# УМЕНИЯ ЗА ОБРАБОТКА НА ДАННИ С MS EXCEL ЗА НУЖДИТЕ НА ФАРМАКОЕПИДЕМИОЛОГИЯТА

*Ас. Е. Насева*
*Доц. И. Гетов*

## Въведение в статистиката. Видове данни и скали за измерване

Биостатистиката е медицинска наука, която изучава **количествената** страна на **масовите явления** в областта на медицината и здравеопазването.

Случаите (обектите на изследване) в статистическите съвкупности притежават многобройни характеристики, наречени признаци. **Статистически признаци (променливи)** са качества, особености, характеристики, по които се извършва статистическото изучаване. Те не могат да обхванат всички качества и свойства на изучаваните единици.

В зависимост от типа на значенията си, изучаваните признаци могат да бъдат:

- количествени (вариационни, метрирани) – имат числови значения;
- качествени (категорийни, атрибутивни, неметрирани) – нямат числови значения, а категории.

Скалите за измерване се формират на база характера (типа) на значенията на изучаваните признаци:

- номинална скала (класификационна, скала на наименованието) – използва се когато дадено свойство не се поддава на непосредствено измерване (категорийни променливи). В нея не се съдържа никаква информация за величината на измервания признак, а само се различават отделните му класове или категории, като те се отбелязват със символи (семейно положение – омъжена/ женен, разведен/а, вдовец/ вдовица; кръвна група – 0, A, B и AB). Частен случай на номиналната скала е дихотомната – когато значенията на наблюдавания признак са само две, например полът – мъж или жена, и са взаимноизключващи се;
- ординална скала – различават се отделни класове, които обаче могат да се сравняват помежду си (отново категорийни променливи). Отделните категории се подреждат по тяхната тежест (например променливата образование: начално, основно, средно, висше). Ординалната скала има три разновидности – полуподредена (тук разновидностите на признака се изразяват чрез термините – голям-малък, евтино-скъпо, ниско-високо и т.н.); рангова (при нея разновидностите на признака се изразяват с термините първо, второ, трето и т.н.) и бална (например скалата за оценяване знанията – слаб, среден и т.н.);
- интервална скала – променливата се характеризира с наличието на единици на измерване и начална точка. Могат да се оценяват разликите между отделните класове. Различията се представят във вид на интервал между две точки от скалата (температура, надморска височина);
- скала на отношенията (абсолютна; пропорционална скала) – задава се

абсолютна начална точка и абсолютна единица на измерване. Може да се сравнява колко пъти дадено измерване е по-голямо или по-малко от друго (ръст, тегло). При интервалната скала нулата е сложена от човека, докато при скалата на отношенията тя е абсолютна – пример е t° по Целзий и по Келвин.

Категорийните (качествените) данни се представят на слабите скали – номинална, ординарна и рангова, а количествените – на интервалната и на скалата на отношенията.

Независимо от какъв тип са данните, те трябва да се кодират количествено, когато се обработват на компютър. Например двете значения (категории) на променливата пол могат да се кодират така: 1=мъж, 2=жена (или обратното – няма значение, важното е да е еднозначно). Количествените данни нямат нужда от кодиране.

## Организация на данните. Въвеждане на данни.

Един масив с данни винаги трябва да е разположен по определен начин – по редовете са обектите, за които събираме информация (хора, болници и т.н.), а по колони са променливите.

При обработка на данните с Excel има два варианта за разполагането им във файла. Единият е за всяка променлива да има само една колона, вторият е всяка променлива да има толкова колони, колкото възможни отговори има тя (за категорийните променливи) или само по една колона (за количествените променливи). Първият вариант е за предпочитане, защото дава възможност данните да бъдат пренесени в друга програма – например SPSS, както и да се извършва истинска статистическа обработка.

## Предварителна обработка на данните

За **неотговорилите** има **два подхода**: единият е да се кодират. Обикновено ги кодираме с 9 или 99 или 999 (обикновено това се задава в самия въпросник – *само при въпросник*!, а ако интервюиращият пропусне да отбележи „без отговор”, то наемаме допълнителен човек, който преглежда всеки въпросник и коригира такива грешки преди въпросниците да се дадат за въвеждане).

Вторият подход е да се оставят като празноти.

След въвеждането на данните от едно проучване, тези данни трябва да бъдат прегледани основно и „**изчистени**” от грешки при събирането на данните и от грешки при въвеждането им в компютъра.

## Преобразуване на данните

След като данните са изчистени, понякога се налага да се извършат някои преобразувания с тях, за да са ни по-удобни за работа.

**СОРТИРАНЕ** на данните. Маркира се целия масив, включително и имената на променливите и от менюто се избира Data-Sort (на версия 2007 Home-Sort and Filter)

Образуване на **НОВА** променлива, чиито стойности за всеки отделен случай са резултат от аритметични и логически операции с други променливи, т.е. пресмятането се отнася за цялата колона, а не само за една клетка.

**РЕКОДИРАНЕ**. Рекодирането е процедура, при която въведените стойности на дадена променлива се преобразуват според избрана от изследователя схема.

## Таблично представяне на данните – честотни разпределения

Честотните разпределения са най-простият и най-често използваният начин за обобщено представяне на данните в даден масив. Чрез честотните разпределения виждаме абсолютната честота, с която са се срещали категориите на променливите. По-простичкото определение е такова подреждане на данните, обикновено в табличен вид, което показва колко пъти дадена стойност или група от стойности са се срещали, наблюдавали, повтаряли в извадката.

Най-често използваните честотни разпределения са т.нар. **едномерни** и **двумерни разпределения.**

В Excel може да се построи чрез т.нар. Pivot table. Data – Pivot Table and Pivot Chart Report, провлачваме променливата до Drop Row fields here, а променливата а (условната променлива) до Drop Data items here. Можем да избираме да се визуализират само мъжете или само жените, например. (на версия 2007 от Insert-Pivot table)

**Друг начин**, който е валиден само за числови данни: Insert-Function-Frequency

**Двумерни разпределения**: Pivot table.

## Графично представяне на данните

Графичните изображения са необходимо средство за провеждането на анализ. Те са един от най-добрите методи за онагледяване на тенденции, връзки, структури, разпределения и др.

Видове графични изображения:

- диаграми – данните се представят чрез линии (линейни диаграми), чрез плоскостни фигури (плоскостни диаграми) и чрез пространствени геометрични тела (стереограми);
- картограми – стилизирани карти, изразяващи териториално разпределение на изучаваните явления. Строят се върху географска карта на съответна територия;
- картодиаграми – съчетават елементи на картограмата с елементи на плоскостни диаграми.

Хистограма и полигон с помощта на Analysis toolpack

Tools-Data Analysis-Histogram (на 2007 – Data-Data Analysis

## Използване на формули за изчисляване на обобщаващи характеристики на количествени променливи. Вариационен анализ

Вариационният анализ е анализ на вариацията на дадена променлива. Дескриптивна (описателна) статистика – другото име на вариационния анализ. Той се прилага само за **количествени** променливи!

Средната аритметична **Function-Insert function – Statistical – Average**

**Медиана**. **Function-Insert function – Statistical – Median**

**Мода**. **Function-Insert function – Statistical – Mode**

Ако редът има повече от една мода, програмата ни показва само тази с най-малка стойност.

**Показатели за разсейване.**

**Размах** = MAX(област с данни) – MIN(област с данни) .

**Стандартно отклонение. Дисперсия.**

**Function-Insert function – Statistical – Var**

**Function-Insert function – Statistical – Stdev**

**Асиметрия**
**Function-Insert function – Statistical – Skew**

**Ексцес.**
**Function-Insert function – Statistical – Kurt**

## Точкови и интервални оценки. Доверителен интервал.

**Стандартна и максимална грешка на средната.**
**Function-Insert function – Statistical – Confidence** – максималната грешка на средната

Всичко това може да се изчисли и с Tool-Data Analysis- Descriptive Statistics

## Проверка на хипотези.

### Сравняване на средни стойности. Зависими и независими извадки

Видове проверка на хипотези:

- за две средни от две независими извадки (с достатъчно голям обем) – z-Test: two samples for means
- за две средни от две независими извадки с равни дисперсии – t-Test: two sample assuming equal variance
- за две средни от две независими извадки с различни дисперсии t-Test: two sample assuming unequal variance
- за две средни от свързани извадки – когато имаме данни в началото и в края на проучването t-Test: paired two sample for means

*Кога извадките са независими и кога са свързани?*

*Как да определим дали са еднакви или различни дисперсиите?*
Tools – Data analysis – F test – Two Sample for Variances
H0: няма разлика между двете дисперсии
H1: има разлика между двете дисперсии

### Корелационен и регресионен анализ. Същност. Коефициенти на корелация.

Tool-Data Analyses-Regression

### Зависимости между количествена и качествена променливи. Дисперсионен анализ.

Tools-Data analyses-ANOVA