ハロートレーニング
急がば学べ

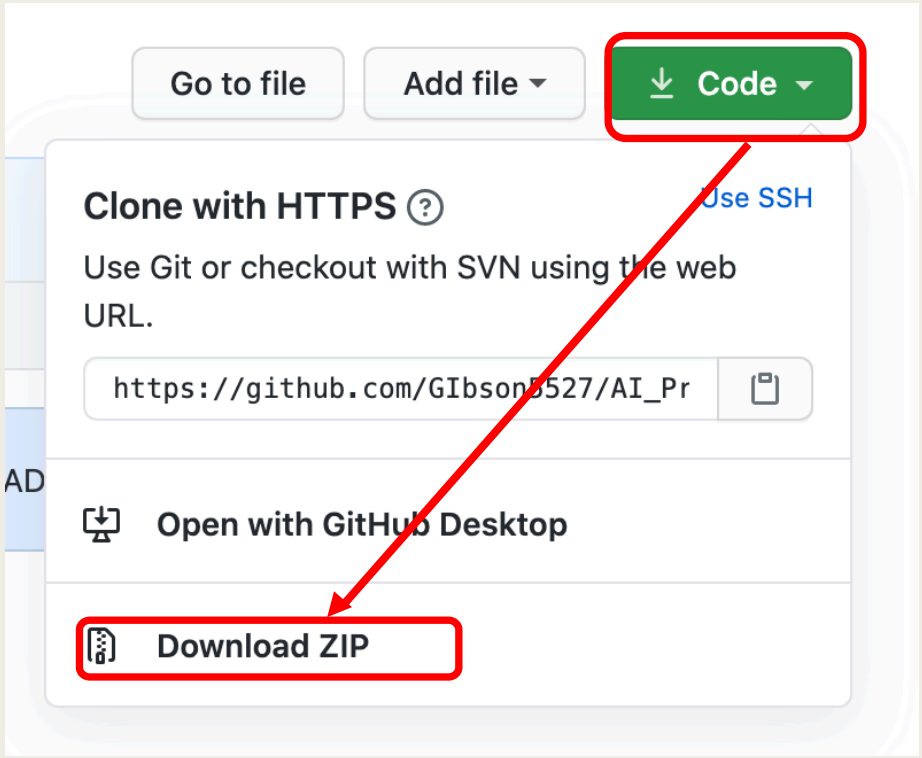
技能向上訓練
データサイエンス
プログラミング
コース

基本統計学

データのダウンロード

- 以下のURLより、データのダウンロードをしてください。
- https://github.com/GIbson5527/DS_1022.git

- ダウンロード後、ZIPファイルを展開し、
- 自身のGoogleアカウントのGoogleDriveに保存



ノートブックを開く

- GoogleDrive上の「bike.ipynb」を開いてください。

統計学

- 収集したデータを活用し、未知のデータの推測や、将来の動きを予測するもの

標本と母集団

- 標本：手持ちのデータ
- 母集団：手に入れてない未知のデータも含んだ、全てのデータ

例：　釣り堀にいる全ての魚　→母集団
　　釣り人が釣った魚　　→標本

統計学の分類

- 記述統計学
- 推測統計学
- ベイズ統計学

記述統計学

- 分かりにくいデータを、分かりやすいデータに変換して表現する。
- 収集したデータの特徴を、平均や分散、標準偏差などから求めます。

推測統計学

- 限られたデータから調査したい母集団全体の特徴を推測する。

- 収集したデータの特長を、サンプルデータをもとに推測します。例えば国民1万人のアンケート結果から、国民全体1億人の動向を推測します。

ベイズ統計学

- ベイズ統計学は必ずしも標本となるデータを必要とせず、データ不十分でも何とかして確率を導きます。

ベイズ統計の利用例

■ 迷惑メール判別

- ユーザーがスパムとしたメール（以下、スパムメール）と、スパムではないとしたメール（以下、正常なメール）から、タイトル、本文に含まれる語句ごとの出現確率（＝特徴）を抽出、点数をつけ、スパムと正常なメールを判別するための閾値を導き出します。新規メールを受信したら、そのタイトルや本文を自然言語処理（＝単語に分割）し、閾値と照らし合わせ、スパムメールである確率が高ければスパムメールとして振り分けます。

基本統計量

- 分布の中心を示す
  - 平均値
  - 中央値（メジアン）
  - 最頻値（モード）
- 分布の広がりを示す
  - 範囲
  - 平方和
  - 分散
  - 不偏分散
  - 標準偏差
  - 不偏標準偏差

Google Colaboratoryの準備

Pythonにおいて、統計指標をとるためのライブラリをインポートします。

平均値

相加平均

- N個のデータの総和を求め、Nで割る。
  （一般的な平均）

相乗平均（幾何平均）

- N個のデータを総乗（全てかける）し、N乗根（$\sqrt[N]{X}$）を求める。

データの作成

```
data = [8, 17, 0, 11, 6, 21, 16, 6, 17, 11, 7, 9, 6, 13, 12, 16, 3, 14, 13, 12]
```

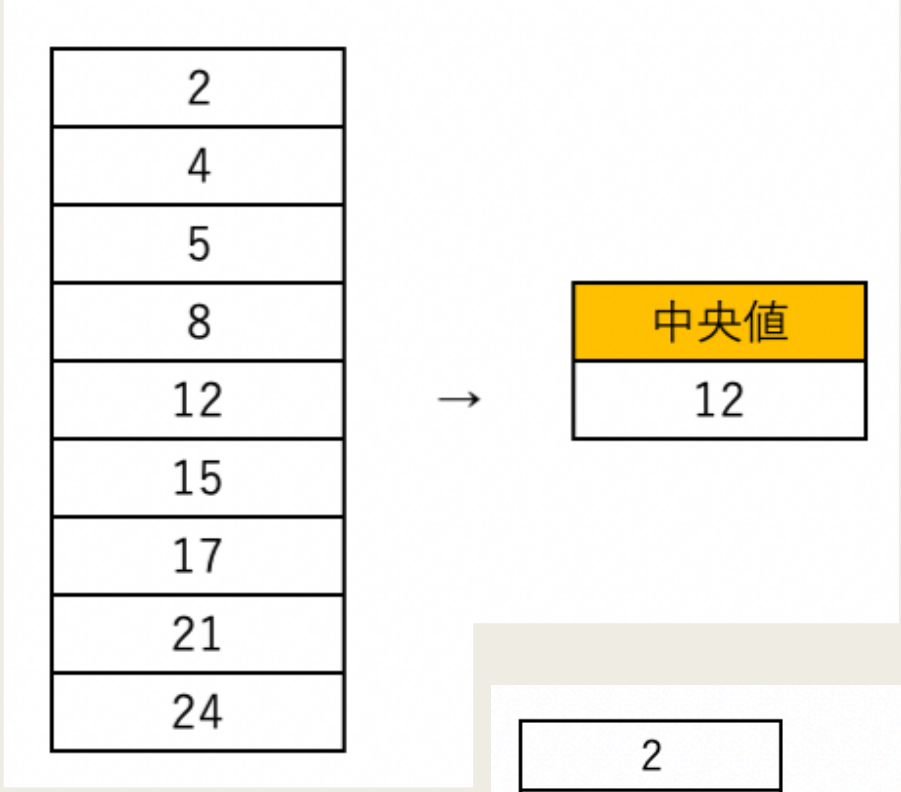
- 上記のコードを実行し、データを作成しておく。

平均（相加平均）を求める。

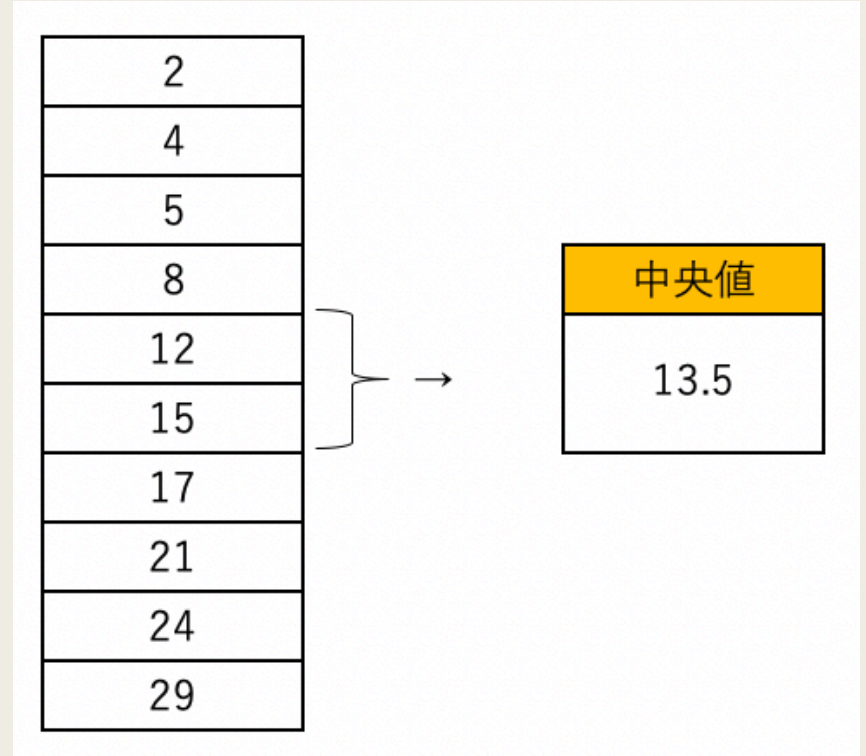
```
print("平均:", st.mean(data) )
```

```
平均: 10.9
```

中央値（メジアン）



- 中央値とは
  - 複数のデータを整列した時に、中央にくる値
  - データ数が偶数の場合、中央の値２つの平均



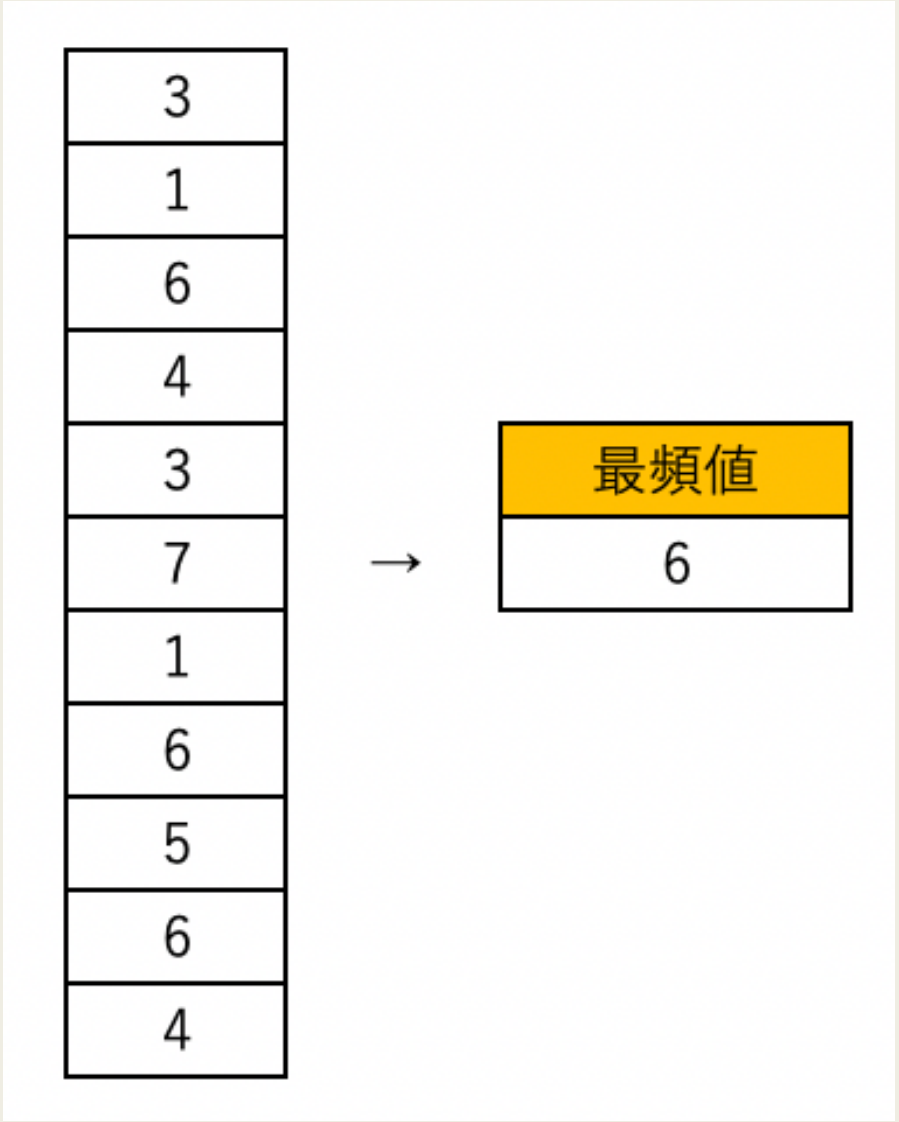
中央値を求める。

```
print("中央値:" , st.median(data))
```

```
中央値: 11.5
```

最頻値（モード）

- 最頻値とは
  - データの中で、最も多く出現するデータ



最頻値を求める。

```
print("最頻値:" , st.mode(data))
```

```
最頻値: 6
```

最大値、最小値を求める

```
print("最大値:" , max(data)) #最大値を求める処理を追加
print("最小値:" , min(data)) #最小値を求める処理を追加
```

```
最大値: 21
最小値: 0
```

分散

- データと平均の差を２乗を平均したもの
- データ全体が、平均からどの程度散らばっているか

- 母集団の分散を「母分散」
- 標本の分散を「標本分散」「不偏分散」

(標本の分散は、データ数−１で割る）

分散を求める。

```
print("母分散:" , st.pvariance(data))
print("標本分散" , st.variance(data))
```

```
母分散: 26.49
標本分散 27.884210526315789
```

- pvariance : データを母集団として捉え、分散を求める
- variance : データを標本として捉え、分散を求める

標準偏差

- 分散の平方根
- 分散で求めた散らばり具合を、元のデータの単位に近づける

- 母集団の標準偏差を「母標準偏差」
- 標本の分散を「標本標準偏差」「不偏標準偏差」

標準偏差を求める

```
print("母標準偏差:" , st.pstdev(data))
print("標本標準偏差" , st.stdev(data))
```

```
母標準偏差: 5.146843692983108
標本標準偏差 5.280550210566679
```

- pstdev : データを母集団として捉え、標準偏差を求める
- stdev : データを標本として捉え、標準偏差を求める