Упражнение 10. Дескриптивни статистики. Точкови оценки.

Генерална съвкупност и извадка.

Генералната съвкупност е множеството от всички обекти на измерване. Например една генерална съвкупност е всички мъже в България. Когато искаме да имерим някоя количествена характеристика на всеки член на генералната съвкупност, то на тази генерална съвкупност може да се гледа като на като всички възможни изходи от даден опит. Например ако измерваме ръст на мъжете в България, тогава генералната съвкупност може да се отъждестви с всички височини на всички мъже в България. Естествено, че това са всички възможни реализации на сл.в. *'ръст на българските мъже'*. Ясно е, че ако наистина вземем един по един всички елементи на генералната съвкупност и измерим дадена характеристика, то тогава ние ще можем да намерим точно закона за разпределение на тази характеристика тъй като ще знаем всички нейни възможни стойности и техните честоти, но такъв подход не е винаги възможен, а и не винаги е оправдан. Поради това обикновено се извършват наблюдения над *случайно и независимо избрани елементи* на *генералната съвкупност*, които съставляват така наречената *извадка.*

Извадката се състои от наблюдения (измервания) на характеристики на *случайно и независимо един от друг избрани елементи от генералната съвкупност.* С други думи повторен е опитът краен брой пъти и са записани резултатите от тези повторения. Характеристиките могат да бъдат *количествени* (ръст, тегло, заплата и др.), *качествени* (мъж-жена-дете, здрав-болен и т.н.)
Тук ще се спрем на количествени характеристики.
Основната задача, която се решава в статистиката може да бъде описана, като задача обратна на тази, която се решава в теория на вероятностите. Там се счита, че законът за разпределение на една сл. величина е известен и се решава задача за определяне на вероятността да се сбъдне определено събитие, свързано с тази сл. величина. Сега по направената извадка можем да видим дали дадено събитие се е сбъднало, да определим закона за разпределение на сл. величина, да намерим оценки за параметрите на този закон, да намерим оценки за математическото очакване, дисперсията, модата, медианата, квантилите на закона за разпределение, въз основа на направените наблюдения в извадката.
Такива оценки сигурно могат да се построят по различен начин например, като средно можем да вземем средно аритметично, средно квадратично, средно претеглено с дадени тегла и т.н. Естествено, възникават въпроси свързани с това коя оценка е за предпочитане. Създадени са критерии за сравнение на различни оценки.

А. Дескриптивни статистики.

В раздела **Дескриптивна (описателна) статистика** по данните от извадката се пресмятат основни количествени характеристики на тази извадка.

Нека извадката се състои от независими наблюдения $x_1, x_2, \dots x_N$ над една сл. величина X.

Числото N се нарича обем на извадката.
Определяме:

Средно на извадката: $\bar{X} = \frac{1}{N}\sum_{i=1}^{N} x_i$

Дисперсия на извадката: $\sigma_X^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{X})$ или $s_X^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{X})$.

Вариационен ред наричаме подредената извадка $x_{(1)} \le x_{(2)} \le \dots \le x_{(N)}$.

Медиана на извадката се дефинира чрез

$$Me(X) = \begin{cases} x_{(\frac{N+1}{2})}, & N \text{ нечетно}, \\ \frac{x_{(N/2)} + x_{(N/2+1)}}{2}, & N \text{ четно}. \end{cases}$$

Медианата разделя числата в извадката по равен брой в ляво и в дясно от нея.

Долен квартил Q_1 е медианата на числата в извадката, които са по-малки от медианата $Me(X)$. Вляво от Q_1 са 25% от наблюденията.

Горен квартил Q_3 е медиана на числата, които са по-големи от медианата $Me(X)$. Вляво от Q_3 са 75% от наблюденията.

p% квантил е числото, вляво от което са p% от наблюденията.

Най-малко, най-голямо и размах: $x_l = \min x_i, x_r = \max x_i, R = x_r - x_l$.

Мода на извадката $Mod(X)$ е онова x_i, което се повтаря най-много пъти в извадката.
Друг начин, който е пригоден и за непрекъснати сл. величини, където всяка стойност би се случила по един път, е да се раздели интервалът $[x_l, x_r]$ на равни подинтервали и да се преброят числата от извадката във всеки интервал. Да се построи хистограмата на честотите и за мода да се вземе средата на интервала, в който хистограмата е най-висока.

Групирани данни. Понякога, при извършване на опита не получаваме точно едно число при измерването над конкретен член на генералната съвкупност, а резултатът е някакъв интервал (a, b) в който лежи точната стойност на измерваната характеристика. Тогава казваме, че имаме групирани данни. В този случай разгледаните по-горе характеристики на извадката леко се модифицират.

Формула за квартилите:

$$Q_j = L_l + \frac{N \times j/4 - F_l}{f_l} \times I, \quad j = 1, 2, 3$$

Тук L_l е левия край на интервала, в който е квартила. Интервалът в който е квартила се определя, като интервала преди който сумата от честотите не надминава $N \times j/4$, а като се добави и честотата в него сумата става по-голяма от $N \times j/4$.
F_l е сумата от честотите в интервалите преди L_l
f_l е честотата в интервала на медианта.
I е ширината на интервала, съдържащ медианата.
Същата формула работи и за други квантили.
Формула за средното $\bar{X} = \frac{1}{N} \sum_{j=1}^{M} f_j \times (L_j + R_j)/2$ като $\sum_{j=1}^{M} f_j = N$.

Графично представяне на данни от извадка.

Представянето на данни всъщност е основна задача както на изчерпателната така и на извадъчната статистика. Информацията, която се съдържа в милионите числа трябва да бъде представена в обозрима форма, така че всеки да си представи основните качества на множеството обекти. Главна роля в това кондензиране на информация има графичното представяне. То е ефектно и в минимална степен при него се губи информация.

Boxplot диаграма

Това е графичен начин за представяне на данни от извадка, чрез 5 числа: най-малкото наблюдение X_l, долния квартил Q_1, медианата $Me(X) = Q_2$, горния квартил Q_3 и най-голямото наблюдение X_r. Тези диаграми са полезни, когато искаме да сравним разлики между две генерални съвкупности без да предполагаме някакви разпределения.

Хистограма на наблюденията Хистограмата е основният вид за представяне на информацията за наблюдения върху числов признак. Тя се строи по просто правило. Избират се обикновено не много на брой (5 - 20) еднакво големи прилежащи интервала покриващи множеството от стойности на наблюдавания признак. Те се нанасят върху оста x. След това всеки от обектите на извадката се премерва и получената стойност попада в някой от интервалите.

Задачи

Зад 1. *От 14 наблюдения за брой минути чакане на даден автобус в един и същи сутрешен час на една и съща спирка, са получени следните данни: 10, 2, 17, 6, 8, 3, 10, 2, 9, 5, 9, 13, 1, 10. Да се определи средното време за чакане на автобус(чрез средно и медиана). Същото да се пресметне, ако извадката се допълни с числото 48. Да се пресметнат първият и третият квартил, модата. Да се построи boxplot диаграмата.*

Решение:

Извадката е

$X = \{10, 2, 17, 6, 8, 3, 10, 2, 9, 5, 9, 13, 1, 10\}$

Подредени данни (вариационен ред)

$X = \{1, 2, 2, 3, 5, 6, 8, 9, 9, 10, 10, 10, 13, 17\}$

$\bar{X} = 7.5$

$Me(X) = 8.5 = (8 + 9)/2.$

$Mod(X) = 10$

$Q_1 = 3$

$Q_3 = 10$

Променена извадка (48 може да се счита за грешно наблюдение).

$X = \{10, 2, 17, 6, 8, 3, 10, 2, 9, 5, 9, 13, 1, 10, 48\}$

Подредени данни (вариационен ред)

$X = \{1, 2, 2, 3, 5, 6, 8, 9, 9, 10, 10, 10, 13, 17, 48\}$

$\bar{X} = 10.2$

$Me(X) = X_8 = 9$

Вижда се че медианата се влияе по-слабо от евентуални грешни наблюдения, които много се отличават от другите.

$Mod(X) = 10$

$Q_1 = 4$

$Q_3 = 10$

Зад 2. *При анкетиране на 100 случайно избрани разведени жени относно възрастта, на която са се развели са получени следните резултати.*

възраст	*честота*
15 - 29	*11*
30-39	*26*
40-44	*21*
45-49	*18*
50-54	*11*
55-64	*13*

Да се пресметне извадъчното средно, медианата, 70-тия процентил, модалния интервал.

Решение:

L	R	f	$(L+R)/2$	$f \times (L+R)/2$
15	29	11	22	242
30	39	26	34.5	897
40	44	21	42	882
45	49	18	47	846
50	54	11	52	572
55	64	13	59.5	773.5
		$N = 100$		$\bar{X} = 42.125$

Намираме $F_l = 11 + 26 = 37$, $L_l = 40$, $f_l = 21$, $I = 4$. Тогава

$$Me(X) = Q_2 = 40 + \frac{100/2 - 37}{21} \times 4 = 42.47619048.$$

За P_{70} имаме по същия начин: Намираме $F_l = 11 + 26 + 21 = 58$, $L_l = 45$, $f_l = 18$, $I = 4$. Тогава

$$P_{70} = 45 + \frac{0.7 \times 100 - 58}{18} \times 4 = 49.44444444.$$

Модалният интервал е $[30, 39]$, защото съдържа най-много наблюдения.

Б. Точкови оценки.

Когато имаме предположение за закона за разпределение въз основа на извадката, ние искаме да намерим оценки за неговите параметри. Така, например при нормалния закон параметрите са 2: средното a и дисперсията σ^2; при експоненциалното разпределение - параметърът е един и т.н.

Оценките, които се получават от данните в извадка са всъщност функции на наблюденията. Такива функции се наричат *статистики.*

За построяването на такива оценки се използват метода на моментите и метода на максималното правдоподобие. В следващите задачи ще използуваме метода на максималното правдоподобие за намиране на оценки за параметрите на някои разпределения. Такива оценки се наричат максимално правдоподобни оценки.

Зад 3. *Нека $x_1, ..., x_N$ са независими наблюдения над сл.в. $\xi \sim N(a, 1)$. Да се намери м.п.о. за неизвестния параметър a.*

Решение:
Нека $f(x, a, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{(x-a)^2}{2\sigma^2})$ е плътността на нормалния закон $N(a, \sigma^2)$.
Да разгледаме съвместната плътност на N на брой независими еднакво разпределедни сл. величини $X_i \sim N(a, \sigma^2)$. Тя е

$$L(x_1, x_2, \ldots, x_N, a, \sigma^2) = f(x_1, a, \sigma^2) \times f(x_2, a, \sigma^2) \times \ldots \times f(x_N, a, \sigma^2)$$

$$= \frac{1}{\sqrt{2\pi}^N \sigma^N} \exp\left(- \sum_{i=1}^{N} \frac{(x_i - a)^2}{2\sigma^2} \right)$$

и в статистиката се нарича *функция на правдоподобие.*

Считаме, че числата x_i, които за нас са всъщност данните в извадката са известни, считаме също, че дисперсията е известна и е равна на 1. Така функцията на правдоподобие е

$$L(x_1, x_2, \ldots, x_N, a, 1) = \frac{1}{\sqrt{2\pi}^N} \exp\left(- \sum_{i=1}^{N} \frac{(x_i - a)^2}{2} \right).$$

Търсим най-голямата стойност на тази функция като функция само на a. Ясно е, че тя ще се получи, когато се получава и най-голямата стойност на логаритъма на функцията на правдоподобие

$$l(x_1, \ldots, x_N, a, 1) = \log L(x_1, \ldots, x_N, a, 1) = \frac{N}{4} \log 2 \times \sum_{i=1}^{N} (x_i - a)^2.$$

Или, което е същото, търсим НГС на $g(a) = \sum_{i=1}^{N} (x_i - a)^2$.
Намираме производната $g'(a) = 2 \sum_{i=1}^{N} (x_i - a) = 0$. Следователно най-голямата стойност на $g(a)$ се достига при

$$\hat{a} = \frac{1}{N} \sum_{i=1}^{N} x_i.$$

Така получената функция на наблюденията (статистиката) се нарича *максимално правдоподобна оценка* за математическото очакване на нормалния закон при известна дисперсия. (Може да се види, че $\sigma^2 = 1$ не е съществено за сметките.)

Зад 4. *Нека $x_1, ..., x_N$ са независими наблюдения над сл.в. $\xi \sim N(a, \sigma^2)$. Да се намери м.п.о. за неизвестния параметър σ^2.*

Решение:
Да решим сега задачата за намиране на максимално правдоподобна оценка за дисперсията, ако считаме, че средното е известно. Отново разглеждме логаритъма на функцията на правдоподобие

$$l(x_1, \ldots, x_N, a, \sigma^2) = \log L(x_1, \ldots, x_N, a, \sigma^2)$$

$$= -\frac{N}{2}(\log 2 + \log \sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{N} (x_i - a)^2.$$

Като диференцираме спрямо σ^2 намираме

$$\frac{dl}{d\sigma^2} = -\frac{N}{2} \frac{1}{\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^{N} (x_i - a)^2 = 0.$$

От тук намираме

$$\hat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - a)^2,$$

което е максимално правдоподобната оценка за дисперсията при известно математическо очакване a.

Зад 5. *Нека $x_1, ..., x_N$ са независими наблюдения над сл.в. $\xi \sim N(a, \sigma^2)$. Да се намери м.п.о. за неизвестния параметър $\theta = (a, \sigma^2)$.*

Решение:
Нека сега векторният параметър $\theta = (a, \sigma^2)$ е неизвестен. Разглеждаме логаритъма на функцията на правдоподобие

$$l(x_1, \ldots, x_N, a, \sigma^2) = -\frac{N}{2}(\log 2 + \log \sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(x_i - a)^2$$

като функция на две променливи a и σ^2. Намираме системата:

$$\frac{\partial l}{\partial a} = \frac{1}{\sigma^2}\sum_{i=1}^{N}(x_i - a) = 0$$

$$\frac{\partial l}{\partial \sigma^2} = -\frac{N}{2}\frac{1}{\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{N}(x_i - a)^2 = 0$$

Първото уравнение дава

$$\hat{a} = \frac{1}{N}\sum_{i=1}^{N} x_i.$$

Като заместим a с намереното във второто, получаваме

$$\hat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \hat{a})^2.$$

Така за векторния параметър θ имаме

$$\hat{\theta} = \left(\frac{1}{N}\sum_{i=1}^{N} x_i, \frac{1}{N}\sum_{i=1}^{N}(x_i - \hat{a})^2\right).$$

Зад 6. *За да се оцени броят на рибите в едно езеро се постъпва по следния начин: улавят се M на брой риби маркират се и се връщат обратно в езерото. След известен период от време се улавят n риби оказва се, че m от тях са маркирани. Да се намери м.п.о. за общия брой на рибите в езерото N.*

Решение:
Имаме едно наблюдение на биномно разпределена сл. величина с параметри n и $p = \frac{M}{N}$ като наблюдаваната стойност е m. Функцията на правдоподобие ще е $L(m, n, p) = \binom{n}{m} p^m (1-p)^{n-m}$, а логаритъмът и е

$$l(m, n, p) = \log \binom{n}{m} + m \log p + (n-m) \log(1-p).$$

Тогава

$$\frac{dl}{dp} = \frac{m}{p} - \frac{n-m}{1-p} = 0$$

или

$$(1-p)m - p(n-m) = 0 \quad m - mp - np + mp = 0.$$

Следователно $p = \frac{m}{n}$, т.е. $\frac{M}{N} = \frac{m}{n}$ или $\hat{N} = M\frac{n}{m}$.

Зад 7. *Нека $x_1, ..., x_N$ са независими наблюдения над сл.в. $\xi \sim Bi(1, p)$. Да се намери м.п.о. за неизвестния параметър p.*

Решение:
Да означим биномните вероятности с $b(n, k, p) = \binom{n}{k} p^k (1-p)^{n-k}$. Имаме $n = 1$ (следователно $k = 0, 1$) и N независими наблюдения. (С други думи имаме Бернулиеви сл. в.). Съвместното разпределение (функцията на правдоподобие) е (като сме използували, че $\binom{1}{1} = \binom{1}{0} = 1$)

$$L(k_1, k_2, \ldots, k_N, p, 1) = \prod_{i=1}^{N} b(1, k_i, p) = p^{k_1+k_2+\ldots+k_n} (1-p)^{N-k_1-k_2\ldots-k_N}.$$

Логаритъмът на функцията на правдоподобие е

$$l(k_1, k_2, \ldots, k_N, p, 1) = (k_1+k_2+\ldots+k_n) \log p + (N - ((k_1+k_2+\ldots+k_n)) \log(1-p).$$

Сега

$$\frac{dl}{dp} = \frac{1}{p}(k_1 + k_2 + \ldots + k_n) - \frac{1}{1-p}(N - ((k_1 + k_2 + \ldots + k_n) = 0.$$

Така

$$(1-p)(k_1 + k_2 + \ldots + k_n) + p(k_1 + k_2 + \ldots + k_n) - pN = 0,$$

$$(k_1 + k_2 + \ldots + k_n) = Np,$$

$$\hat{p} = \frac{(k_1 + k_2 + \ldots + k_n)}{N}.$$

В числителя се получава сума равна на броя на успехите.

Зад 8. *Нека $x_1, ..., x_N$ са независими наблюдения над сл.в. $\xi \sim U[0, b]$. Да се намери м.п.о. за неизвестния параметър b.*

Решение:

Имаме N независими наблюдения над сл. величина с равномерно разпределение $U[0,b]$. Плътността на X_i е съответно $f(x_i) = \frac{1}{b} \;\; x_i \in [0,b]$, т.е. $f(x_i) = \frac{1}{b} I_{\{0 \leq x_i \leq b\}}$. Следователно функцията на правдоподобие е

$$L(x_1, x_2, \ldots, x_N | b) = \frac{1}{b^N} \prod_{i=1}^{N} I_{\{0 \leq x_i \leq b\}} = \frac{1}{b^N} I_{\{\max\{x_1, \ldots, x_N\} \leq b\}}.$$

Сега имаме, че ако $b < \max\{x_1, x_2, \ldots, x_N\}$ то $L(x_1, x_2, \ldots, x_N | b) = 0$, докато при $b \geq \max\{x_1, x_2, \ldots, x_N\}$, т.е. имаме $L(x_1, x_2, \ldots, x_N | b) = \frac{1}{b^N}$. Очевидно най голямата стойност на тази функция в интервала $b \in [\max\{x_1, x_2, \ldots, x_N\}, \infty)$ се достига в левия край, при $b = \max\{x_1, x_2, \ldots, x_N\}$, което е и максимално правдоподобната оценка за b.