	Брой маршрути
Пирин	259	15
Родопи	503	29
Рила	236	10
ЦБ	243	42

В работната таблица в SPSS се работи с данните за всеки маршрут от всяко място.

Ако знаем площта на изминатите маршрути, можем да определим плътността на популацията (брой индивиди на км^2). Искаме да представим и сравним плътността на популациите на дивите кози за есента на 2010 в различни планински находища. (Тъй като в момента не разполагам с данни за точната площ, нека приемем, че всички маршрути са с големина 10 км^2). Данните не са с нормално разпределение и поради тази причина използваме непараметричния еквивалент на еднофакторен дисперсионен анализ – Kruskal-Wallis (Табл. 12). В SPSS се изисква данните да са в една колона, а в съседната – групиращата променлива (в случая групите са 4 = четирите планини).

Таблица 12. Доклад на SPSS за проведен непараметричен еднофакторен дисперсионен анализ Kruskal-Wallis.

Ranks

	Groups	N	Mean Rank
Брой	Пирин	15	57,03
	Родопи	29	58,09
	Рила	10	50,70
	Централен Балкан	42	38,31
	Total	96	

Test Statistics[a,b]

	Брой
Chi-Square	10,591
df	3
Asymp. Sig.	,014

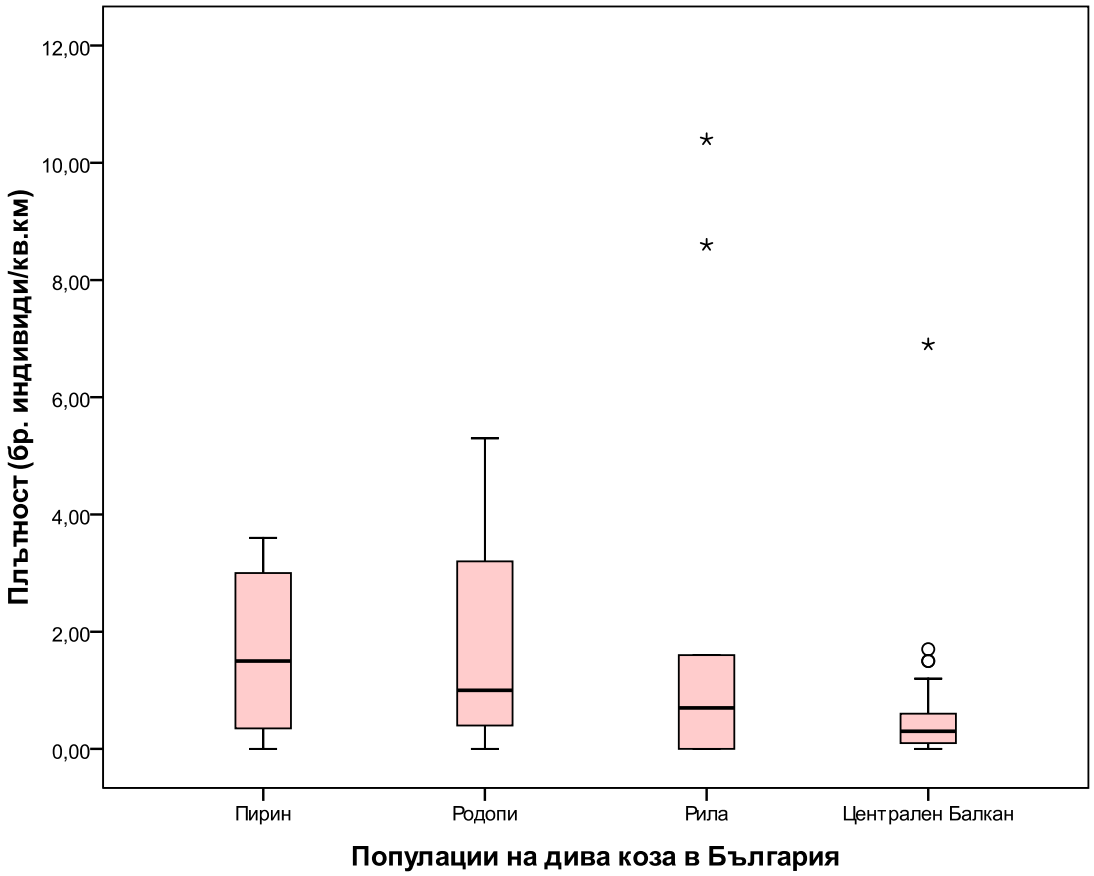
a. Kruskal Wallis Test

b. Grouping Variable: Groups

Обратна връзка:

Резултатите от теста показват достоверни различия в средната плътност на популациите по маршрути в различните планини. Установяваме по-големи разлики между групите (планините), отколкото вътре в групите (маршрутите) P=0.014.

Графично резултатите от непараметрични тестове се представят с бокс-плот графики, които сравняват медианите, дават границите на вариране на белега и симетрията в разпределението.



Фиг. 21. Плътност на наблюдаваните популации на дива коза в България през есента на 2010.

От графиката се виждат разликите в разпределението на стойностите – плътността, симетрията и границите на вариране за всяко от изследваните места. С най-голяма средна плътност (медиана) е НП Пирин, следвана от Родопите, Рила и Централен Балкан.

При доказване на достоверни различия, както е в случая е необходимо да се направи *post-hoc* тест по двойки.

Тъй като SPSS не ги предлага като опция при непараметрична ANOVA е необходимо да се направят отделно за всяка двойка. Подходящ тест за неравни извадки е тестът на

Dunnet. В случая обаче не тества всички двойки, поради тази причина двойките са сравнени ръчно с непараметричен тест - Mann-Witney (Табл.13)

Таблица 13. Резултат от проведен тест Mann-Witney. Сравняване по двойки на средната плътност на дивите кози в изследваните райони.

Сравнение	P=
Пирин - Родопи	0.931
Пирин - Рила	0.594
Пирин - Централен Балкан	0.027
Родопи - Рила	0.572
Родопи - Централен Балкан	0.003
Рила - Централен Балкан	0.248

Достоверни различия в средната плътност има между Пирин и Централен Балкан и между Родопи и Централен Балкан.

Изследваните места може да се представят и чрез **срещаемост на вида** (честота, *frequency*) – броят на пробите, в които даден вид се среща спрямо общия брой проби (Табл. 14). Зависи от плътността и разпределението на вида и служи за оценка на вероятността за намиране на екземпляри на даден вид.

Таблица 14. Срещаемост на дивата коза в районите на провеждан мониторинг

Район	Срещаемост
Пирин	73%
Родопи	86%
Рила	70%
Централен Балкан	90%

Може да се представи и графично. Срещаемостта на вида във всички находища е висока.

3.1.3 Регресионни анализи

Използват се за представяне на тенденциите на различните показатели с времето. Преди да приложим конкретен анализ е необходимо да се провери на кой модел най-много съответства връзката на дадения показател с времето.

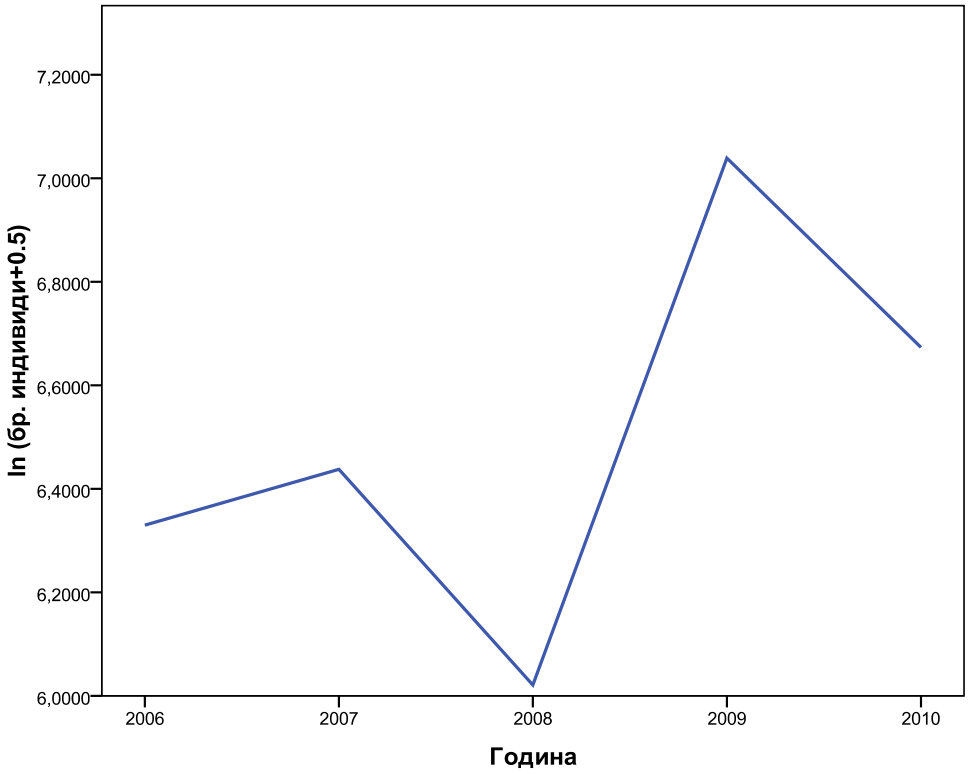
В SPSS това става от функцията *Curve estimation*.

- Линейна регресия – Използва се при проследяване на тенденциите в метрични показатели или меристични с нормално разпределение. За да се приложи за численост най-често се налага трансформация на данните ln (n+0.5).

Пример: Средно зимно преброяване на *Pelecanus crispus* в България.

Таблица 15. Изходни данни от Средно зимно преброяване на *Pelecanus crispus* в България.

Година	2006	2007	2008	2009	2010
Брой индивиди	561	625	412	1140	791
Ln(n+0.5)	6.33	6.44	6.02	7.04	6.67



Фиг. 22. Графика на промените в числеността на *Pelecanus crispus* в България за периода от 2006 до 2010 година.

Таблица 16. Доклад от SPSS за проведен анализ - линейна регресия за числеността на *Pelecanus crispus* в България спрямо годините на наблюдение.

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,534[a]	,285	,046	,3726487

a. Predictors: (Constant), Година

b. Dependent Variable: ln (бр. индивиди+0.5)

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	,166	1	,166	1,195	,354[a]
	Residual	,417	3	,139		
	Total	,583	4			

a. Predictors: (Constant), Година

b. Dependent Variable: ln (бр. индивиди+0.5)

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Correlations		
		B	Std. Error	Beta			Zero-order	Partial	Part
1	(Constant)	-252,150	236,627		-1,066	,365			
	Година	,129	,118	,534	1,093	,354	,534	,534	,534

a. Dependent Variable: ln (бр. индивиди+0.5)

Обратна връзка:

Резултатите показват, че моделът на регресия не обяснява вариацията на броя индивиди с годините. Не обяснява връзката между двете променливи. От стойностите на корелационния коефициент можем да видим, също че моделът не отразява връзка между двете променливи. Т.е. няма достоверна тенденция във времето.

- GENLIN модели. Тъй като повечето следени показатели обикновено са меристични белези в повечето случай се налага прилагане log-линейна регресия, а при анализ на присъствие и отсъствие логит-логлинейна. Тези анализи присъстват и в Generalized

Linear Models, където ще изчислим отново тенденциите в числеността без предварителна трансформация за *Pelecanus crispus,* както и възможностите за предвиждане на стойности от модела. Приемаме, че отклоненията са с Поасоново разпределение.

Тблица 17. Доклад от SPSS за проведен анализ – GENLIN (log-модел, Поасоново разпределение) за числеността на *Pelecanus crispus* в България спрямо годините на наблюдение.

Continuous Variable Information

		N	Minimum	Maximum	Mean	Std. Deviation
Dependent Variable	Брой индивиди	5	412,0000	1140,0000	705,800000	278,2098129
Covariate	Година	5	2006	2010	2008,00	1,581

Goodness of Fit[b]

	Value	df	Value/df
Deviance	285,029	3	95,010
Scaled Deviance	285,029	3	
Pearson Chi-Square	282,599	3	94,200
Scaled Pearson Chi-Square	282,599	3	
Log Likelihood[a]	-163,360		
Akaike's Information Criterion (AIC)	330,721		
Finite Sample Corrected AIC (AICC)	336,721		
Bayesian Information Criterion (BIC)	329,940		
Consistent AIC (CAIC)	331,940		

Dependent Variable: Брой индивиди

Model: (Intercept), VAR00001

a. The full log likelihood function is displayed and used in computing information criteria.

b. Information criteria are in small-is-better form.

Omnibus Test[a]

Likelihood Ratio Chi-Square	df	Sig.
135,250	1	,000

Dependent Variable: Брой индивиди

Model: (Intercept), VAR00001

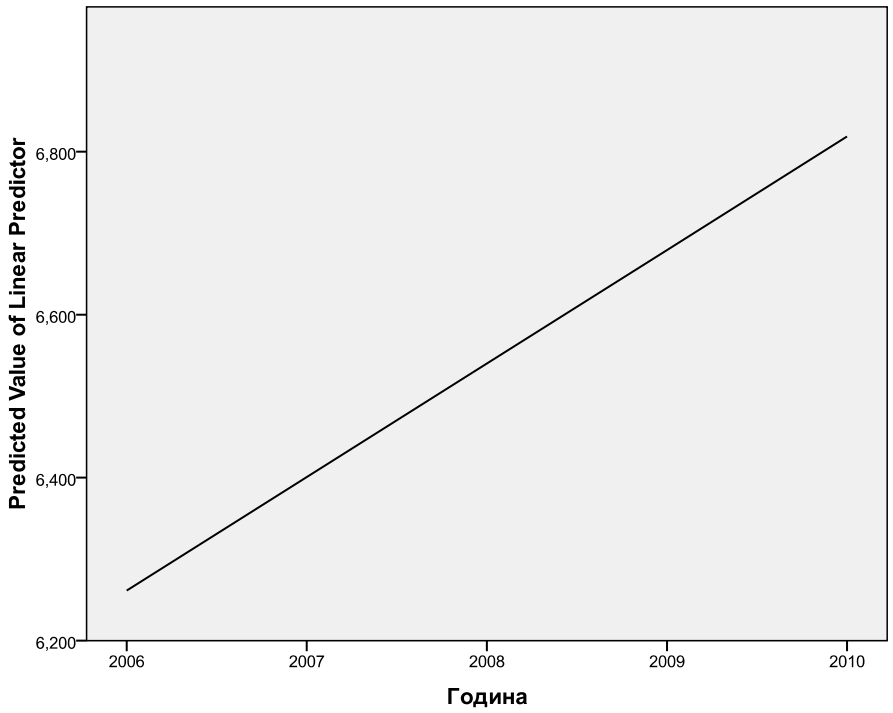
a. Compares the fitted model against the intercept-only model.

Tests of Model Effects

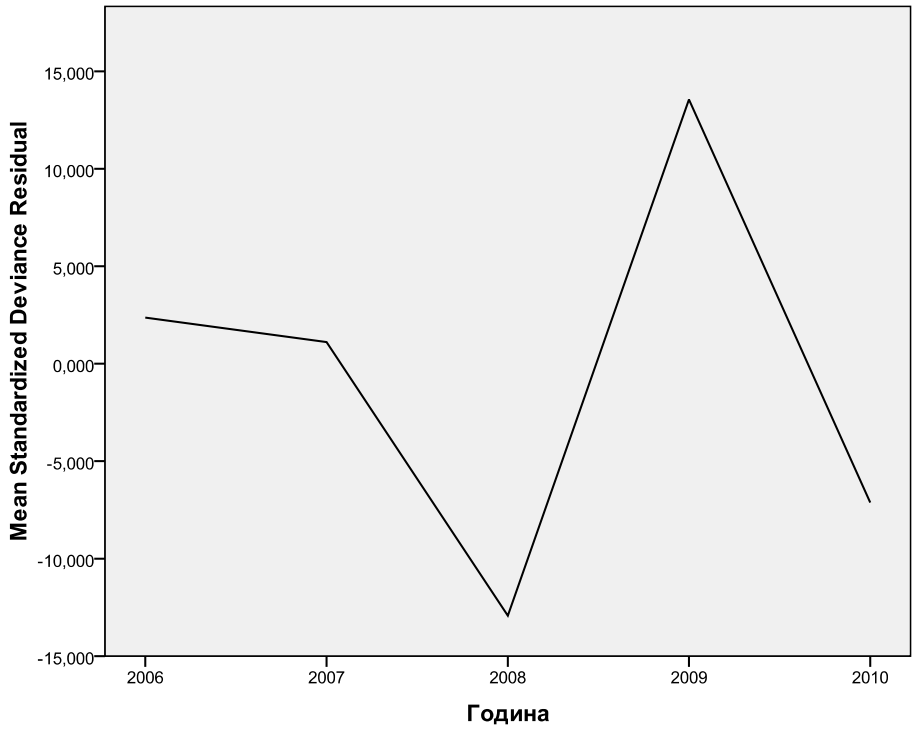
Source	Type III		
	Wald Chi-Square	df	Sig.
(Intercept)	127,358	1	,000
VAR00001	133,566	1	,000

Dependent Variable: Брой индивиди

Model: (Intercept), VAR00001



Фиг. 23. Графика на линейната регресия предсказана от модела.



Фиг. 24. Графика на стандартизираните отклонения от модела.

По данните от статистиката изглежда сякаш този модел пасва на данните за числеността през годините. Наклонът B е = 0.139. За да сме сигурни обаче трябва да сравним стойностите на log-likelihood за този модел с друг т.нар. нулев модел. В случая за меристични белези е този за отрицателно биномно разпределение.

Таблица 18. Доклад на SPSS за проведен анализ log- линеен модел с отрицателно биномно разпределение.

Goodness of Fit[b]

	Value	df	Value/df
Deviance	,400	3	,133
Scaled Deviance	,400	3	
Pearson Chi-Square	,376	3	,125
Scaled Pearson Chi-Square	,376	3	
Log Likelihood[a]	-37,705		
Akaike's Information Criterion (AIC)	79,409		
Finite Sample Corrected AIC (AICC)	85,409		
Bayesian Information Criterion (BIC)	78,628		
Consistent AIC (CAIC)	80,628		

Dependent Variable: Брой индивиди

Model: (Intercept), VAR00001

a. The full log likelihood function is displayed and used in computing information criteria.

b. Information criteria are in small-is-better form.

Model Information

Dependent Variable	Брой индивиди
Probability Distribution	Negative binomial (1)
Link Function	Log

Omnibus Test[a]

Likelihood Ratio Chi-Square	df	Sig.
,191	1	,662

Dependent Variable: Брой индивиди

Model: (Intercept), VAR00001

a. Compares the fitted model against the intercept-only model.

Tests of Model Effects

Source	Type III		
	Wald Chi-Square	df	Sig.
(Intercept)	,184	1	,668
VAR00001	,193	1	,660

Dependent Variable: Брой индивиди

Model: (Intercept), VAR00001

От това, че стойностите на *Log Likelihood* в отрицателно биномния модел са по-високи разбираме, че предишния модел всъщност не пасва по-добре от този.

Резултатите от този модел обаче показват, че ефектът на променливата време върху числеността не е достоверен. Това отново показва, че няма изразена тенденция. Наклонът B е = 0.138

Тъй като не винаги регресионните модели са чувствителни към промените във времето, в такъв случай би трябвало да се използва двуфакторен дисперсионен анализ за определяне на вариацията между различните места за мониторинг и годините. Резултатите също показват, че няма достоверни различия, както между годините така и между групите.

Source of Variation	**DF**	**SS**	**MS**	**F**	**P**
Район	3	318517.750	106172.583	2.500	0.109
Година	4	77400.700	19350.175	0.456	0.767
Residual	12	509704.500	42475.375		
Total	19	905622.950	47664.366		

Друг пример от литературата: Изследвана е популация на полски мишки. Данните за плътността за няколко години са следните:

Таблица 19. Изходни данни за плътността на популация на полски мишки.

Година	**Брой индивиди/ha**
1991	**485,35**
1992	**515,13**
1993	**411,65**
1994	**391,86**
1995	**361,50**

Таблица 20. Доклад на SPSS за проведен анализ – линейна регресия за тенденциите в плътността на популацията на полски мишки във времето.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,907[a]	,822	,762	31,5410283

a. Predictors: (Constant), Година

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	13761,874	1	13761,874	13,833	,034[a]
	Residual	2984,509	3	994,836		
	Total	16746,383	4			

a. Predictors: (Constant), Година

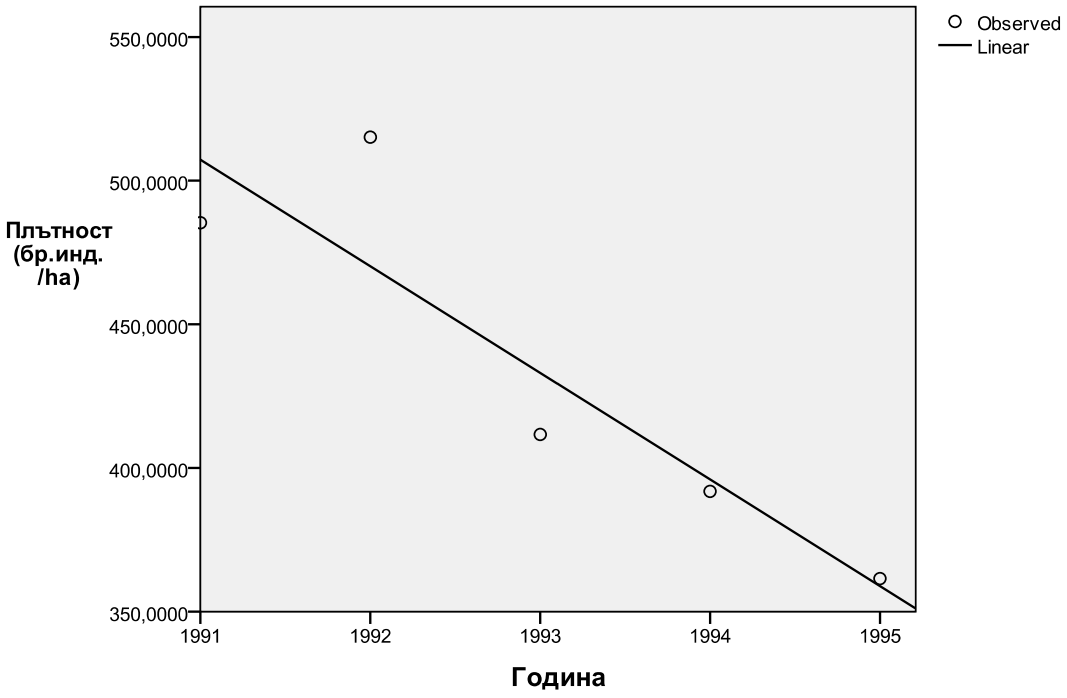
b. Dependent Variable: Плътност (бр.инд./ha)

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	74367,419	19878,484		3,741	,033
	Година	-37,097	9,974	-,907	-3,719	,034

a. Dependent Variable: Плътност (бр.инд./ha)

От статистиката на модела се вижда, че има ясна тенденция в плътността на полските мишки и тя е да намалява с годините. Моделът на регресия е достоверен – F статистика (P= 0.034), а коефициентът на корелация близък до 1. След време, ако тенденцията се запази ще доведе до измиране на популацията. Наклонът, на правата (коефициентът B) показва скоростта, с която плътността намалява е -37,097.



Фиг. 25. Линия на регресия, показваща отрицателна тенденция в плътността на полската мишка.

- **експоненциална регресия** – отново се използва за отчитане на тенденции и предвиждане на стойности на показател които намаляват или нарастват експоненциално с времето.

- За проследяване на влияние на фактори на околната среда в промените на числеността в различните периоди също може да се използва **GENLIN** в зависимост от разпределението на стойностите и спазването на условията за различните модели. (численост и време ще са зависима и независима променлива, факторите – са променливите оказващи влияние на числеността в различните периоди – може да си види кой фактор има най-голям принос за вариацията през годините.

- рандомизирани методи

Използват се, когато не са спазени изискванията на регресионните модели за хомогенност на дисперсията и нормално разпределение. Целта е чрез генериране на случайни комбинации на показателите и построяване на регресионни линии да се види колко пъти би се повторил по случайност същия наклон. Ако нашия наклон получен с оригиналните данни е по-малък от новия изчислен наклон означава, че съществува истинско намаляване на популацията. Изчислява се вероятността на нашия наклон.

За малки извадки с обем <6, рандомизацията не е с достатъчна сила, тъй като извадъчното разпределение на дадения параметър е малко.

3.1.4. Индекси за разнообразие

Използват се в мониторингови схеми, следящи разнообразието на определени съобщества в местата за мониторинг.

Индекси за α – биоразнообразие се използват за оценка на разнообразието в дадено съобщество от определено място и се използват за сравнение между различните съобщества или на едно и също съобщество за различни периоди.

Най-често използвани индекси са:

- Брой видове (видово богатство) – S,
- Общ брой индивиди – N
- Доминантност = 1 - индекс на Simpson. Варира от 0 (всички видове са еднакво представени) до 1 (един вид доминира напълно в съобществото)

$D=\Sigma(n_i/n^2)$, където n_i е брой индивиди от вида i.

- Индекс на Simpson = 1 – Доминантност. Измерва изравнеността на видовете по численост. Варира от 0 до 1 (пълна изравненост).

- Индекс на Shanon. Взима предвид както броя индивиди, така и броя видове. Варира от 0 – съобщества от 1 вид до високи стойности за съобщества с много видове с по малко индивиди.

$H= \Sigma((n_i/n)\ln(n_i/n))$

- Индекс на Margalef за видово разнообразие: (S-1)/ln(n), където S е брой видове, а n брой индивиди.

- Equitability. Индексът на Шанон разделен на логаритъм от броя видове. Оценява изравнеността с която индивидите са разпределени между видовете.

- Индекс на Fisher α. S=a*ln(1+n/a), където S е брой видове, n – брой индивиди и „a" е Fisher α.

Доверителни интервали на индексите се определят чрез Bootstrap метода.

Тези индекси не са налични в SPSS, но има freeware програми, с които могат да се изчисляват лесно.

Пример: Изследвано е растително дървесно съобщество в даден резерват в 5 кръгли полигона определени случайно с GPS точки и с по 5 случайно избрани квадрати в тях. Интересува ни разнообразието на дърветата и храстите в тези райони. Отчитан е видът и броя дървета и храсти на квадрат.

Таблица 21. Индекси на разнообразие за дървесните съобщества в резерват

	Полигон 1	Полигон 2	Полигон 3	Полигон 4	Полигон 5
Taxa_S	9	10	9	8	10
Individuals	124	270	89	121	144
Dominance_D	0.2215	0.2184	0.1428	0.2196	0.1126
Shannon_H	1.722	1.73	2.06	1.708	2.233
Simpson_1-D	0.7785	0.7816	0.8572	0.7804	0.8874
Evenness_e^H/S	0.6218	0.5642	0.8718	0.69	0.9325
Margalef	1.66	1.608	1.782	1.46	1.811
Equitability_J	0.7837	0.7514	0.9376	0.8216	0.9697
Fisher_alpha	2.23	2.045	2.5	1.924	2.443

Обратна връзка:

От S и N виждаме съответно броят видове и броят индивиди във всеки полигон.

Индексите на Shanon и Margalef за разнообразие отразяват съотношението на броя видове към тяхната представеност в съобществото. И двата индекса показват най-високо разнообразие в полигон 5 (най-голям брой видове, изравнени по обилие), както и относително високо разнообразие във всички полигони. Индексите за изравеност Evenness, Simpson и Equitability са обратно пропорционални на индекса за доминиране Dominance и показват висока изравеност във всички съобщества и най-висока съответно в полигон 5. Високата изравеност на видовете по обилие е характерна за старите и стабилни съобщества.

С промяната на индексите може да се следи за промените в съобществото. Появата на един или няколко силно доминиращи вида (високи стойности на Dominance), означава, че върху съобществото е имало някакво въздействие от външни фактори.

Промените в индексите за разнообразие през годините също може да се обработват статистически с дисперсионен и регресионен анализ.

3.1.5. Класификация и ординационни анализи на съобщества

В мониторинговите програми тези методи може да се използват при анализи на растителни съобщества или за определяне сходство на различни местообитания на видовете.

Едни от най-често използваните анализи са клъстерните, където чрез стойности на индекси за сходство става групиране на обектите по численост или по присъствие и отсъствие на вида. Графично се изобразяват чрез дендрограма. Освен клъстерните анализи други най-често използвани анализи са следните:

TWINSPAN (Two-way Indicator Species Analysis) – използва се при изследване основно на растителни съобщества, но също и при редица групи животни. Прави класификация на местата със съпровождаща класификация на видовете. В този анализ е развита концепцията за псевдовидовете. Числеността на всеки вид се кодира в брой на псевдовидовете. Напр. 1 отговаря на относителна численост от 0 до 2% ; 2 на 3 до 5 % относителна численост и т.н., 5 отговаря на обилие над 20%. По този начин псевдовидовете позволяват да се анализира едновременно присъствието/отсъствието на вида и неговата численост, тъй като са отделни променливи. С този анализ може да се определят индикаторни видове за дадено съобщество.

Ординационните анализи целят да организират (подредят) изследваните обекти (напр. пробни площадки, индивиди, видове) по значим (достоверен) градиент. Използват се за количествена оценка на връзките между голям брой взаимозависими променливи и да ги обясни чрез съставяне на по-малка мрежа от подчинени стойности. Подходът включва кондензиране на информацията, съдържаща се в оригиналните променливи в по-малка мрежа от стойности – компоненти така, че да има минимална загуба на информация. Графично се изобразяват като координатна система с разположени точки, отразяващи групирането на обектите по градиентите.

PCA (Principal components analysis) създава линейни комбинации от оригиналните променливи (т. нар. Principal components), които са ориентирани в направления, които описват максималната вариация между обектите. Правейки това обектите се подреждат по непрекъснати градиенти, определени от главните компоненти и се опитват да открият източниците на най-голяма вариация между наблюденията. Обектите представляват единична случайна извадка от познат или непознат брой генерални съвкупности. Данните трябва да се състоят от два или повече непрекъснати количествени белега или от комбинация от качествени и количествени белези. В тези анализи не се прави разлика между зависими и независими променливи.

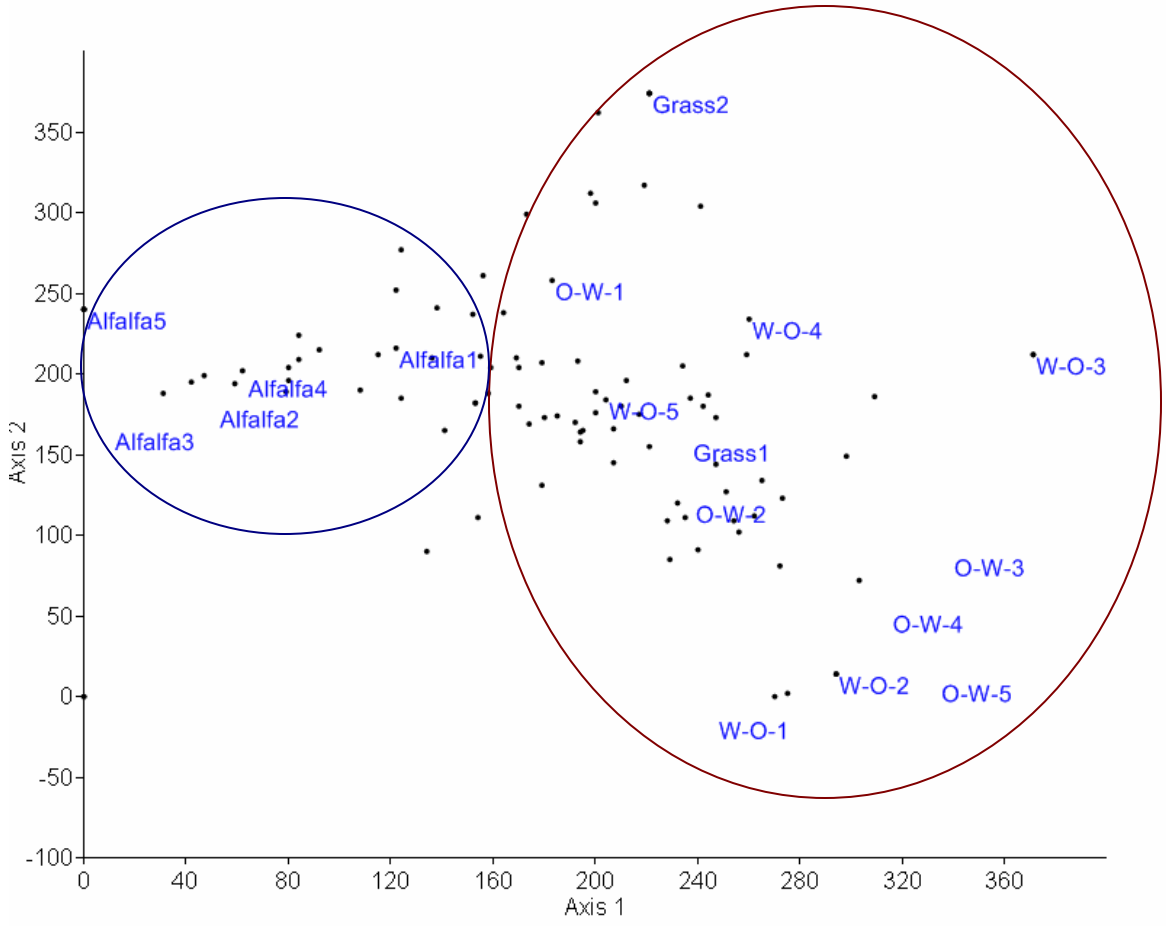
Eigenvalue – стойностите представляват дисперсията на съответния главен компонент. Дават оценка на степента на вариация между обектите по протежение на главния компонент. Всяка eigenvalue е свързана с един главен компонент. Колкото е по-висока тази стойност, толкова компонента е по-значим (обяснява по-голямо количество вариация). Стойности близки до 0 показват, че съответния компонент има минимална обяснителна стойност.

CA (Correspondence анализ) – подобен на PCA, но се използва предимно за качествени белези. eigenvalue – стойностите отразяват степента на сходство между променливите и обектите. Висока eigenvalue показва дълъг и достоверен градиент.

DCA (Detrended correspondence анализ) – подобен на CA – анализа. В основата си е кореспондентен анализ с две допълнителни стъпки, които премахват т. нар. arch-ефект на изкривяване на разпределение на стойностите и т. нар. компресия на осите. Използва се, когато белезите нямат нормално разпределение.

CA-CCA (Canonical correspondence анализ) – Използва се при анализи на численост и други характеристики на вида с включване на характеристики на местообитанията. Представлява хибриден метод между ординация и множествена регресия.
Много често тези анализи се прилагат при много голямо количество променливи, в случаи, в които не могат да се приложат другите стандартни анализи.

Пример: Данните са по дисертация на Костова (2004). Изследване на съобщества от бръмбари бегачи (Coleoptera) в различни агроценози – зимна пшеница, люцерна и овес. Втората година е извършен сеитбооборот на двете житни култури. Данните са от 17 пробни площадки с по 10 почвени капана в тях за 2 години пробовземане. Направен е Detrended correspondence анализ на видовете по площадки и численост, тъй като стойностите не са с нормално разпределение.



Фиг. 26. Detrended correspondence анализ на видовете бръмбари бегачи по площадки и относителна численост.

Oc	Eigenvalue
1	0.378
2	0.205
3	0.07

Обратна връзка:

С най-висока обяснителна стойност са градиентите по първата и втората ос, третата е с незначителна обяснителна стойност.

С точки е дадено подреждането на видовете, а с етикети на полигоните.

Групирането по първата ос е по тип растителност - в една група се обединяват полигоните от житните култури (пшеница първа година, овес втора W-O; овес първа година, пшеница втора O-W и синурите с преобладаваща тревиста житна растителност Grass), а във втората са полигоните от люцерната Alfalfa.

Втората ос не дава видимо биологично обяснение на групирането. Когато се поставят етикетите с подреждането и на видовете до някъде може да се предположи, че градиента е свързан с влажността на местообитанията.

По принцип от тези анализи също може да се видят кои видове са свързани с определено място, могат да го характеризират и да се използват като индикатори за дадения хабитат.

3.1.6. Оценка на ефективността

Прави се за всеки изследван вид или хабитат. Постигнати ли са целите на мониторинга. Сравними ли са получените резултати. Оценка на използваните ресурси за постигане на целите – оценка на ефективността на използваните методи (с криви на ефективността от т.1), индикаторни видове, човешки ресурс и финанси.

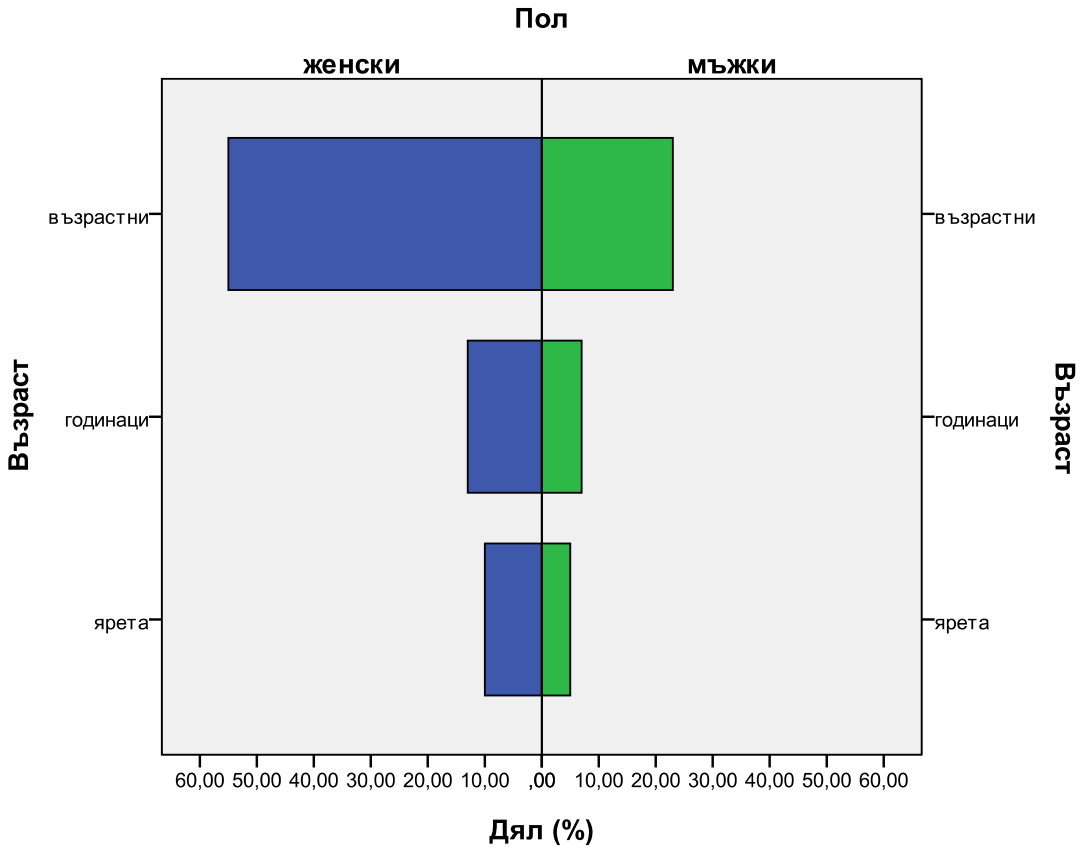
3.2. Полова и възрастова структура на популациите на отделни видове

Подходящи данни за тези анализи са от формулярите за земноводни и влечуги, дива коза и благороден елен.

3.2.1. Графично представяне и сравняване на полова и възрастова структура.

Важна характеристика на популацията, която оказва влияние на раждаемостта и смъртността в една популация.

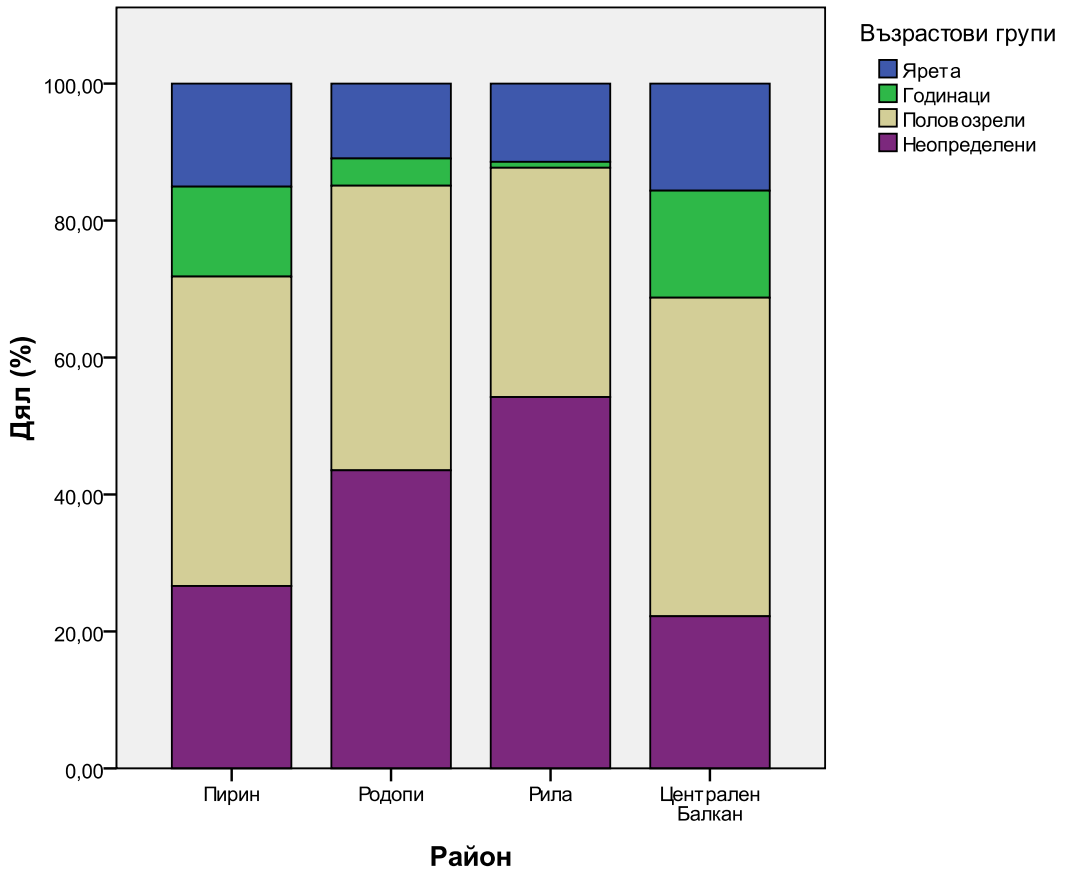
Най-често графично възрастовата структура се представя чрез възрастови пирамиди изразяващи процентното съотношение между индивидите по пол в отделни възрастови групи.



Фиг. 27. Разпределение на възрастовите категории по пол на дивата коза в Пирин планина.

Обратна връзка: Такъв тип пирамида показва бавно нарастване на популацията. Пирамидата ще има друг вид, ако категорията възрастни индивиди се раздели на в репродуктивна възраст и пострепродуктивни. Делът на младите индивиди е по-голям при бързо нарастващи популации.

Чрез построяване на графика може да се проследи разликата в дяловете в отделните популации на вида:



Фиг. 28. Разпределение на възрастовите групи в популацията на диви кози в България.

От графиката се вижда процентното разпределение на възрастовите категории.
Дяловете се сравняват чрез **z-тест,** за да се покаже дали нарастват или намаляват през годините.

3.2.2. Жизнени таблици

Половата, възрастовата структура, раждаемост и смъртност на популациите се анализират чрез жизнени таблици.
Жизнените таблици основно се базират на популационни данни организирани по възраст, представена обикновено чрез отделни възрастови класове, най-често с еднаква дължина. Могат да бъдат конструирани и за данни организирани около жизнени

стадии (яйце, ларва и т.н.), които могат и да не бъдат с еднаква продължителност. В други случай индивидите могат да бъдат групирани по размери, тогава съответните изчисления са базирани на вероятността индивидите да оцелеят от един клас размери до следващия и средния брой индивиди от поколенията продуциран от индивидите в даден клас размери. Често към тези данни се прибавя и смъртността за даден период, възраст, клас.

В зависимост от характера на данните могат да се направят два типа жизнени таблици: *cohort life table* или *statistical life table*.

Кохортните (демографски) жизнени таблици се изработват, когато могат да се наблюдават всички събития в живота на група индивиди, родени по едно и също време през целия им живот. При такива демографски изследвания за всеки от индивидите в кохортата (групата) се знае възрастта на първия репродуктивен цикъл, броя на репродуктивните цикли и възрастта, при която настъпва смърт. Броят индивиди, продуциран от един индивид в групата може да бъде установен директно чрез наблюдение или може да се знае по принцип познавайки биологията на вида.

За съжаление, особено за дълго живеещите видове, не винаги е възможно да се съберат такива данни. Познанията за връзките между възрастта и смъртността могат да бъдат получени чрез изследване на възрастовата структура на популацията - построяват се т. нар. статистически жизнени таблици.

Основните изчисления при статистическите и кохортни жизнени таблици са едни и същи. Данните се подреждат в колони (Табл.22).

Таблица 22. Принцип на изчисленията в жизнените таблици, включващи смъртност и преживяемост.

Възрастов клас (напр.години) (x)	Брой живи индивиди (n_x)	Преживяемост ($l_x=n_x/n_0$)	Брой умрели индивиди през дад. възрастов интервал ($d_x = n_x-n_{x+1}$)	Възрастово специфична смъртност ($q_x = d_x/n_x$)	$\log_{10} n_x$	Killing power ($k_x= \log_{10} n_x - \log_{10} n_{x+1}$)
0	250	250/250=1.00	250-120=130	130/250=0.52	2.398	2.398-2.079= 0.319
2	120	120/250=0.48	120-75=45	45/120=0.38	2.079	2.079-1.875= 0.204
3	75	75/250=0.30	75-55=20	20/75=0.27	1.875	1.875-1.740= 0.135
4	55	55/250=0.22	55-0=55	а) 55/55=1.00	1.740	б)
5	0	0/250=0.00	-	-	-	в)Σk_x=0.658

а) Тъй като всички индивиди в този период (4) са умрели, вероятността за смърт е = 1.00

б) Не може да се изчисли, тъй като $\log_{10} 0$ е неопределимо

в) Σk_x – обща смъртност за поколението (групата)

Възрастово специфичната смъртност е равна на вероятността един индивид да умре в рамките на даден възрастов клас.

Общата смъртност за поколението K (Σk_x) се получава от сумирането на т.нар. *killing power* (к-фактор) за всеки възрастов клас. Този адитивен показател е особено полезен, когато са получени стойностите на общата смъртност за поколението и *к*-фактора за даден жизнен стадий или възрастов клас за няколко групи. Чрез корелация между стойностите на *к*-фактора и общата смъртност на поколението е възможно да се установи относителния принос на всеки от тях към общата смъртност на популацията. Корелация между общата плътност на популацията, плътността на дадения жизнен стадий на групата или възрастов клас и общата смъртност на групата и отделните *к*-фактори може да осигури доказателства за зависима смъртност. Загубата на потенциално поколение може също да се разглежда като фактор за смъртност. *к*-факторът за редуцирана плодовитост е равен на $\log_{10}$ (потенциална плодовитост) - $\log_{10}$ (реална плодовитост). Чрез вкарване на редукции в плодовитостта при к-факторния анализ, тя може да се анализира по подобен начин както другите форми на смъртност, които се появавт през живота на дад. индивид.

Ако са налични данни за броя на поколенията, които се продуцират от индивидите във всеки възрастов клас, може да се направи таблица на плодовитостта.

Таблица 23. Принцип на изчисленията в жизнените таблици, включващи плодовитост и популационен растеж.

Възрастов клас (напр.години) (x)	Брой живи индивиди (n_x)	Преживяемост (l_x)	Среден брой женски в поколение (m_x), продуцирани от една женска на възраст x	Принос на всеки възрастов клас към R_0 (l_x m_x)	x l_x m_x
0	250	1.00	0	0	0
2	120	0.48	125	60	120

Възрастов клас (напр.години) (x)	Брой живи индивиди (n_x)	Преживяемост (l_x)	Среден брой женски в поколение (m_x), продуцирани от една женска на възраст x	Принос на всеки възрастов клас към R_0 (l_x m_x)	x l_x m_x
3	75	0.30	300	90	270
4	55	0.22	175	38.5	154
5	0	0.00	-	0	0

$R_0=\Sigma\ l_x\ m_x=248.5$ $\Sigma x\ l_x\ m_x=544$

Колоната m_x съдържа **възрастово специфичната плодовитост** или раждаемост, които са равни на средния брой женски от поколение, продуцирано от една женска на възраст *x*. При наличие на данни за възрастово специфичната плодовитост може да се изчисли **чиста (нетна) репродуктивна стойност** (R_0).

Влиянието на смъртността върху възпроизводството се включва чрез стойностите за преживяемост – *lx*. Общият брой на продуцираните поколения зависи от броя оцелели членове на групата до всеки възрастов клас и средния брой индивиди от поколението продуцирано от женските в тази възраст. За видовете, които се размножават един път, R_0 съответства на R (**годишна репродуктивна стойност**). Стабилен размер на популацията се появява тогава, когато R=1, ако R>1 популацията ще се увеличи и ако R<1 популацията ще намалее. За популации със застъпващи се поколения, изчислението на R е малко-по-сложно. **Средното време за поколение** (Tc), което може да се определи като средна възраст на родителите на всички поколения продуцирани от членовете на групата се изчислява по формулата:

$$Tc = \frac{\sum x l_x m_x}{\sum l_x m_x} = \frac{\sum x l_x m_x}{R_0}$$

За да се установи R е нужно да имаме предвид популационния растеж. Взимайки най-простия модел на популационен растеж, размерът на популацията на следващата година ще зависи от броя индивиди през настоящата година и средния

брой индивиди от поколението продуцирано от един индивид; на математически език:

$$N_1 = N_0R$$

$$N_2 = N_1R = N_0R^2$$

$$N_3 = N_2R = N_0R^3$$

$$[1] \ldots \ N_T = N_0R^T$$

Където степента на нарастване на поколението (R_0) и времето за едно поколение (T_C) са познати, размерът на популацията след едно пълно поколение се дава от:

$$N_T=N_0R_0$$

Използвайки уравнение [1], може да се изведе второ уравнение за същата популация от гледна точка на годишната степен на нарастване (R) и времето за 1 поколение (Tc): $N_T=N_0R^{Tc}$

При сравнение на двете уравнения е ясно, че:

$$N_T=N_0R_0= N_0R^{Tc} \text{ и } R_0= R^{Tc}$$

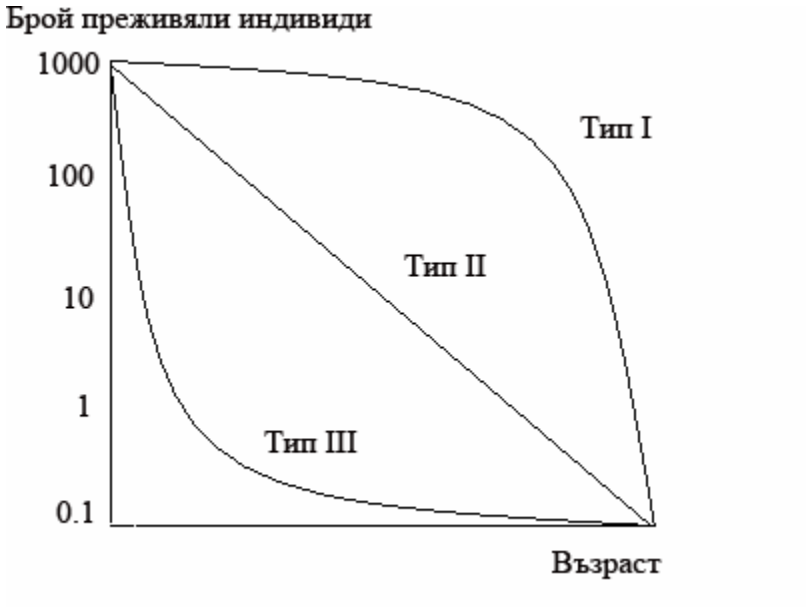
Логаритмувайки от двете страни на уравнението с натурален логаритъм се получава **специфичната скорост на нарастване** (r), константата в **диференциалното уравнение за популационен растеж**: dN/dt = rN:

$$lnR_0= T_clnR$$

$$lnR= lnR_0/T_c$$

- Съотнасяне на възрастта към смъртността

Общата връзка между смъртността и възрастта може да се види от кривите на преживяемост, построени чрез съпоставяне на логаритъм от броя оцелели индивиди (n_x) или стойностите на lx срещу възрастта (x) или времето (t). Стойностите на n_x и lx често се поставят, така че първата стойност при x=0 или t=0 е 1000. Три типа криви:



Фиг. 29. Хипотетични криви на преживяемост. Тип I – преживяемостта първоначално е голяма, след което рязко намалява (с такъв тип обикновено са бозайници и други висши животни, които проявяват грижа за поколението. Тип II – рискът от смърт е еднакъв през цялото време (с такъв тип са много видове птици, както и тревисти растения в стабилни хабитати). Тип III – висока смъртност в ранните възрастови класове (някои насекоми, видове с пелагични ювенилни стадии като бентосни безгръбначни, мекотели, ракообразни и риби).

3.2.3. Анализ на преживяемост и оценка на риска с SPSS:

За да се използват статистическите методи за преживяемост и оценка на риска в SPSS трябва да разполагаме със следния вид на извадките:

- Начален брой индивиди от даден възрастов клас – Напр. при коза годинаци, ярета, възрастни (женски и мъжки). Таблиците може да се правят за всеки възрастов клас, може да се правят и сравнения на преживяемост например по пол.
- за всеки един индивид се отчита статус за целия период на изследване – 0 (жив) и 1 (умрял). За умрелите индивиди се знае в кой период се е случило крайното събитие, т.е. за колко време са били живи. Тези, за които не се знае дали са живи или мъртви са цензурирани случаи. За тях отново се отчита времето, в което са били налични.

- Целият период на изследване да е подразделен на под периоди – напр. – изследване на преживяемост на възрастните индивиди за общ период 5 години се подразделя на сезони – пролет и есен - общо 10 сезона за петте години. При яретата и годинаците периодът на изследване ще е максимум една година, а отделните подпериоди ще са месеци.

Пример. Взимаме за пример хипотетични данни за козата, основани на данни от мониторинга 2009, 2010. Тъй като мъжките са солитарни, те биха били проследени доста трудно. Спираме се на стада от женски с малки в Рила и Пирин планина. Отчитаме женските на възраст над 2 години. Начален брой Рила есен 2009 – 69 женски. В края на 5 годишния период (2013 год.) от тях са останали 75 женски, като за 4 женски няма информация след определен период от време. Начален брой в Пирин – 51, в края на пет годишния период са останали 22 женски, от тях за 5 женски няма информация.

Таблица 23. Изходна таблица за анализ в SPSS за изследваната група диви кози в Рила планина:

Женска	Брой сезони, в които е била жива или до които е наблюдавана	Статус (жива мъртва)
1	5	0
2	2	0
3	10	1
4	5	1*
5	10	1
6	10	1
7	10	1
8	10	1
9	1	0
10	2	0
11	3	0
,	,	,
,	,	,
,	,	,
63	8	0
64	8	0
65	10	1
66	10	1
67	8	0
68	10	1
69	6	0

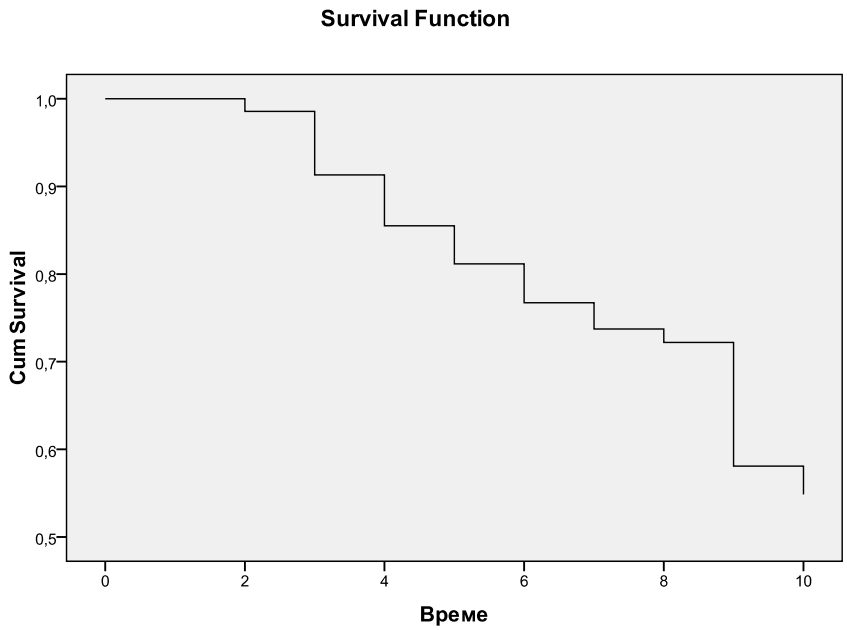
* Цензуриран случай – след петия период не знаем какво се е случило с този индивид.

Можем да използваме или обикновените жизнени таблици, или Kaplan-Meier в зависимост от това дали ни интересува точният момент на настъпване на крайното събитие и на цензурираните случаи. Резултатите от двата анализа са сходни:

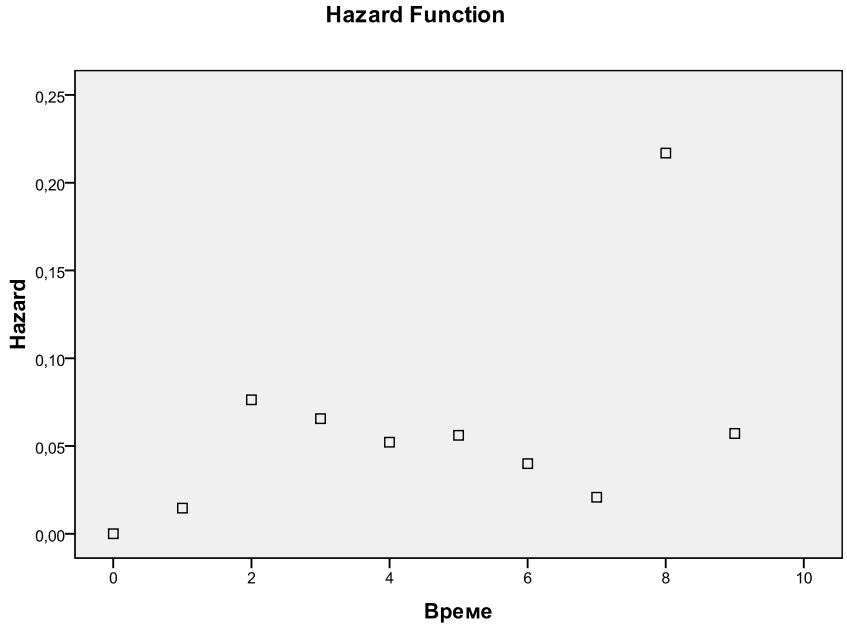
Таблица 24. Доклад на SPSS от *Life table* анализ.

Interval Start Time	Number Entering Interval	Number Withdrawing during Interval	Number Exposed to Risk	Number of Terminal Events	Proportion Terminating	Proportion Surviving	Cumulative Proportion Surviving at End of Interval	Std. Error of Cumulative Proportion Surviving at End of Interval	Probability Density	Std. Error of Probability Density	Hazard Rate	Std. Error of Hazard Rate
Интервали начално време	Брой индивиди В началото на интервала	Брой изгубени (цензурирани) индивиди по време на интервала	Брой индивиди изложени на риск	Брой на крайното събитие (смърт) за интервал	Дял на умрелите индивиди	Дял на преживелите индивиди	Кумулативен дял на оцелелите в края на интервала	Стандартна грешка	Плътност на вероятността	Стандартна грешка	Степен н ариск	Стандартна грешка
0	69	0	69,000	0	,00	1,00	1,00	,00	,000	,000	,00	,00
1	69	0	69,000	1	,01	,99	,99	,01	,014	,014	,01	,01
2	68	0	68,000	5	,07	,93	,91	,03	,072	,031	,08	,03
3	63	0	63,000	4	,06	,94	,86	,04	,058	,028	,07	,03
4	59	0	59,000	3	,05	,95	,81	,05	,043	,025	,05	,03
5	56	2	55,000	3	,05	,95	,77	,05	,044	,025	,06	,03
6	51	0	51,000	2	,04	,96	,74	,05	,030	,021	,04	,03
7	49	1	48,500	1	,02	,98	,72	,05	,015	,015	,02	,02
8	47	2	46,000	9	,20	,80	,58	,06	,141	,044	,22	,07
9	36	0	36,000	2	,06	,94	,55	,06	,032	,022	,06	,04
10	34	34	17,000	0	,00	1,00	,55	,06	,000	,000	,00	,00

a. The median survival time is 10.000



Фиг. 30. Крива на преживяемостта за група от женски индивиди от популацията на Рила планина.



Фиг. 31. Разпределение на риска за група от женски индивиди от популацията на Рила планина.

Обратна връзка: средното време за преживяемост (медиана) е през всичките 10 периода. До края на периода от 5 години преживяват 55 % от индивидите. Най - висока степен на риск има в осми период – (есен 2012)

Таблица 25. Резултати от Kaplan-Meier анализ за една група женски индивиди от популацията на Рила планина.

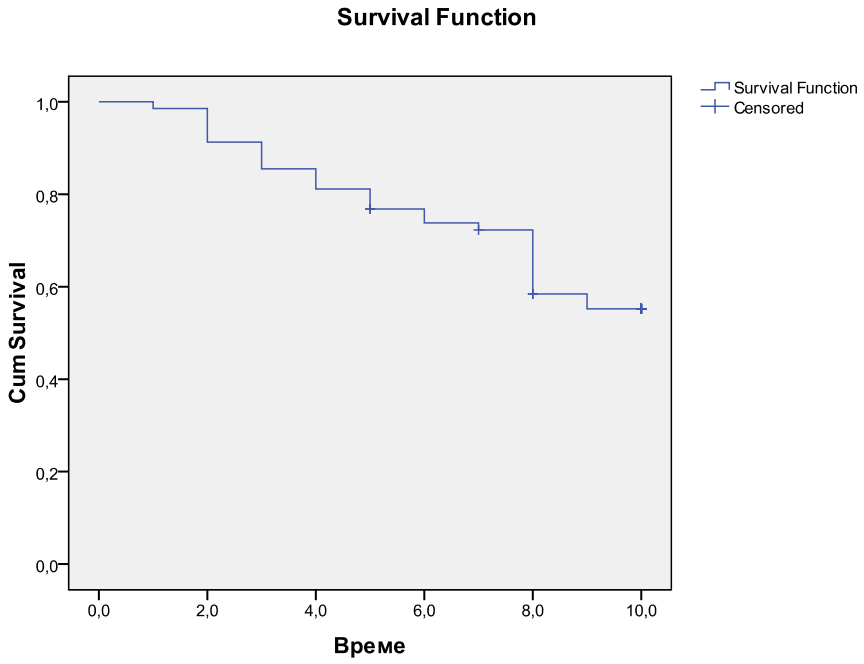
Case Processing Summary

Total N	N of Events	Censored	
		N	Percent
69	30	39	56,5%

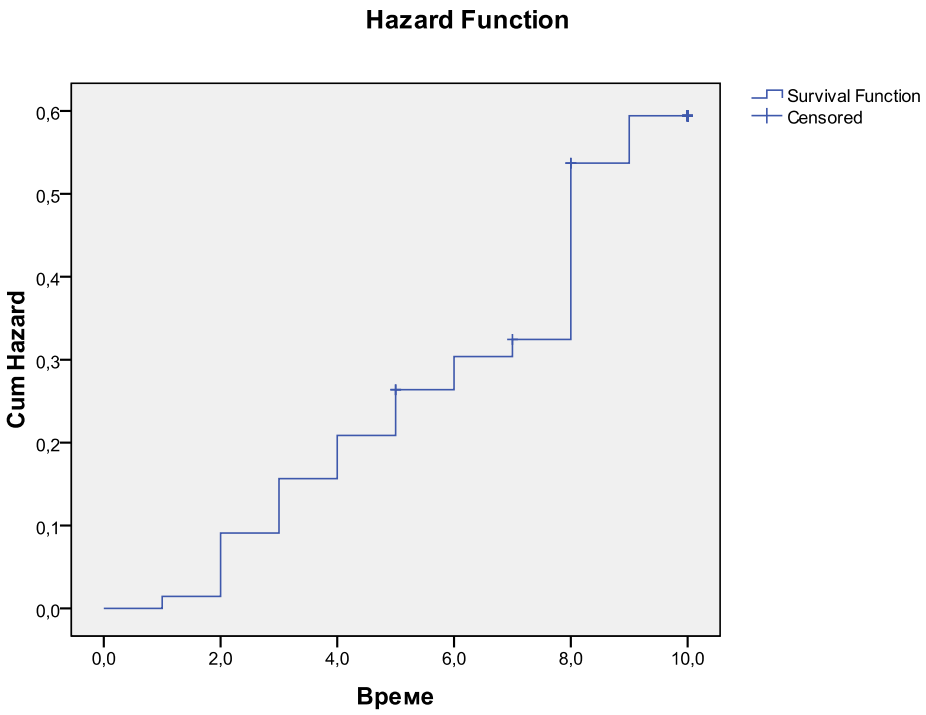
Means and Medians for Survival Time

Mean[a]				Median			
		95% Confidence Interval				95% Confidence Interval	
Estimate	Std. Error	Lower Bound	Upper Bound	Estimate	Std. Error	Lower Bound	Upper Bound
7,931	,346	7,252	8,610			.	.

a. Estimation is limited to the largest survival time if it is censored.



Фиг. 32. Kaplan-Meier крива на преживяемостта за група от женски индивиди от популацията на Рила планина.



Фиг. 33. Kaplan-Meier крива на риска за група от женски индивиди от популацията на Рила планина.

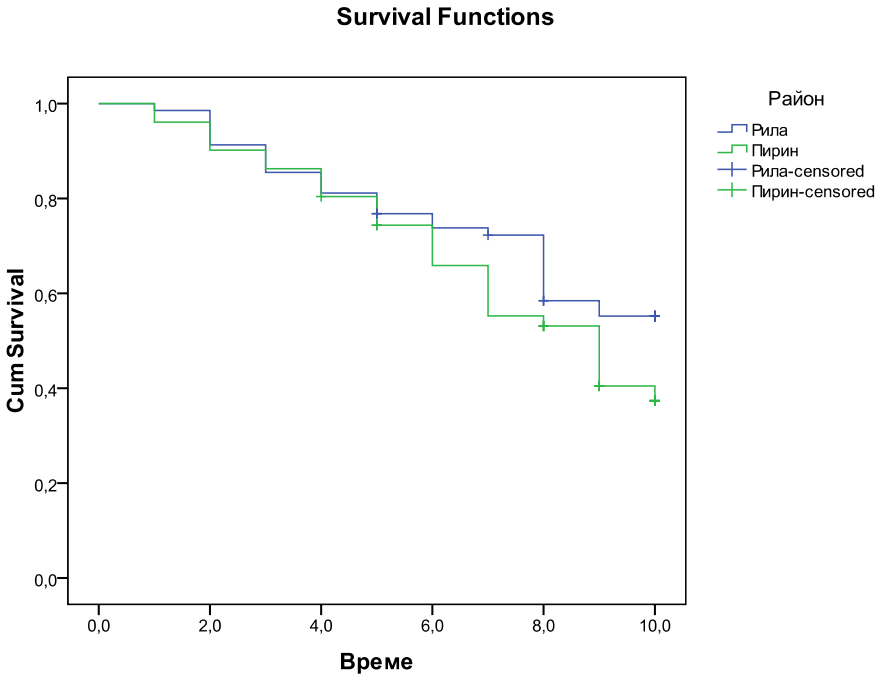
Обратна връзка: средното време за преживяемост (средна аритметична) е 8 периода. До 10-ти период са преживели 56,5 % от индивидите. Най- висока степен на риск има в осми период – (есен 2012)

Ако искаме да сравним преживяемостта на дивата коза в два района отново може да използваме двата вида анализи. Прибавяме данните за втората група (Пирин) под първата като в отделна групираща променлива се посочват районите, в диалоговия прозорец тази променлива се посочва като фактор.

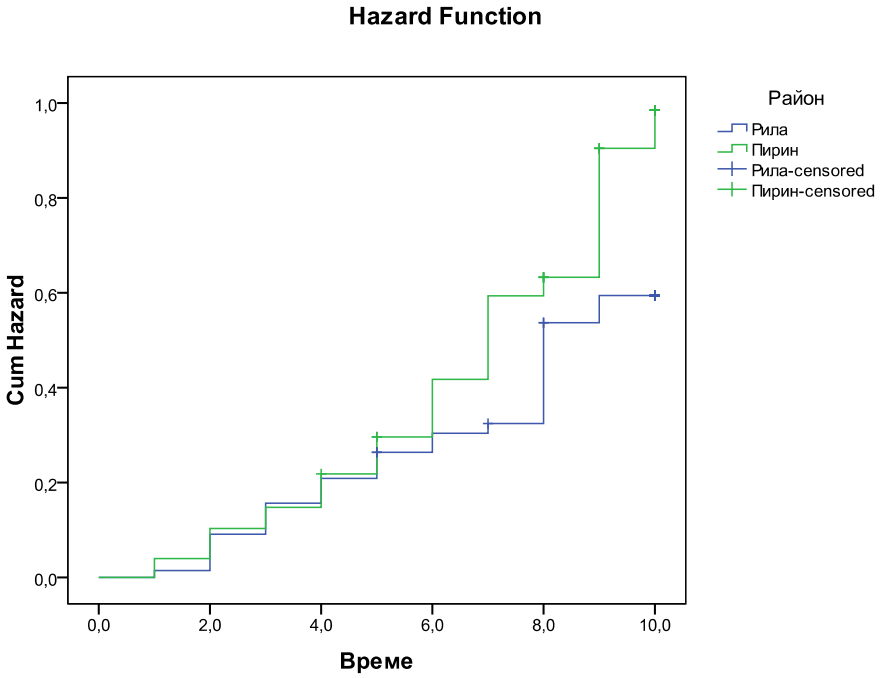
Таблица 26. Резултати от анализа Kaplan-Meier за преживяемост на женските в две групи диви кози – в Рила и Пирин.

Case Processing Summary

Grupa	Total N	N of Events	Censored	
			N	Percent
Рила	69	30	39	56,5%
Пирин	51	29	22	43,1%
Overall	120	59	61	50,8%



Фиг. 34. Kaplan-Meier крива на преживяемостта за 2 групи от женски индивиди от 2 планински популации – Рила и Пирин.



Фиг. 35. Kaplan-Meier крива на риска за група от женски индивиди от 2 планински популации – Рила и Пирин.

Обратна връзка:

От таблицата се вижда процента оцелели индивиди до края на периода за двата района по отделно – съответно Рила – 56,5 % и Пирин – 43,1%, както и общата – 50,8%. От графиките се вижда, че преживяемостта на групата в Рила е по-висока от тази в Пирин, виждат се и периодите на най-голям риск за групите.

При всички тези анализи допълнително могат да се посочват фактори и да се сравнява преживяемостта по действието на различни фактори на средата.

Повечето жизнени таблици се правят за маркирани животни, тъй като това дава възможност за точно проследяване на индивидите. Такива таблици е възможно да се правят и за растения за оценка на риск (от болест, или друг подтискащ фактор). (Виж т.2)

3.3. Литература:

Boitani L., T. K. Fuller. 2000. Research Techniques in Animal Ecology. Columbia University Press. 191-209, 288-327

Burden F., D. Donnert, T. Godish, I. McKalvie. 2004. Environmental monitoring handbook. Water.Part 4 Data analisys. Downloaded from Digital Engineering Library @ McGraw-Hill (www.digitalengineeringlibrary.com) p. 571-629

Chanin P. 2003. Monitoring the Otter *Lutra lutra*. Conserving Natura 2000 Rivers Monitoring Series No. 10, English Nature, Peterborough.

De Laat N. 2010. Monitoring Biodiversity in Asubima Forest Reserve, Ghana. Internship Report. 48pp.

Goverse E., G.F.J. Smit, A. Zuiderwijk, T. van der Meij. 2006.The national amphibian monitoring program in the Netherlands and NATURA 2000. In: M. Vences, J. Köhler, T. Ziegler, W. Böhme (eds): Herpetologia Bonnensis II.

McComb B., B. Zuckerberg, D. Vesely, C. Jordan. Monitoring Animal Populations and Their Habitats. 2010. CRC Press, pp. 190-228

Müller F., C. Baessler, H. Schubert, S. Klotz. 2010. Long-Term Ecological Research. Springer. pp. 91-129

Proceedings of the 13th Congress of the Societas Europaea Herpetologica. pp. 39-42

Костова Р. 2004. Фаунистични и екологични изследвания на съобщества от бръмбари бегачи (Coleoptera: Carabidae) от агроценози около София. Дисертационен труд.

Одум Ю. 1971. Основы экологии. 740 pp.

Изготвил: гл. ас. д-р Румяна Костова