Основные понятия математической статистики: оценки параметров распределения, проверка гипотез, системы случайных величин: корреляция, регрессия

Лекция 17

Способы организации выборки

1. **Вариационный ряд** – элементы выборки упорядочивают по величине: $x^{(1)} < x^{(2)} < \cdots < x^{(n)}$ → $\boldsymbol{x^{(1)} = min\,\{x_n\}}$
$\boldsymbol{x^{(n)} = max\{x_n\}}$

2. **Размах выборки -** разность между максимальным и минимальным элементами выборки $\boldsymbol{w = x^{(n)} - x^{(1)}}$

3. **Пусть выборка содержит $\boldsymbol{k}$ различных элементов.**

Частота элемента выборки $\boldsymbol{n_i}$ - число раз, которые данный элемент встречается в выборке

4. **Мода** – элемент выборки **с наибольшей частотой**

5. **Статистический ряд – таблица:** $\{\boldsymbol{x_i} \rightarrow \boldsymbol{n_i}\}$

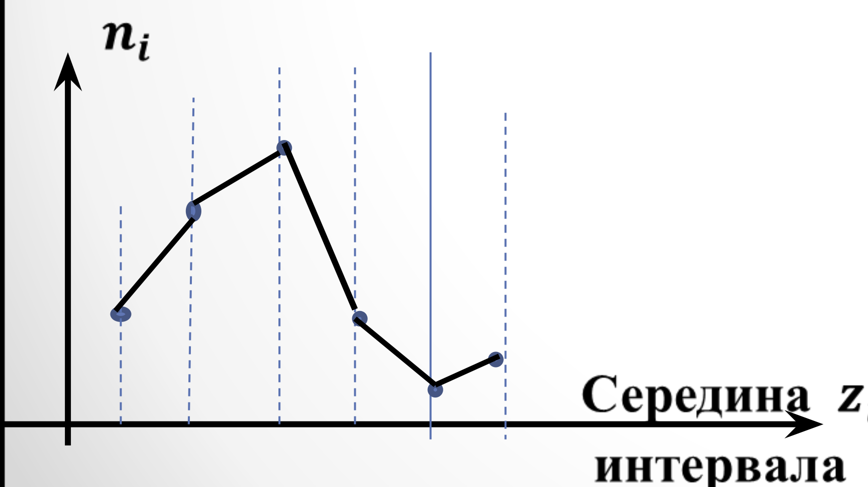
			....	
			...	

$\boldsymbol{\sum_{i=1}^{k} n_i = n}$

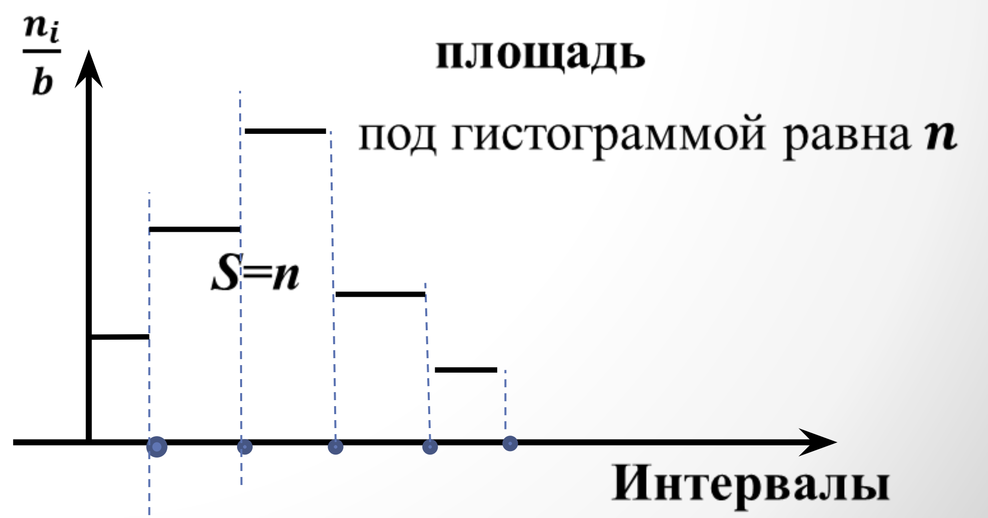
сумма частот всех элементов равна объему выборки

Способы описания выборки

При **большом** объеме выборки ее элементы объединяют в группы (разряды, карманы): **выбирают ширину** интервала $\boldsymbol{b = \frac{w}{k}}$, где $k \approx \sqrt{n}$, или $\boldsymbol{b = \frac{w}{1+3{,}2lgn}}$, **а частота** $\boldsymbol{n_i}$ **- количество элементов выборки, попавшее в** $\boldsymbol{i}$ **–й интервал** (элемент, совпадающий с внешней границей интервала считают в последующем). Кроме того вычисляю **середину каждого интервала** $\boldsymbol{z_i = \frac{x_i + x_{i+1}}{2}}$ **и относительную частоту** $\boldsymbol{\frac{n_i}{n}}$ **-** ***оценку вероятности*** **попадания значения случайной величины в данный интервал :** $\boldsymbol{\sum_{i=1}^{k} \frac{n_i}{n} = 1}$ **(Образец табл. Стр. 181).** **Графическое представление** – *полигон частот* (или относительных частот) и *гистограмма* – статистические аналоги функции распределения $f(x)$



Полигон частот



Гистограмма

Числовые характеристики выборки

Выборочное среднее $\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$ **или** $\overline{x} = \frac{\sum_{i=1}^{n} n_i x_i}{n}$

Выборочная дисперсия $S^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n}$ **или** $S^2 = \frac{\sum_{i=1}^{n} n_i (x_i - \overline{x})^2}{n}$.

Для выборок **малого объема** $(n < 30)$ вводят **исправленную дисперсию**

$\widetilde{S}^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}$ **Пример.** Для выборки из 5 чисел **3, 5, 5, 8, 4** получаем

$$\overline{x} = \frac{3+5+5+8+4}{5} = 5; \quad \widetilde{S}^2 = \frac{(3-5)^2+(5-5)^2+(5-5)^2+(8-5)^2+(4-5)^2}{4} = 3,5$$

- **Excel надстройки «Пакет анализа» Описательная статистика**
- Среднее 5
- Стандартная ошибка 0,836660027 $\varepsilon 1 = \frac{s}{\sqrt{n}}$
- Медиана 5
- Мода 5
- Стандартное отклонение 1,870828693 S
- Дисперсия выборки 3,5 S^2
- Эксцесс 2
- Асимметричность 1,145405322
- Интервал 5
- Минимум 3
- Максимум 8
- Сумма 25
- Счет 5
- Уровень надежности(95,0%) 2,322940635 - $\varepsilon = \frac{s}{\sqrt{n}} t_{0,05}(n-1)$

Статистическое оценивание. Точечные оценки

Точечной оценкой $\tilde{\theta}$ неизвестного параметра Θ называют приближенное значение этого параметра, полученное по выборке $\tilde{\theta} = \tilde{\theta}\,(x_1,\ x_2, \dots x_n)$ *или «статистика»*.

Качество оценок.

1. Состоятельность. Оценка параметра сходится по вероятности к самому параметру при $n \to \infty$ $\lim_{n\to\infty} P\left(\left|\widetilde{\theta_n} - \theta\right| < \varepsilon\right) = 1$.

Или чем больше объем выборки, тем точнее оценка

Пример. $\boldsymbol{M[X] = \bar{x}}$ выборочное среднее – состоятельная оценка математического ожидания (теорема Чебышева)

2. Несмещенность. Математическое ожидание оценки параметра равно самому параметру : $M\left[\tilde{\theta}\right] = \boldsymbol{\theta}$. **Пример 1.** Выборочное среднее является несмещенной оценкой математического ожидания. **Пример 2. Выборочная дисперсия** $\boldsymbol{S^2 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n}}$ **является смещенной оценкой (** $\boldsymbol{M\left[S^2\right] = \frac{n}{n-1}S^2}$**) , а оценка** $\boldsymbol{\tilde{S}^2 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}}$ **является несмещенной**

3. Эффективность. Оценка должна обладать наименьшей дисперсией

Интервальные оценки. Уровень значимости

Интервальные оценки или. **доверительные интервалы вводятся с целью определения *точности* оценки.**

Доверительным интервалом для параметра $\boldsymbol{\theta}$ называют интервал (θ_1, θ_2), содержащий истинное значение параметра

с заданной вероятностью $\boldsymbol{P = 1 - \alpha}$:

$$\boldsymbol{P(\theta_1 < \theta < \theta_2) = 1 - \alpha.}$$

$(1 - \alpha)$ – *доверительная вероятность;*

α – *число - вероятность , которую называют* ***уровнем значимости,*** *характеризует точность оценивания. Обычно выбирают* $\alpha = 0{,}1; \quad 0{,}05$

Пример. Доверительный интервал для математического ожидания при неизвестной заранее дисперсии:

$$\boldsymbol{P(\overline{x} - \varepsilon < M[X] < \overline{x} + \varepsilon) = 1 - \alpha}, \quad \textbf{где} \quad \boldsymbol{\varepsilon = \frac{S}{\sqrt{n}} t_{1-\frac{\alpha}{2}}(n-1)}$$

- **Число $\boldsymbol{t_{1-\frac{\alpha}{2}}(n-1)}$**- квантиль распределения Стьюдента находим по статистическим таблицам или в Excel (функции→ статистические → Стьюдентраспобр)

Чем больше уровень значимости, тем выше точность оценивания

Проверка статистических гипотез

Статистическая гипотеза H – это предположение относительно параметров или вида распределения (проверяемая гипотеза называется нулевой):

Пример 1. $\boldsymbol{H_0}$: $M[X] = m \Longrightarrow \boldsymbol{H_1}$: $M[X] \neq m$ (альтернативная гипотеза)

Пример 2. $\boldsymbol{H_0}$: случайная величина распределена по нормальному закону

$\boldsymbol{H_1}$: случайная величина не распределена по нормальному закону

Критерий - правило, согласно которому принимается решение принять или отвергнуть нулевую гипотезу.

Перед проверкой задается **малая вероятность $\boldsymbol{\alpha}$ – уровень значимости, которая определяет размер критической области $\boldsymbol{V_k}$ статистики критерия $\boldsymbol{Z}$.**

Если выборочное значение статистики критерия попадает в критическую область $\boldsymbol{Z_{\text{выб}} \in V_k}$, гипотеза $\boldsymbol{H_0}$ отклоняется , то есть $\boldsymbol{\alpha}$ – вероятность совершить ошибку , отвергнув правильную гипотезу.

О достоверности выводов, полученных при заданном уровне значимости:

$\alpha \geq 0{,}1$ **высокий** уровень значимости $\longrightarrow$ данные согласуются с $\boldsymbol{H_0}$

$\boldsymbol{\alpha = 0,05}$ значимость H_0 возможна, но есть сомнения в истинности

$\alpha = 0{,}02$ имеют место сильные доводы против $\boldsymbol{H_0}$

$\alpha \leq 0{,}01$ основная гипотеза $\boldsymbol{H_0}$ наверняка ложная

Выборочный коэффициент корреляции. Оценка

Для системы случайных величин $\{\boldsymbol{X},\ \boldsymbol{Y}\}$ вводится характеристика – **ковариация** (**корреляционный момент**):

$$cov[X,Y] = K_{XY} = M[(X - M[X])(Y - M[Y])] = M[XY] - M[X]M[Y]$$

Для независимых $X,\ Y$ ковариация $\boldsymbol{cov[X,Y] = 0.}$

Коэффициент корреляции $\boldsymbol{r_{XY} = \frac{K_{XY}}{\sigma_X \sigma_Y}}$ **-** безразмерный коэффициент, который определяет степень линейной корреляционной зависимости между случайными величинами.

Свойства $\boldsymbol{r_{XY}}$ **:** **1)** если $\boldsymbol{Y = AX + B}$ **, то** $\boldsymbol{|r_{XY}| = 1}$

2) $\boldsymbol{|r_{XY}| \leq\ 1}$ **3)** если $\boldsymbol{r_{XY} = 0,}$ **то** случайные величины $\boldsymbol{X}$ **,** $\boldsymbol{Y}$ называют **некоррелированными.** Независимые случайные величины являются некоррелированными.

Оценкой коэффициента корреляции является выборочный коэффициент корреляции $\widetilde{r_{XY}} = \dfrac{\sum_{k=1}^{n}(x_k-\bar{x})(y_k-\bar{y})}{\sqrt{\sum_{k=1}^{n}(x_k-\bar{x})^2 \sum_{k=1}^{n}(y_k-\bar{y})^2}}$

Excel функции статистические корреляция

Регрессионные модели

Пусть коэффициент корреляции между двумя случайными величинами **значимо отличается от нуля и близок к единице.**

Выдвигаем гипотезу: случайные величины связаны линейной корреляцинной зависимостью $\boldsymbol{Y = AX + B \ + \ \varepsilon}$

Это уравнение называют **уравнением линейной регрессии.**

Регрессия – оптимальная зависимость, которая обеспечивает аппроксимацию опытных данных с наибольшей точностью, то есть с минимальной случайной ошибкой $\boldsymbol{\varepsilon}$. Наилучшие оценки для коэффициентов регрессии $\boldsymbol{A}$, $\boldsymbol{B}$ получают по **методу наименьших квадратов :**

$$\boldsymbol{S(\widetilde{A}, \widetilde{B}) = \sum_{k=1}^{n}\left(y_k - \left(\widetilde{A}x_k + \widetilde{B}\right)\right)^2 \text{ - } min}$$

Excel $\longrightarrow$ **точечная диаграмма** $\longrightarrow$ **линия тренда с указанием уравнения и качества аппроксимации - коэффициент детерминации** $\boldsymbol{R^2 = \frac{\sum_k^n\left((\widetilde{A}\,x_k + \widetilde{B}) - \bar{y}\right)^2}{\sum_{k=1}^{n}(y_k - \bar{y})^2}}$

(*правой кнопкой на точку).* На заключительной стадии обязательно проверяют **статистическую значимость** (можно доказать , что в доверительный интервал для коэффициента $\boldsymbol{A}$ не содержит $A = 0$) и **адекватность модели** (случайные ошибки наблюдений – **остатки распределены с нулевым средним** $\boldsymbol{\bar{\varepsilon} = 0}$).

Excel «Анализ данных» Регрессия