Основы математической статистики.

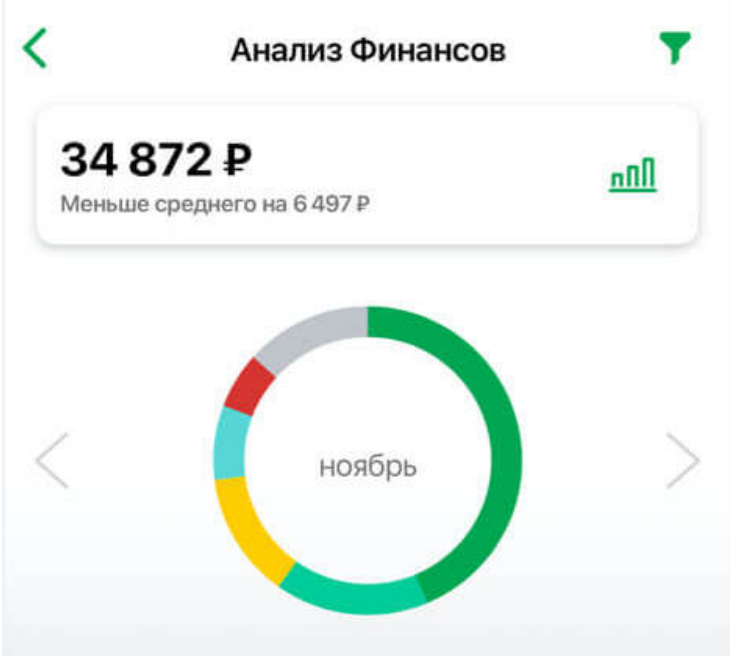
Математическая статистика изучает методы сбора и обработки статистической информации для получения научных и практических выводов. Под **статистической информацией** подразумеваются первичные данные о состоянии различных явлений, формирующиеся в процессе статистического наблюдения, которые затем подвергаются систематизации, сводке, анализу и обобщению.

Примеры:

1. Социологические исследования. Ответы респондентов, являющиеся в данном случае первичными данными, помогают выявить тенденции, закономерности и особенности в поведении людей, их мнениях, отношениях и восприятии окружающего мира. Например, краткая аналитика из исследования стриминговых сервисов музки:

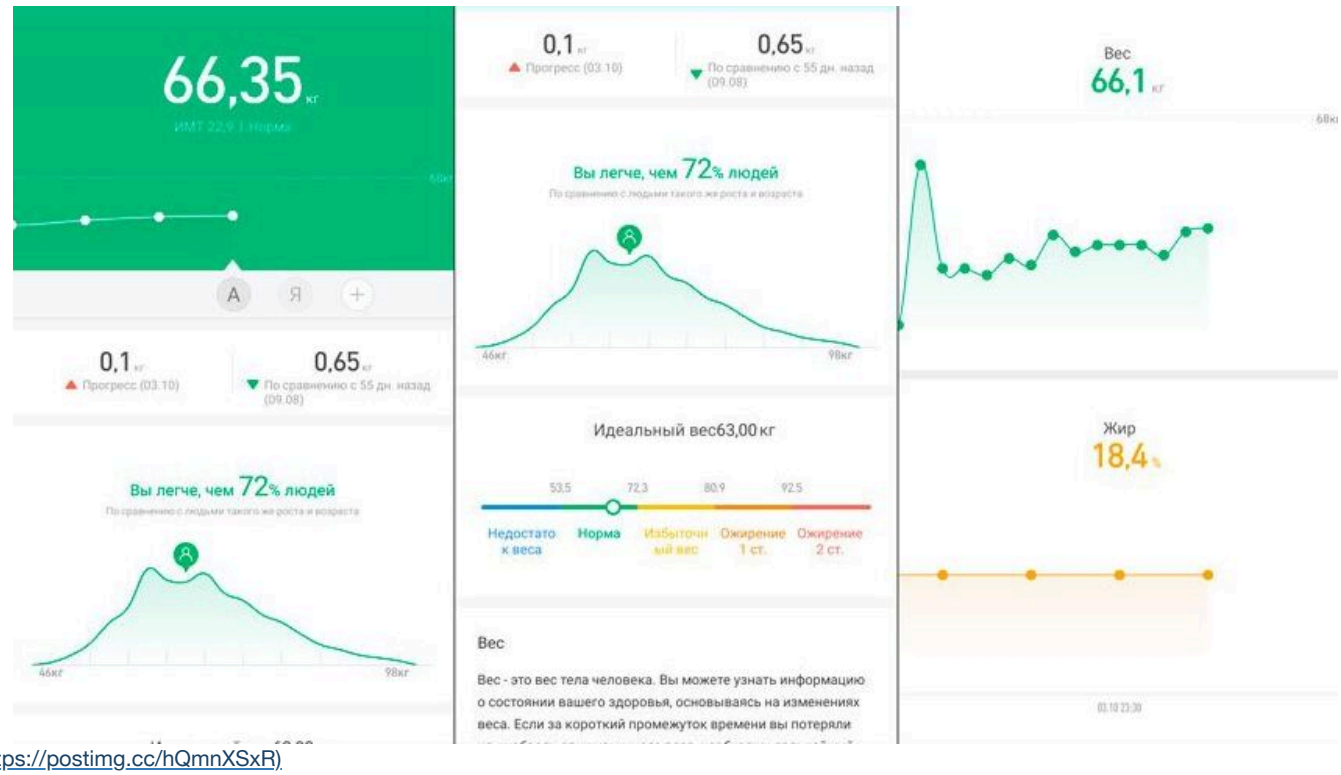
   > Платную подписку оформляют чаще всего на «Яндекс.Музыке» (63%). Также 18% пользуются платно сервисом «VK Музыка», а 6% – «МТС Music». Несмотря на сложности с оплатой, 11% приобретают подписку на Apple Music, 6% – YouTube Music (например, оплачивают со счета мобильного телефона). Отметим, что 4% опрошенных обходят блокировку и продолжают пользоваться Spotify, и 1% – платформой Deezer, которая тоже ограничила доступ российским пользователям прошлой весной.

2. Анализ финансов в приложении банка. В данном примере статистическая информация - транзакции с карты, которые в последсвии распределяются по категориям и обобщаются в диаграмму, по которой можно сделать выводы о тратах.



(https://postimg.cc/FdRsH5Dh)

3. Измерения умных весов. Первичными данными в этом примере являются измерения веса, которые обобщаются в графики динамики веса, показатели ИМТ и пр.



(https://postimg.cc/hQmnXSxR)

Основные статистические понятия

Генеральная совокупность - совокупность всех потенциально возможных вариантов, которые можно получить при одинаковых условиях.

Выборка - данные, которые получены при наблюдениях. Размер выборки конечен и ограничен критериями — методами отбора.

Репрезентативность - понятие, которое говорит, насколько показательна выборка, реалистично ли в ней распределены варианты. Выборка считается репрезентативной, если в ней учтено множество параметров и она достоверно отражает генеральную совокупность.

Распределение — это описание того, с какой частотой в переменной встречаются определённые значения.

Числовые характеристики

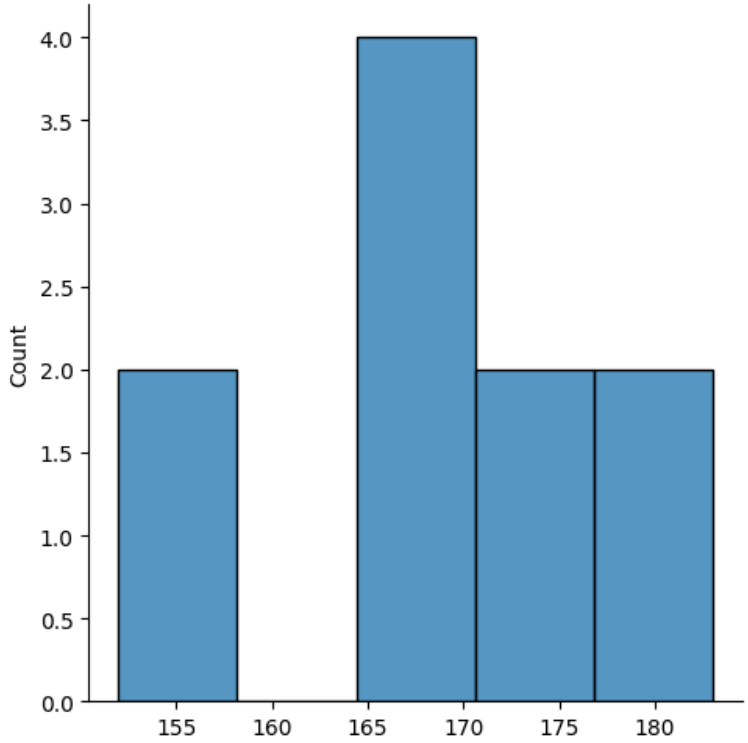
Среднее арифметическое — усредненное значение среди всех показателей.
Медиана — значение, которое находится посередине распределения.
Мода — значение, которое встречается в выборке чаще всего.
Размах — разница между минимальным и максимальным значением.
Дисперсия — отклонение значений от среднего арифметического.

Ниже рассмотрим на примере основные статистические понятия:

In [1]:
```
# Проведем исследование роста студентов Москвы. В данном случае генеральная совокупность – рост с
тудентов во всех вузах Москвы.
# Выборка – рост студентов в аудитории (будет ли данная выборка репрезентативной?)
import numpy as np # Импорт библиотеки для работы с массивами
sample = np.array([166, 183, 152, 165, 171, 167, 180, 170, 155, 173]) # Сбор данных
```

In [2]:
```
import matplotlib.pyplot as plt # Загрузка библиотек визуализации (установка через pip install "н
аименование пакета" в ячейке)
import seaborn as sns
# Визуализация графика распределения роста
sns.displot(data=sample)
```

Out[2]:
```
<seaborn.axisgrid.FacetGrid at 0x1230bb210>
```



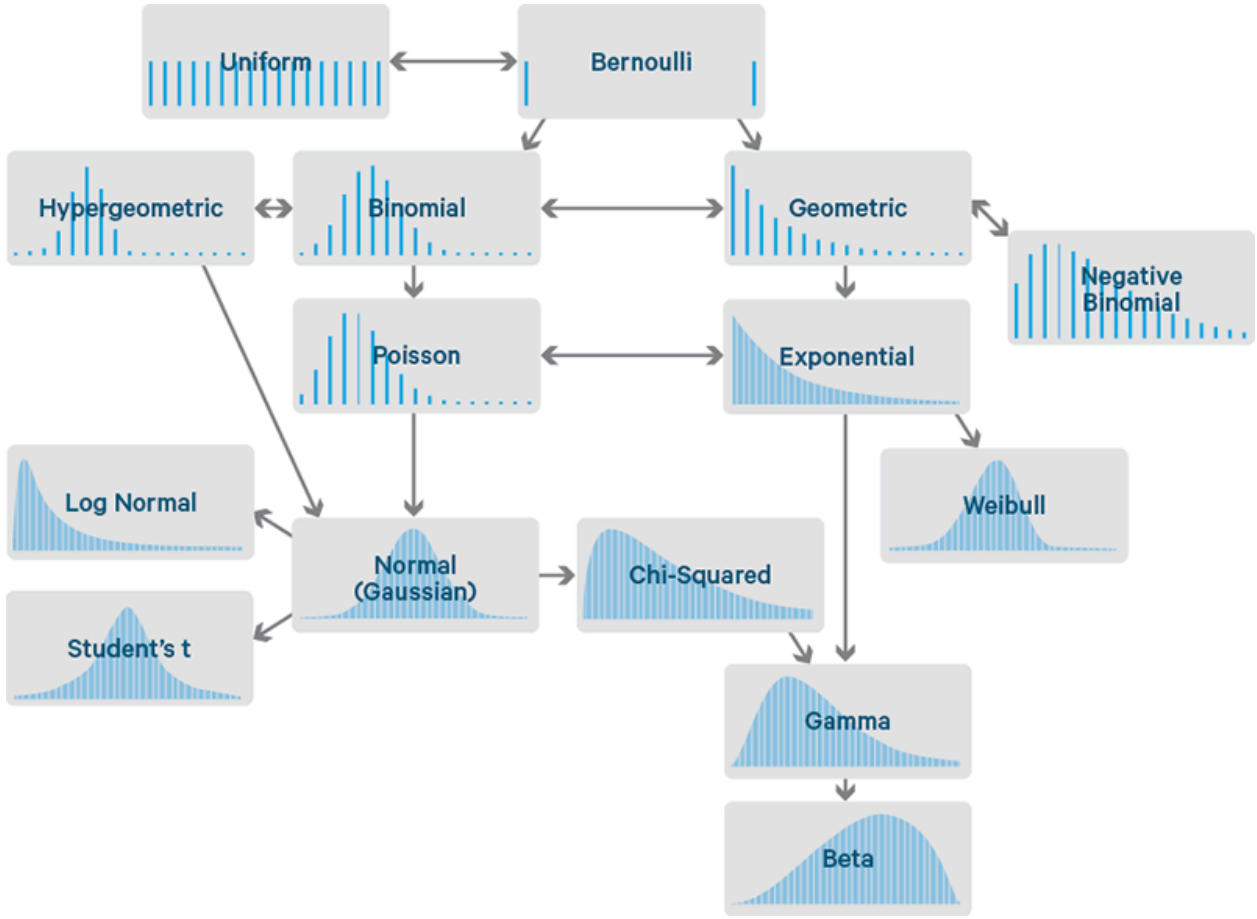
In [3]:
```
print('Среднее:', np.mean(sample)) # Вычисление среднего значения
print('Медиана:', np.median(sample)) # Вычисление медианы
print('Стандартное отклонение:', np.std(sample)) # Вычисление стандартного отклонения
print('Дисперсия:', np.var(sample)) # Вычисление дисперсии
```

```
Среднее: 168.2
Медиана: 168.5
Стандартное отклонение: 9.195651146058118
Дисперсия: 84.55999999999999
```

Виды распределений.

Распределение — это функция, которая показывает возможные значения переменной и частоту их появления. Она обеспечивает математическое описание поведения данных, указывая, где сосредоточено большинство точек данных и как они распределены. Распределения делятся на дискретные и непрерывные:

1. Дискретные распределения используются для описание событий с конечным или счетным числом исходов, например, успех или неудача, целое число (например, игра в рулетку, в кости), орёл или решка и т.д.
2. Непрерывные распределения используются для данных, значения которых могут принимать любое значение в некотором интервале. Например: температура, рост, вес, и т.д.
   Примеры основных распределений:



(https://postimg.cc/6TvyTxdJ)

Распределение данных играет важную роль в выборе методов для исследования. Выделяют параметрические и непараметрические методы:

1. Параметрические методы — это статистические методы, которые основаны на определённых предположениях о распределении данных. Эти методы обычно предполагают, что данные подчиняются известному распределению вероятностей, например нормальному распределению, и оценивают параметры этого распределения с помощью имеющихся данных.
2. Непараметрические методы — это статистические методы, которые не основаны на конкретных предположениях о распределении изучаемой совокупности. Эти методы часто называют «свободными от распределения», поскольку они не предполагают никаких предположений о форме распределения.
   Работу с распределением данных рассмотрим в следующем пункте.

Алгоритм проведения анализа данных.

Основные этапы анализа данных:

1. Сбор данных. На данном этапе определяются источники данных и методы сбора. (Например, для социологического опроса подбирается группа респондентов и составляются вопросы)
2. Предобработка данных. Этап включает в себя чистку данных, анализ пропущенных значений, дубликатов, преобразование типов и т.д. (Например, при исследовании среднего дохода населения важно удалить выбросы - доход олигархов во избежание завышенных статистических показателей и представления смещенных результатов)
3. Анализ данных. Данный этап включает в себя применение статистических методов для выявления закономерностей в данных, классификаций и пр. (Например, прогнозирование дохода магазина на следующий год)
4. Визуализация данных и интерпретация результатов для принятия решений. (Например, в результате анализа были выявлены потенциальные клиенты для рассылки промокодов и привлечения мх в магазин)

In [4]:
```
# Рассмотрим алгоритм проведения анализа данных на примере
# основная задача исследования - определить существует ли разница между весом пингвинов самок и самцов
import pandas as pd #Импортируем библиотеку для работы с таблицами
df = sns.load_dataset("penguins") #Загружаем датасет с исследованием пингвинов из библиотеки seaborn
df.head() #Выводим первые строки датасета для проверки импорта методом .head()
```

Out[4]:

	species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex
0	Adelie	Torgersen	39.1	18.7	181.0	3750.0	Male
1	Adelie	Torgersen	39.5	17.4	186.0	3800.0	Female
2	Adelie	Torgersen	40.3	18.0	195.0	3250.0	Female
3	Adelie	Torgersen	NaN	NaN	NaN	NaN	NaN
4	Adelie	Torgersen	36.7	19.3	193.0	3450.0	Female

In [5]:
```
# анализ пропущенных значений
df.info() # Выводим основную информацию о датасете, в колонке Non-Null Count можно заметить наличие пропущенных значений
# Для работы с пропущенными значениями можно использовать метод .fillna(),
#  либо удалить пропущенные значения если их мало и невозможно восстановить
df = df[~df['sex'].isna()] #  df['sex'].isna() - отбирает пропущенные значения по полю 'sex',
# ~ перед условием обозначает логическое отрицание, df[<условие>] возвращает необходимые строки
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 344 entries, 0 to 343
Data columns (total 7 columns):
 #   Column             Non-Null Count  Dtype  
---  ------             --------------  -----  
 0   species            344 non-null    object 
 1   island             344 non-null    object 
 2   bill_length_mm     342 non-null    float64
 3   bill_depth_mm      342 non-null    float64
 4   flipper_length_mm  342 non-null    float64
 5   body_mass_g        342 non-null    float64
 6   sex                333 non-null    object 
dtypes: float64(4), object(3)
memory usage: 18.9+ KB
```

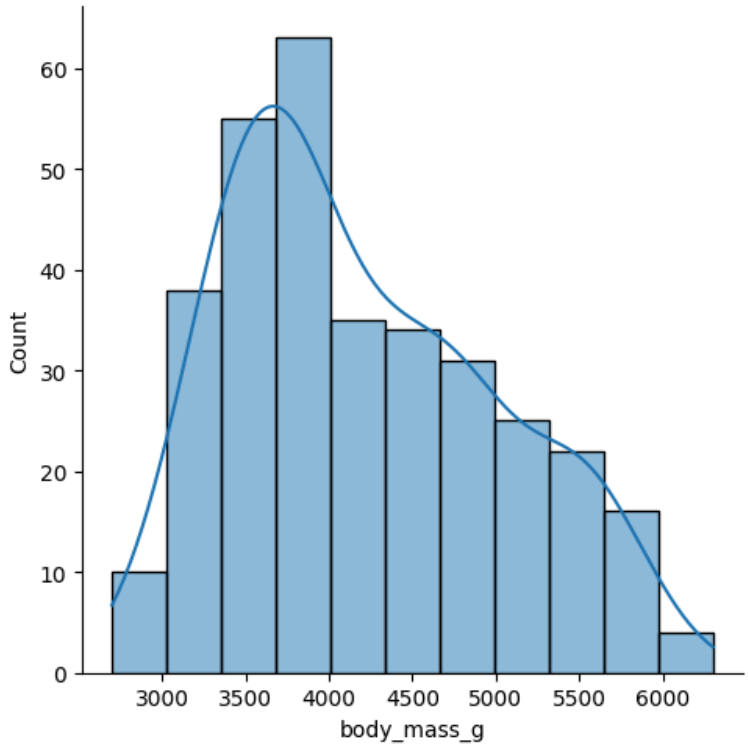
In [6]:
```
# вывод общей статистической информации о данных
df.describe()
```

Out[6]:

	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g
count	333.000000	333.000000	333.000000	333.000000
mean	43.992793	17.164865	200.966967	4207.057057
std	5.468668	1.969235	14.015765	805.215802
min	32.100000	13.100000	172.000000	2700.000000
25%	39.500000	15.600000	190.000000	3550.000000
50%	44.500000	17.300000	197.000000	4050.000000
75%	48.600000	18.700000	213.000000	4775.000000
max	59.600000	21.500000	231.000000	6300.000000

In [7]:
```
# Для применения статистических методов проверяем распределение данных на нормальность
# Строим график распределения массы пингвинов (визуальный вариант проверки)
sns.displot(data=df, x="body_mass_g", kde = 'True') # Указываем источник data=df, ось x - "body_mass_g"
# Параметр kde = 'True' показывает график ядерной оценки плотности - сглаживает данные и помогает оценить распределение наглядно
```

Out[7]:
```
<seaborn.axisgrid.FacetGrid at 0x1237d9710>
```



In [8]:
```
# Проверяем данные на нормальность тестом Шапиро-Уилка
# В данном тесте выдвигается нулевая гипотеза H0: "случайная величина распределена нормально", альтернативная H1: "распределение не нормальное"
import scipy # Импортируем библиотеку со статистическими методами
stat, p = scipy.stats.shapiro(df['body_mass_g']) # проводим тест Шапиро-Уилка, он возвращает значение статистики и p-value
# P-value - это минимальный уровень значимости, на котором нулевая гипотеза может быть отвергнута.
# другими словами, p-value - это вероятность того, что нулевая гипотеза жизнеспособна,
# чем выше p-value, тем лучше, если мы хотим, чтобы нулевая гипотеза не была отвергнута.
print('Statistics=%.3f, p-value=%.3f' % (stat, p))
alpha = 0.05 # alpha - уровень значимости, который задается самостоятельно, обычно используют alpha = 0.05
if p > alpha:
    print('Принять гипотезу о нормальности')
else:
    print('Отклонить гипотезу о нормальности')
# p-value=0, следовательно, нет оснований принимать нулевую гипотезу. В исследовании используем непараметрические методы.
```

```
Statistics=0.958, p-value=0.000
Отклонить гипотезу о нормальности
```

In [9]:
```
# Далее нам необходимо определить существует ли разница между весом пингвинов самок и самцов
# для этого можно расчитать средние значения
df[['body_mass_g','sex']].groupby(by='sex').mean()
# видно, что вес самцов больше чем у самок, но можно ли данны результат назвать статистическим значимым?
# т.е. можем ли мы сказать, что такие результаты не случайны и при проведении других экспериментов не окажется, что вес самки больше?
```

Out[9]:

sex	body_mass_g
Female	3862.272727
Male	4545.684524

In [10]:
```
# для задач проверки средних значений используется t-тест (для нормальных данных) и U-критерий Манна – Уитни (если распределение неизвестно)
# H0: две группы имеют одинаковое распределение, H1: одна группа имеет большие (или меньшие) значения, чем другая
male = df['body_mass_g'][df['sex']=='Male'] # массив с самцами
female = df['body_mass_g'][df['sex']=='Female']  # массив с самками
stat, p = scipy.stats.mannwhitneyu(male,female) # Используем метод mannwhitneyu для проверки гипотезы
print('Statistics=%.3f, p-value=%.3f' % (stat, p))
alpha = 0.05
if p > alpha:
    print('Принять гипотезу об отсутсвии различий')
else:
    print('Отклонить гипотезу об отсутсвии различий')
# отклоняем нулевую гипотезу, следовательно, между весом самца и самки есть статистическая разница
```

```
Statistics=20845.500, p-value=0.000
Отклонить гипотезу об отсутсвии различий
```

In [11]:
```
# Визуализировать разницу можно при помощи графика ящика с усами
sns.boxplot(data=df, x='sex', y='body_mass_g')
```

Out[11]:
```
<Axes: xlabel='sex', ylabel='body_mass_g'>
```