Тема 1 Случайные величины. Дискретные и непрерывные СВ. Генеральная совокупность и выборка. Выборочные оценки параметров положения.

ПЛАН:

1. Определение случайной величины (СВ). Множество значений. Канал наблюдения.
2. Генеральная совокупность и выборка.
3. Параметры, описывающие генеральную совокупность:

- Параметры положения СВ (мода, медиана и среднее значение, квартили, процентили);
- Параметры разброса СВ (дисперсия и стандартное отклонение, разброс);

4. Графические модели выборочного распределения:

- Диаграмма размаха (ящик с усами).
- Гистограмма как форма визуализации распределения выборочных значений СВ.

5. Выборочные оценки параметров положения.

1.1. Случайная величина (СВ)

Для того, чтобы можно было использовать вероятностные распределения, удобно закодировать результаты наблюдения параметров числами (даже если параметр является категориальным) и считать, что в результате эксперимента какая-то величина (параметр) принял какое-то значение.

Вероятностная модель процесса нам может понадобиться для того, чтобы:

- пользоваться моделями для генерации данных и моделировании и исследовании свойств процесса;
- уметь пользоваться свойствами этих моделей для оценки неопределенности результата прогноза, анализа аномалий и выбросов;
- уметь по выборке экспериментальных данных строить модели, описывающие случайную составляющую наблюдений.

Определение СВ. Примеры

> **Опр.** *Случайная Величина* (СВ) - это параметр, значения которого представляют собой закодированные числами исходы некоторого случайного явления или эксперимента (наблюдения). Так как результат эксперимента заранее неизвестен, то и неизвестно заранее какое значечние примет данная величина. Поэтому она называется Случайной Величиной.

> **Опр**. *Непрерывной случайной величиной* (НСВ) будем называть случайную величину, которая принимает произвольное числовое значение в некотором интервале (в объединении интервалов).

> **Опр**. *Дискретной случайной величиной* (ДСВ) будем называть случайную величину, которая принимает конечное или счетное количество возможных числовых значений.

Это нас не ограничивает при работе с категориальными параметрами. Мы можем каждый исход для категориального параметра закодировать числом.

Обозначения

> Случайные величины обозначают заглавными латинскими буквами (X, Y, Z), а их значения - строчными буквами (x_i, y_i, z_i).

Пример 1.1. Пол пациента.

- категориальный параметр, может принимать значение М или Ж (Male, Female);
- значения можно закодировать числами 0 или 1;
- ДСВ.

Пример 1.2. Возраст пациента, полных лет

- числовой параметр, может принимать значение от 14 до 150 (если это не детская больница);
- ДСВ.

Пример 1.3. Вес пациента, кг

- числовой параметр, может принимать значения от 20 до 200 кг (не детская больница);
- НСВ.

Модели СВ

> **Лингвистическая (описательная) модель СВ** задает *множество ее возможных значений* и описывает смысл величины, что она показывает.

> **Модель данных СВ** – это когда вы к лингвистической модели добавили описание *канала наблюдения* параметров и добавили наблюдаемые данные.

> **Канал наблюдения СВ**. Данные о процессе мы получаем с помощью неидеальных каналов наблюдения. Наблюдаемое значение параметра отличается от истинного измеряемого свойства объекта.

> **Пример канала измерения роста**. При измерении роста, мы округляем его значение с точностью до 1 см. Кроме того, результат измерения может быть разным утром и вечером.

Поэтому очень важно при описании Модели СВ указать не только множество возможных значений, но и точно описать канал измерения/наблюдения параметра:

- Как он измеряется?
- С помощью какого инструмента?
- В какие моменты времени?
- В каких условиях?

В дальнейшем мы всегда будем работать с полученными в результате наблюдения данными, т.е. с моделью данных.

1.2. Генеральная совокупность и выборка

Предположим, мы захотели узнать информациб о весе пациентов, лечащихся в больнице, построить распределение СВ "вес пациента". Как быть? Если бы смогли провести взвесить всех пациентов, прошедших лечение и будущих, то получили бы **генеральную совокупность** значений СВ и смогли бы на основе этой генеральной совокупности делать выводы.

Но что, если таких пациентов у нас сейчас в болнице более 1000? А в прошлом было более 10000? И не у всех измеряли вес, а только у тех, кто лечился от ожирения или сердечной недостаточности. И тем более каждого будущего пациента взвесить просто нереально.

Опр. Генеральной совокупностью для данной СВ назовем все мыслимое (конечное или бесконечное) множество значений СВ, которое мы могли бы получить при наблюдении данной СВ.

Имея такое множество значений, мы могли бы построить истинный закон распределения СВ. Т.е. можно считать, что:

> ЗНАНИЕ ГЕНЕРАЛЬНОЙ СОВОКУПНОСТИ == ЗНАНИЕ ЗАКОНА РАСПРЕДЕЛЕНИЯ СВ

ВЫБОРКА

А можно ли взвесить из всех только 50-100 пациентов и на основе этих данных построить распределение и его использовать вместо истинного? Насколько сильно данное эмпирическое распределение будет отличаться от истинного? Такая часть из генеральной совокупности (значения СВ, полученные для части объектов) называется **ВЫБОРКОЙ** или выборочными значениями.

И иногда даже нас интересует не все распределение, а только его некоторые характеристики - среднее и разброс значений СВ. И оценки таких характеристик на основе только выборки, как части ГС называются выборочными оценками. И вот они, к нашему счастью, подчиняются определенным распределениям (нормальному, хи-квадрат). А значит, мы можем оценить их поведение и понять как обеспечить их устойчивость и приближение к истинным значениям, которые мы могли бы получить, зная всю генеральную совокупность.

1.3. Параметры, описывающие генеральную совокупность

Генеральная совокупность для СВ – это бесконечное множество значений наблюдений. Мы будем моделировать генеральную совокупность с помощью выборки – конечного числа наблюдений. Для решения задач машинного обучения объем выборки может быть равен 30, а может и 30 млн. (для задач обучения LLM и того больше).

Тем не менее, иногда для описания ГС нам достаточно всего несколько параметров.

Пример 1.4. Количество голов.

При отборе нападающего в команду нам достаточно знать сколько голов он забивает в среднем за матч. Т.е. СВ здесь - "Кол-во голов, забитых нападающем в одном матче". Ясно, что для любого игрока это заранее предсказать невозможно. Но, имея большую выборку, т.е. зная сколько голов он забивал в каждом матче за прошлый сезон, мы можем оценить сколько он в среднем забивает за матч, т.е. найти среднее значение СВ. Тем самым мы нашли удобную характеристику ГС.

Пример 1.5. Температура больного

У больного, находящегося на лечении, температура может меняться каждый час. Мы измеряем ее каждые два часа, чтобы следить за динамикой болезни. Явно, средняя температура больного за день или несколько дней - это не та характеристика, которая будет интересовать нас. Здесь подойдет другая характеристика - максимальная температура или ее колебания, разброс.

Параметры положения

В статистике **параметры положения** (или меры центральной тенденции) — это параметры СВ, которые описывают положение наблюдаемых значений СВ на числовой оси, указывая на "центр" распределения. Они дают представление о том, где концентрируются значения в наборе данных.

К основным параметрам положения относятся:

1. **Среднее арифметическое (M)**: сумма всех значений, делённая на количество этих значений. Характеризует центр распределения.
$$M(X) = \frac{x_1 + x_2 + \ldots + x_n}{n}\ , (n \to \infty)$$
2. **Медиана (Me)**: значение, которое делит упорядоченную выборку пополам, то есть половина значений меньше медианы, а другая половина — больше.
$$Me(X) : p(X \le Me(X)) = p(X \ge Me(X)) = 0.5$$
3. **Квартили (Q1, Q2, Q3)**: это значения, которые делят упорядоченную выборку в отношении 25% выборочных значений меньше него и 75% больше него (**Q1**), 50% - 50% (**Q2**), 75% - 25% (**Q3**).
$$Q1(X) : p(X \le Q1(X)) = 0.25\ ;\ p(X \ge Q1(X)) = 0.75$$
$$Q3(X) : p(X \le Q3(X)) = 0.75\ ;\ p(X \ge Q3(X)) = 0.25$$
4. **Процентили Pk (P1,P5, P10, ..., P99)**: значения, которые делят упорядоченную выборку в отношении k% выборочных значений меньше него и (100 - k)% больше него,
$$P_k(X) : p(X \le P_k(X)) = k/100\ ;\ p(X \ge P_k(X)) = 1 - k/100$$
5. **Мода (Mo)** - это значение, которое чаще всего принимает СВ (наибольшая частота появления в ГС).
$$Mo(X) = argmax(p(X == x))$$

Q2 - и медиана - одно и тоже значение.

ОПРОС 1.1.

Вернемся к Примеру 1.1. СВ "Пол пациента", где 0 - кодирует "Female", а 1 - кодирует "Male".

Мы взяли данные по полу всех пациентов, находящихся сейчас в больнице на лечении. Тогда среднее значение СВ "пол пациента" означает:

> A. Долю женщин, находящихся на лечении в больнице среди всех лечащихся пациентов

> B. Долю мужчин, находящихся на лечении в больнице среди всех лечащихся пациентов

> C. значение не имеет смысла

ОПРОС 1.2.

Вернемся к Примеру 1.1. СВ "Пол пациента", где 0 - кодирует "Female", а 1 - кодирует "Male".

Мы взяли данные по полу всех пациентов, находящихся сейчас в больнице на лечении. Тогда какие значения может принимать медиана СВ "пол пациента":

> A. значение в данном случае не имеет смысла;

> B. может принимать любое значение в интервале от 0 до 1;

> C. может принимать значение или 0 или 1.

Пример 1.6. Зарплата работника отдела "Статистики".

Поясним разницу среднего и медианы на простом примере. Пусть в отделе работает 5 чел. Зарплаты у них следующие:

- Начальник отдела: 200 т.р.
- зам. начальника: 150 т.р.
- инженер 1 кат.: 50 т.р.
- инженер 2 кат. с выслугой лет: 35 тыс. руб.
- инженер 2 кат.: 25 тыс. руб.

Средняя зарплата по отделу: $M(25, 35, 50, 150, 200) = (200 + 150 + 50 + 35 + 25)/5 = 460/5 = 92$тыс. руб.

Медианная зарплата по отделу: $Me(25, 35, 50, 150, 200) = 50$тыс. руб.

1-й квартиль

Параметры разброса

Параметры разброса в статистике — это показатели, которые характеризуют степень индивидуальных отклонений от "центра" ГС.

Основные показатели разброса:

1. ***Дисперсия (Var — varience)*** — это среднее арифметическое квадратов отклонений каждого значения от среднего. Она измеряет степень разброса значений случайной величины относительно её среднего значения. Чем выше дисперсия, тем больше разброс значений.
$$D(X) = \frac{(X_1 - M(X))^2 + (X_2 - M(X))^2 + \ldots + (X_n - M(X))^2}{n}, (n \to \infty)$$
2. ***Стандартное отклонение (Std — standard deviation)*** является квадратным корнем из дисперсии и выражается в тех же единицах, что и исходные данные.
$$Std(X) = \sqrt{D(X)}$$
3. ***Размах (R — range)*** — это самая простая мера разброса (вариации) данных, численно равная диапазону между минимальным и максимальным значением выборки. Работает только для СВ, принимающих конечные значения.
$$R(X) = max(X) - min(X)$$
4. ***Интерквартильный размах (IQR — inter-quartile range)*** — это мера разброса в статистике, которая представляет собой разницу между третьим (Q3) и первым (Q1) квартилями в упорядоченном наборе данных. Он охватывает центральные 50% данных и служит показателем разброса в этой области. Является устойчивым к выбросам, так как не зависит от крайних значений.
$$IQR(X) = Q3(X) - Q1(X)$$

Проиллюстрируем рассмотренные показатели на примере датасета Cluster patients by demographics, health metrics, and engagement patterns. see https://www.kaggle.com/datasets/nudratabbas/patient-segmentation-data.

```
import os
import pandas as pd
import kagglehub
import numpy as np # фреймворк работы с массивами данных
import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)
import matplotlib.pyplot as plt # библиотека для рисования графиков
import seaborn as sns # библиотека для графичесмкого анализа данных

# Download latest version
web_path = kagglehub.dataset_download("nudratabbas/patient-segmentation-data")
csv_file = 'patient_segmentation_dataset.csv'
print(f"Path <{web_path}> to dataset file <{csv_file}>")

try:
    file_path = os.path.join(web_path, csv_file)
    data = pd.read_csv(file_path)
    print(f"Successfully loaded '{csv_file}' into DataFrame 'data'.")
    print("Here's a brief overview of the loaded data:")
    data.info()
except:
    print("No CSV file found in the dataset directory.")
```

```
Using Colab cache for faster access to the 'patient-segmentation-data' dataset.
Path </kaggle/input/patient-segmentation-data> to dataset file <patient_segmentation_dataset.csv>
Successfully loaded 'patient_segmentation_dataset.csv' into DataFrame 'data'.
Here's a brief overview of the loaded data:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2000 entries, 0 to 1999
Data columns (total 16 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   PatientID               2000 non-null   object
 1   Age                     2000 non-null   int64
 2   Gender                  2000 non-null   object
 3   State                   2000 non-null   object
 4   City                    2000 non-null   object
 5   Height_cm               2000 non-null   int64
 6   Weight_kg               2000 non-null   int64
 7   BMI                     2000 non-null   float64
 8   Insurance_Type          2000 non-null   object
 9   Primary_Condition       1505 non-null   object
 10  Num_Chronic_Conditions  2000 non-null   int64
 11  Annual_Visits           2000 non-null   int64
 12  Avg_Billing_Amount      2000 non-null   float64
 13  Last_Visit_Date         2000 non-null   object
 14  Days_Since_Last_Visit   2000 non-null   int64
 15  Preventive_Care_Flag    2000 non-null   int64
dtypes: float64(2), int64(7), object(7)
memory usage: 250.1+ KB
```

The dataset contains 2,000 patient records with comprehensive information including:

> Demographics:

- Age, Gender, Geographic location (State, City)

> Health Metrics:

- Height, Weight, BMI (Body Mass Index)
- Number of chronic conditions
- Primary medical condition

> Healthcare Utilization:

- Annual visit frequency
- Days since last visit
- Average billing amount per visit

> Другие (страховка, ...)

```
from scipy import stats # импорт Stats

# посмотрим на основные характеристики числовых СВ - среднее, стандартное отклонение, мин-макс, квартили
data.describe()
```

	Age	Height_cm	Weight_kg	BMI	Num_Chronic_Conditions	Annual_Visits	Avg_Billing_Amount	Days_Sinc
count	2000.00000	2000.000000	2000.000000	2000.000000	2000.000000	2000.000000	2000.000000	
mean	50.69550	167.907500	85.143500	30.740650	1.080000	5.466500	4000.270500	
std	15.44445	12.684494	20.385428	8.839952	0.890504	3.485965	2463.239215	
min	18.00000	145.000000	50.000000	13.400000	0.000000	1.000000	207.000000	
25%	40.00000	158.000000	67.000000	23.800000	1.000000	3.000000	2061.000000	
50%	51.00000	168.000000	86.000000	30.100000	1.000000	4.000000	3707.250000	
75%	63.25000	177.000000	103.000000	37.025000	1.000000	8.000000	5650.875000	
max	87.00000	195.000000	120.000000	57.100000	3.000000	12.000000	12467.500000	

```
# выделим столбцы с числовыми признаками для расчета характеристик СВ

int_columns = data.select_dtypes('int').columns
float_columns = data.select_dtypes('float').columns
numeric_columns = list(int_columns) + list(float_columns)

int_columns, float_columns
numeric_columns
```

```
['Age',
 'Height_cm',
```

```
 'Weight_kg',
 'Num_Chronic_Conditions',
 'Annual_Visits',
 'Days_Since_Last_Visit',
 'Preventive_Care_Flag',
 'BMI',
 'Avg_Billing_Amount']
```

```
# Вычислим параметры положения для каждой из этих СВ

M = data[numeric_columns].mean()
print(f"{25*'-'} средние СВ:\n {M} \n")

Me = data[numeric_columns].median()
print(f"{25*'-'} медианы СВ:\n {Me} \n")
```

```
------------------------- средние СВ:
 Age                        50.69550
Height_cm                  167.90750
Weight_kg                   85.14350
Num_Chronic_Conditions       1.08000
Annual_Visits                5.46650
Days_Since_Last_Visit      180.08500
Preventive_Care_Flag         0.46400
BMI                         30.74065
Avg_Billing_Amount        4000.27050
dtype: float64

------------------------- медианы СВ:
 Age                        51.00
Height_cm                  168.00
Weight_kg                   86.00
Num_Chronic_Conditions       1.00
Annual_Visits                4.00
Days_Since_Last_Visit      183.00
Preventive_Care_Flag         0.00
BMI                         30.10
Avg_Billing_Amount        3707.25
dtype: float64
```

```
# соберем параметры положения в одну таблицу, чтобы сравнить значения
disp_chars = pd.DataFrame({
    'признак': ('возраст', 'рост', 'вес', 'кол-во хронич.', 'визитов/год', 'дней после визита', 'превент.лечение', 'BMI',
    'среднее': M,
    'медиана': Me
}).set_index('признак')

disp_chars
```

признак	среднее	медиана
возраст	50.69550	51.00
рост	167.90750	168.00
вес	85.14350	86.00
кол-во хронич.	1.08000	1.00
визитов/год	5.46650	4.00
дней после визита	180.08500	183.00
превент.лечение	0.46400	0.00
BMI	30.74065	30.10
плата	4000.27050	3707.25

Далее: New interactive sheet

Процентили — это значения, которые делят упорядоченный набор данных на 100 равных частей. Например, 95-й процентиль отделяет самые низкие 95% значений от самых высоких 5%. 25-й процентиль совпадает с Q1, 50-й процентиль совпадает с Q2 (медианой), а 75-й процентиль совпадает с Q3.

Вычислим в качестве примера 5-й и 95-й процентили.

```
# добавим другие параметры положения

Q1 = data[numeric_columns].quantile(0.25)
Q3 = data[numeric_columns].quantile(0.75)
```

```
P_05 = data[numeric_columns].quantile(0.05)
P_95 = data[numeric_columns].quantile(0.95)
min_val = data[numeric_columns].min()
max_val = data[numeric_columns].max()

# соберем параметры положения в одну таблицу
disp_chars = pd.DataFrame({
    'признак': ('возраст', 'рост', 'вес', 'кол-во хронич.', 'визитов/год', 'дней после визита', 'превент.лечение', 'BMI',
    'среднее': M,
    'медиана': Me,
    'Q1': Q1,
    'Q3': Q3,
    'P_05': P_05,
    'P_95': P_95,
    'минимум': min_val,
    'максимум': max_val
}).set_index('признак')
```

признак	среднее	медиана	Q1	Q3	P_05	P_95	минимум	максимум
возраст	50.69550	51.00	40.0	63.250	24.950	75.000	18.0	87.0
рост	167.90750	168.00	158.0	177.000	148.000	190.000	145.0	195.0
вес	85.14350	86.00	67.0	103.000	53.000	117.000	50.0	120.0
кол-во хронич.	1.08000	1.00	1.0	1.000	0.000	3.000	0.0	3.0
визитов/год	5.46650	4.00	3.0	8.000	1.000	12.000	1.0	12.0
дней после визита	180.08500	183.00	90.0	268.000	19.950	348.000	1.0	365.0
превент.лечение	0.46400	0.00	0.0	1.000	0.000	1.000	0.0	1.0
BMI	30.74065	30.10	23.8	37.025	17.700	46.605	13.4	57.1
плата	4000.27050	3707.25	2061.0	5650.875	592.975	8248.025	207.0	12467.5

Далее: New interactive sheet

Gemini

```
variance = data[numeric_columns].var
std_dev = data[numeric_columns].std
range_val = max_val - min_val
iqr_val = Q3 - Q1

# соберем параметры разброса в одну
dispersion_chars = pd.DataFrame({ __
    'признак': ('возраст', 'рост',
    'дисперсия': variance,
    'станд. отклонение': std_dev,
    'размах': range_val,
    'интерквартильный размах': iqr_v
}).set_index('признак')

dispersion_chars
```

признак	дисперсия	станд. отклонение	размах	интерквартильный размах
возраст	2.385310e+02	15.444450	69.0	23.250
рост	1.608964e+02	12.684494	50.0	19.000
вес	4.155657e+02	20.385428	70.0	36.000
кол-во хронич.	7.929965e-01	0.890504	3.0	0.000
визитов/год	1.215195e+01	3.485965	11.0	5.000
дней после визита	1.095968e+04	104.688484	364.0	178.000
превент.лечение	2.488284e-01	0.498827	1.0	1.000
BMI	7.814475e+01	8.839952	43.7	13.225
плата	6.067547e+06	2463.239215	12260.5	3589.875

Далее: New interactive sheet

1.4. Графические модели выборочного распределения

Удобно анализировать сразу несколько параметров СВ с помощью графического представления выборки. Рассмотрим два таких представления

Диаграмма размаха (ящик с усами).

Основные элементы Boxplot:

- Ящик (Box): Центральная часть, где находится 50% данных, от 25-го (Q1) до 75-го (Q3) процентиля.
- Медиана (Median): Линия внутри ящика, разделяющая данные пополам (50-й процентиль).
- Усы (Whiskers): Линии, идущие от ящика вверх и вниз, показывающие диапазон данных, не считающихся выбросами.
- Выбросы (Outliers): Отдельные точки за пределами «усов», представляющие аномально большие или малые значения.

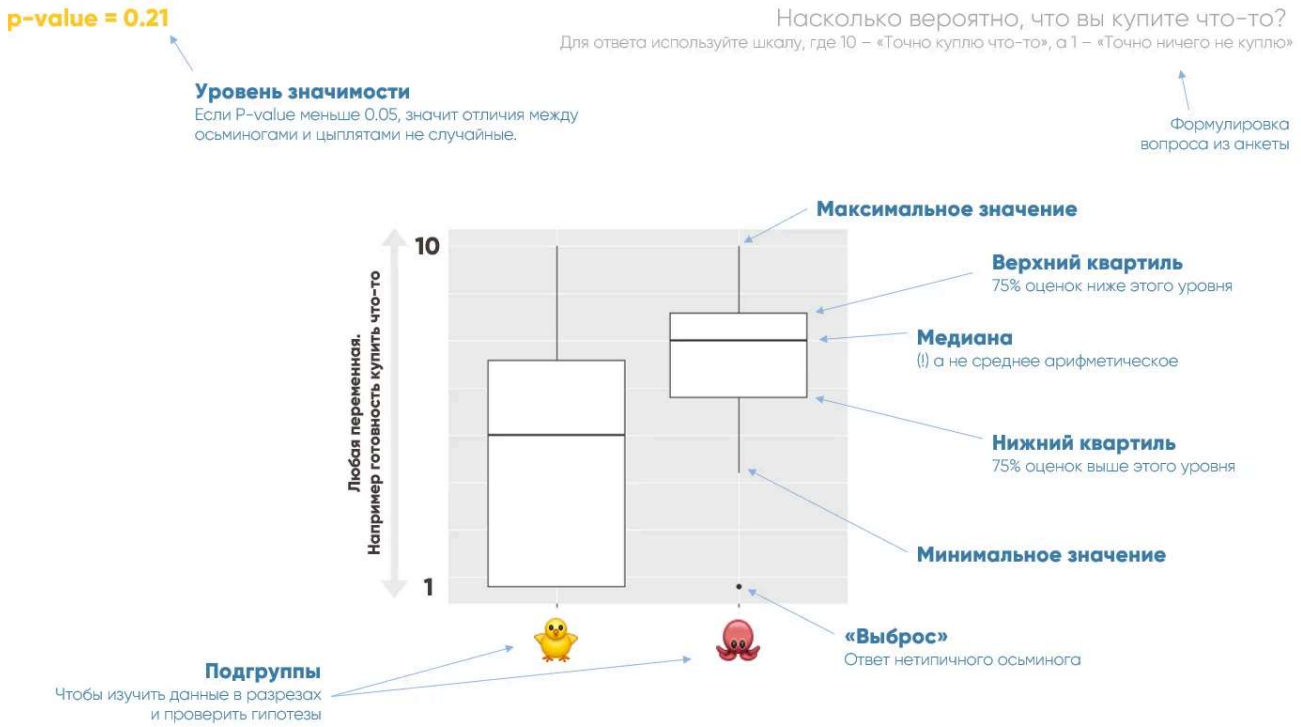
Что можно увидеть с помощью Boxplot:

- Центральные значения: Медиана и квартили (Q1, Q3).
- Разброс данных: Ширина ящика (IQR) показывает плотность данных; длина "усов" – степень вариации.
- Симметричность/Асимметрия: Если медиана смещена к одному из концов ящика или усы разной длины, данные асимметричны.
- Выбросы: Аномальные значения, которые могут требовать дальнейшего исследования.

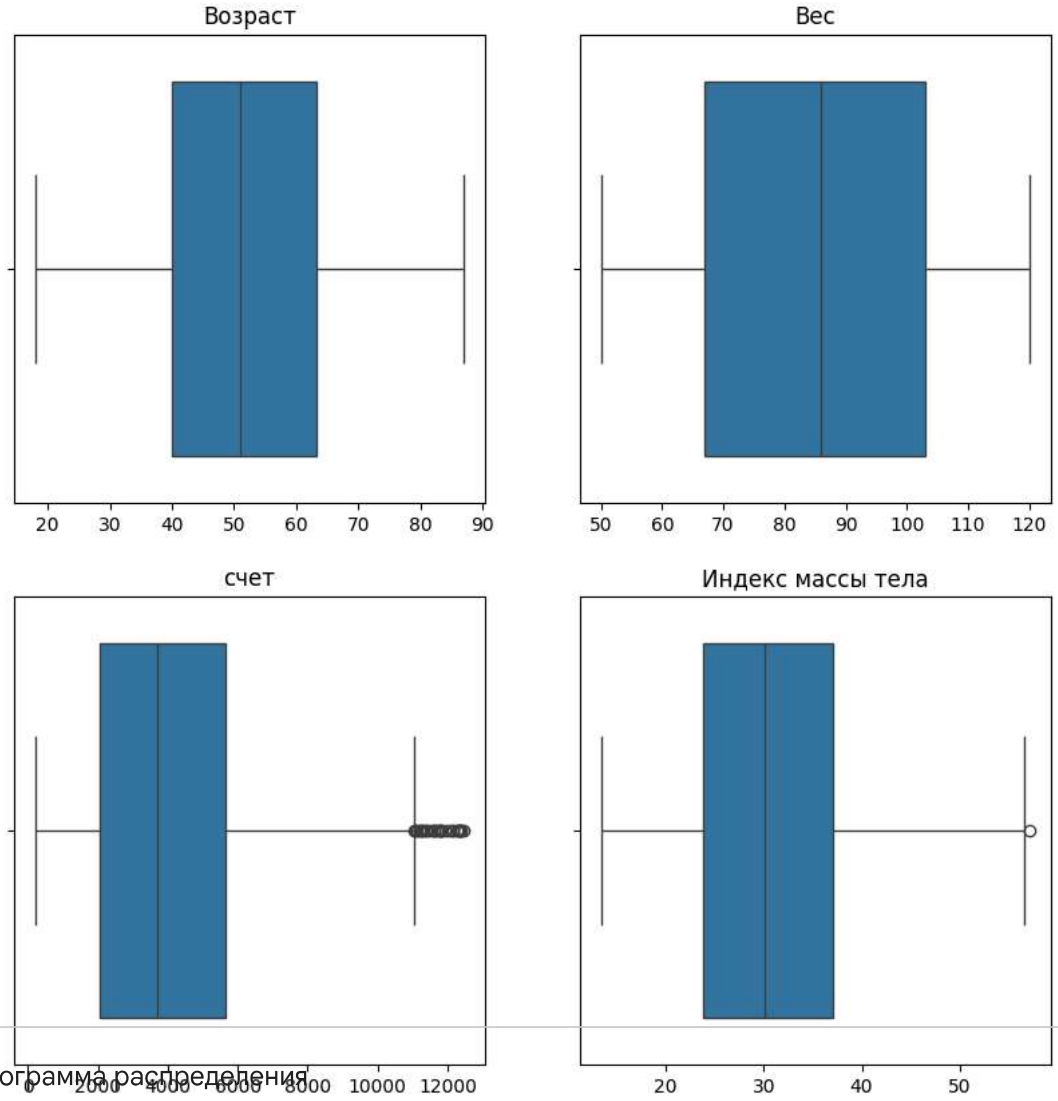
Применение:

Boxplots отлично подходят для сравнения распределений нескольких групп данных (например, продаж разных продуктов, результатов тестов разных классов) и для выявления аномалий

см. https://www.tidydata.ru/boxplot




```
fig, ax = plt.subplots(nrows=2, ncols=2, figsize=(10,10))
ax[0,0] = sns.boxplot(ax=ax[0,0], x = data['Age'])
ax[0,1] = sns.boxplot(ax=ax[0,1], x = data['Weight_kg'])
ax[1,0] = sns.boxplot(ax=ax[1,0], x = data['Avg_Billing_Amount'])
ax[1,1] = sns.boxplot(ax=ax[1,1],x = data['BMI'])
ax[0, 0].set_title('Возраст')
ax[0, 1].set_title('Вес')
ax[1, 0].set_title('счет')
ax[1, 1].set_title('Индекс массы тела')
ax[0, 0].set(xlabel = '') # пустые кавычки удаляют подпись по оси (!)
ax[0, 1].set(xlabel = '')
ax[1, 0].set(xlabel = '')
ax[1, 1].set(xlabel = '');
```



Гистограмма распределения

Гистограмма - это наиболее подробная графическая иллюстрация распределения значений СВ по оси. Применяется в основном для визуализации НСВ (гистограмма распределения).

Опр. *Гистограмма распределения* — это способ визуализации количественных данных, представляющий собой столбчатую диаграмму, где высота каждого соприкасающегося прямоугольника пропорциональна числу значений (частоте) выборки, попавших в соответствующий интервал (бин).

Для построения гистограммы распределения необходимо:

- определить интервал значений СВ X (X_{min}, X_{max});
- определить количество интервалов (бинов), на которые разбивается интервал X (X_{min}, X_{max}); в этом случае все интервалы одинаковы по ширине; если мы хотим разные интервалы, то необходимо задать границы интервалов;
- по оси X откладываются интервалы значений, а по оси Y — плотность попадания туда СВ, т.е. частота, количество попаданий в интервал значений, деленная на ширину интервала;
- общая площадь прямоугольников гисторамммы должна равняться 1 (если используется относительная частота) или количеству наблюдений.

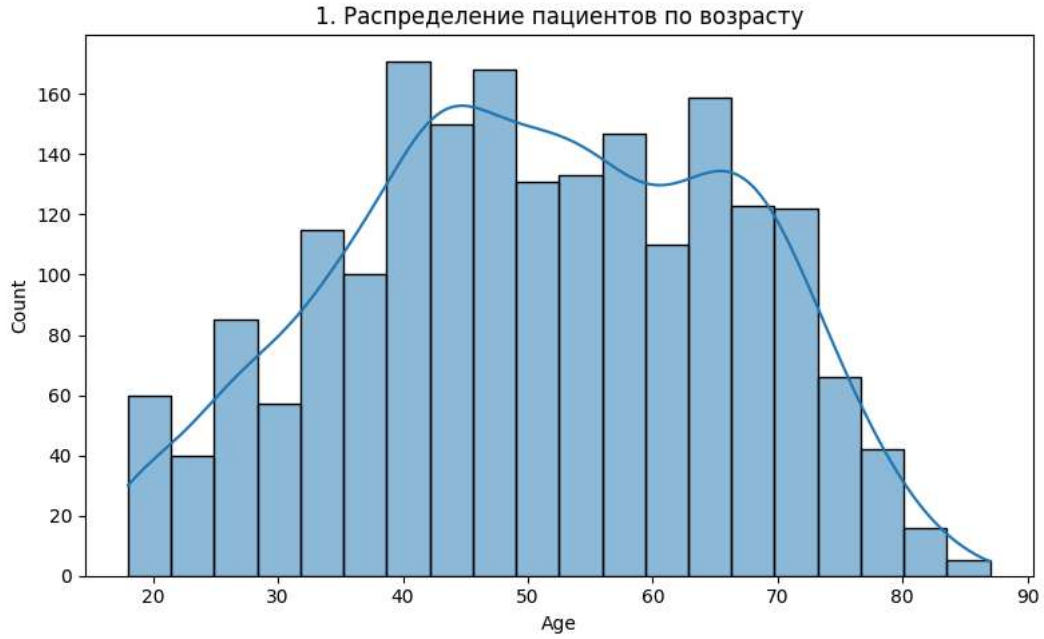
Применение: Используется для анализа непрерывных величин в статистике, контроле качества, и для оценки плотности вероятности.

Что можно увидеть с помощью гистограммы:

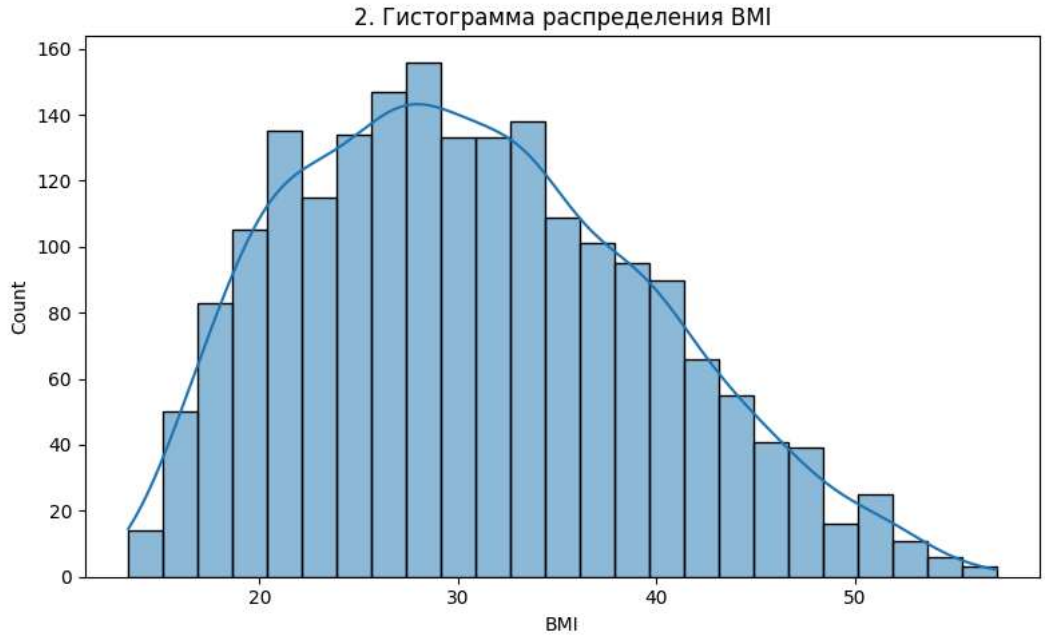
- Она позволяет наглядно оценить структуру данных, их разброс, среднее значение и закон распределения, показывая, как часто значения встречаются в определенных диапазонах.
- Помогает выявить центр распределения, его симметричность и наличие выбросов.
- Гистограммы помогают понять, соответствует ли распределение нормальному закону, при котором большинство значений концентрируется вокруг среднего.

```
def show_fig():
    plt.tight_layout()
    plt.show()
```

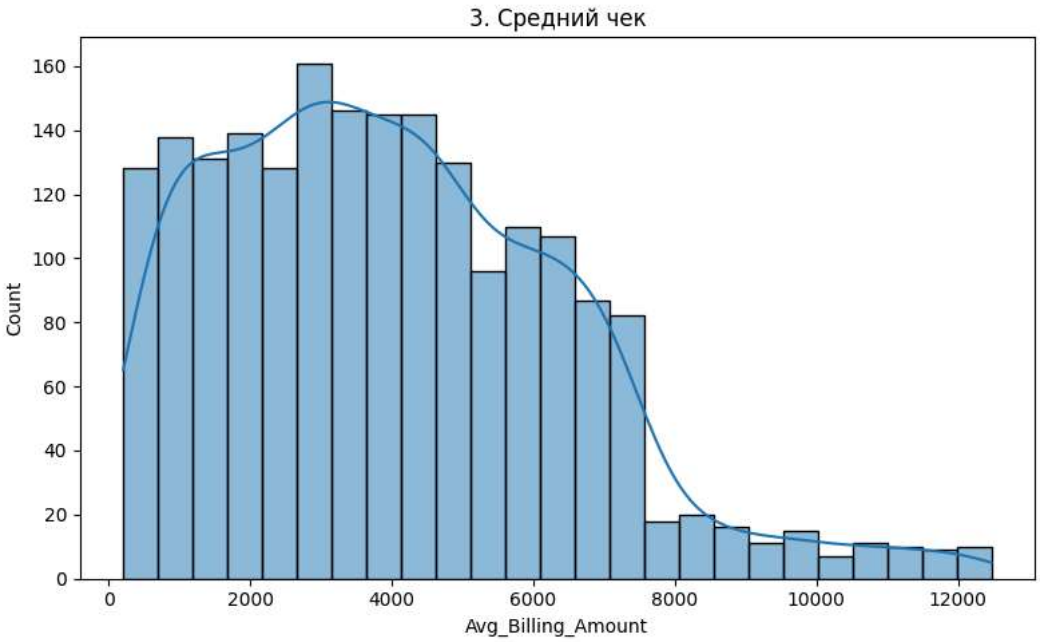
```
fig = plt.figure(figsize=(8,5))
sns.histplot(data['Age'], bins=20, kde=True, )
plt.title(f'{plot_no}. Распределение пациентов по возрасту')
show_fig()
plot_no += 1
```




```
fig = plt.figure(figsize=(8,5))
sns.histplot(data['BMI'], bins=25, kde=True)
plt.title(f'{plot_no}. Гистограмма распределения BMI')
show_fig()
plot_no += 1
```




```
fig = plt.figure(figsize=(8,5))
sns.histplot(data['Avg_Billing_Amount'], bins=25, kde=True)
plt.title(f'{plot_no}. Средний чек')
show_fig()
plot_no += 1
```



ОПРОС 1.3. Анализ гистограммы

Посмотрите на гистограмму среднего чека и определите по ней моду:

A. $Mo(X) \approx 0.5$;

B. $Mo(X) \approx 160$;

C. $Mo(X) \approx 3000$;

D. $Mo(X) \approx 5000$;

1.5. Выборочные оценки параметров распределения СВ

- На практике мы всегда работаем с выборками из ГС объема **n**: $(x_1, x_2, .., x_n)$.
- На основе этих выборок мы можем построить эмпирический закон распределения (например, гистограмму распределения). Для того, чтобы построить эмпирический закон распределения (гистограмму распределения) необходимо большое количество наблюдений (для устойчивости оценок h_i). Не всегда нам доступно большое кол-во наблюдений и тем более не всегда эти наблюдения охватывают весь спектр возможных значений СВ.
- с практической точки зрения гораздо продуктивнее сделать предположение о виде распределения (или смеси распределений), а затем оценить параметры распределения (среднее, дисперсию, ...); но об этом мы поговорим в следующей теме.

Таким образом, мы приходим к необходимости изучения выборочных оценок параметров, которые сами являются СВ. И эти СВ подчиняются определенным законам распределения. Так что мы должны научиться строить эти законы и определять свойства этих оценок.

Опр. Выборкой из ГС размера n будем называть набор значений СВ X, который мы наблюдали или можем наблюдать при n экспериментах.

> Обозначать конкретную выборку из ГС СВ X размером n будем набором $(x_1, .., x_n)$. Обозначать мыслимую выборку из ГС СВ X размером n (векторную СВ) будем большими буквами $(X_1, X_2, \ldots, X_n)$.

Опр. Статистика — числовая функция от выборки:

$$f(X_1, X_2, \ldots, X_n)$$

Разные статистики могут быть оценками одного и того же параметра СВ. Мы рассмотрим наиболее практичные.

Выборочные оценки параметров положения

- в качестве наилучшей статистики, оценивающее среднее СВ является среднее выборочное значение:

$$x_{\text{ср}}(n) = f(X_1, X_2, \ldots, X_n) = \frac{X_1 + X_2 + \ldots + X_n}{n}$$

- если из полученной выборки $(x_1, x_2, \ldots, x_n)$ построить вариацинный ряд, т.е. упорядочить значения X: $(x_{i_1} \leq x_{i_2} \leq \ldots \leq x_{i_n})$, то можно использовать выборочную медиану, в качестве оценки медианы ГС:

$$Me(X) \approx Me(x_1, x_2, .., x_n) = me : \frac{n(x_i \leq me)}{n} = \frac{n(x_j \geq me)}{n} = 0.5$$

здесь $n(x_i \leq me)$ - количество наблюдений x_i в выборке, которые меньше или равны me.

аналогично выборочные минимум и максимум могут быть оценками минимума и максимума ГС. Хотя мы понимаем, что их надо как-то корректировать в зависимости от вида распределения:

$$min(X) \leq min(x_1, x_2, .., x_n)$$
$$max(X) \geq max(x_1, x_2, .., x_n)$$

Выборочные Дисперсия и стандартное отклонение

- выборочная дисперсия:

$$S_X^2(n) = \frac{(X_1 - m_X)^2 + (X_2 - m_X)^2 + \ldots + (X_n - m_X)^2}{n}$$

- или так

$$S_X^2(n) = \frac{(X_1 - x_{\text{ср}})^2 + (X_2 - x_{\text{ср}})^2 + \ldots + (X_n - x_{\text{ср}})^2}{n} = \bar{x^2}(n) - x_{\text{ср}}^2(n)$$

- выборочное СКО:

$$S_X(n) = \sqrt{S_X^2(n)}$$

Надо понимать, что выборочные оценки сами являются СВ и их распределение сильно зависит от n, объема выборки. Ясно, что чем больше **n**, тем ближе оценка к реальному значению параметра.

И если **n** небольшое, то часто выборочные оценки необходимо корректировать, чтобы обеспечить их хорошие свойства.

В следующих уроках мы посмотрим как распределение некоторых оценок зависит от значения **n**.

```
Напишите программный код или сгенерируйте его с помощью искусственного интеллекта.
```