Глава 11

Понятия, формули и дефиниции в описателната статистика

Ще дадем дефиниции на основните понятия в т. нар. описателна статистика – дял на математическата статистиката, който дава първоначална представа за числовите характеристики на данните и тяхното разпределение. Най-общо, задачата на математическата статистика може да се формулира като оценяване неизвестните параметри на разпределението на наблюдавания признак по данните от извадката, която правим, основано на теория на вероятностите. Ако имаме основания да направим някакви първоначални предположения за разпределението, както ще видим по-нататък, то задачата се свежда до оценяване на параметрите на това разпределение. Този дял от математическата статистика, който се занимава с оценката на параметрите, се нарича параметрична статистика. Методите на параметричната статистика са: точково оценяване, доверително оценяване или построяване на доверителни интервали и тестване на хипотези за неизвестните параметри или проверка на хипотези. В тази глава ще се знимаваме със задачите на параметричната статистика.

Определение 11.1 *Случайна извадка с обем n наричаме наблюденията, които правим, означени с $X_1, X_2, \ldots, X_n$.*

Определение 11.2 *Извадкова статистика наричаме произволна функция на наблюденията.*

Определение 11.3 *Извадъчно средно наричаме* $\bar{X} = \sum_{i=1}^{n} \frac{X_i}{n}$.

Определение 11.4 *Извадъчна дисперсия наричаме*

$$s^2 = \sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{n-1} = \frac{\sum_{i=1}^{n}[X_i^2] - n\bar{X}^2}{n-1}.$$

Определение 11.5 *Извадъчно стандартно отклонение наричаме*

$$s = \sqrt{s^2} = \sqrt{\sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{n-1}}.$$

Определение 11.6 *Вариационен ред наричаме стойностите от извадката, подредени по големина:*

$$X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)}.$$

Определение 11.7 *Размах (обхват) наричаме разликата между най-голямата и най-малката стойност в извадката:* $X_{(n)} - X_{(1)}$.

Определение 11.8 *Медиана наричаме стойност, за която 50% от данните са по-малки от нея.*

Определение 11.9 *p-ти процентил наричаме стойност, за която p% от данните са по-малки от нея.*

Определение 11.10 *Долен и горен квартил* (Q_L, Q_U) *наричаме стойности, за които* 25% *и* 75% *съответно от данните са по-малки от нея.*

Определение 11.11 *Интерквартилен размах наричаме разликата между горния и долния квартил:* $Q_U - Q_L$.

Определение 11.12 *Силно отличаващо се наблюдение (outlier) наричаме стойност, която силно се отличава от извадката - необикновено голямо или необикновено малко наблюдение в сравнение с останалите.*

Определение 11.13 *z-score на наблюдение* X_i *наричаме z-score*$(X_i) = \dfrac{X_i - \bar{X}}{s}$.