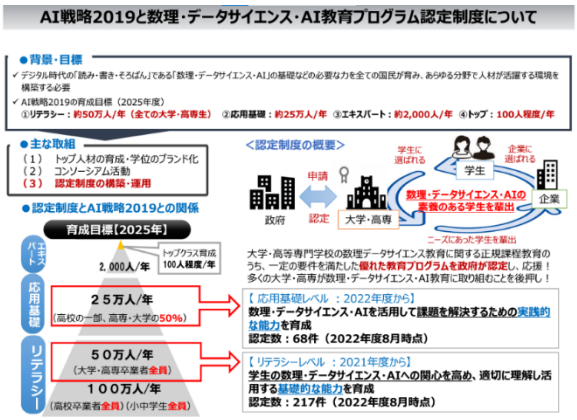
Python で学ぶ
データサイエンス入門
（データサイエンス編）

10.　データの活用

10.1　データサイエンス

日本政府は「AI 戦略 2019」において「読み・書き・そろばん」的な素養として「数理・データサイエンス・AI」の基礎などの必要な力を、すべての国民が育むことを目標としている。ここで、**データサイエンスとはデータから価値を引き出すこと**。そのために数学や統計学、プログラミングなどの様々な手法を用いる。この章では統計学の基礎を学習していく。ただし、数学の授業ではないので、定理の証明を行ったり、公式等を用いて計算を行ったりすることはない。



出典：文部科学省ホームページ　（https://www.mext.go.jp）

問　次の表から B 組と C 組の平均点には偶然ではない差があるといえるか。併せて理由も考える。

番号	1	2	…	40	41	42	43	44	45	平均	標準偏差
B 組	62	81	…	77	73	61	74	54	35	64.2	18.6
C 組	71	54	…	63						68.45	14.17

自分の意見

周囲の人の意見

平均点に 4 点差があるが、この 4 点を大きいとみるか、小さいとみるかは人それぞれ。主観的な主張では論理的とは言えない。これを科学的な視点で考察を行う手法について学習を進めていく。

10.2　統計学の基本

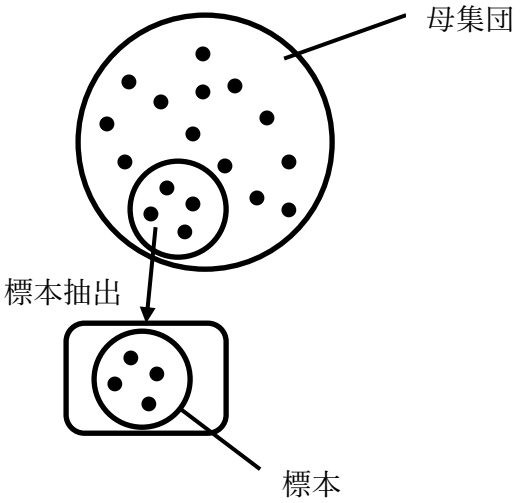
データを収集する方法として、自ら実験を行ったり、アンケートを実施したりする方法がある。他にもインターネット上のデータを用いる方法もあるが、インターネット上のデータはすべて正しいとは限らないため、データの信ぴょう性を確認する必要がある。このようなインターネット上のデータとして**オープンデータ**がある。オープンデータとは国や地方公共団体及び事業者が保有する官民データで、誰もが利用できるように公開されたデータのことである。

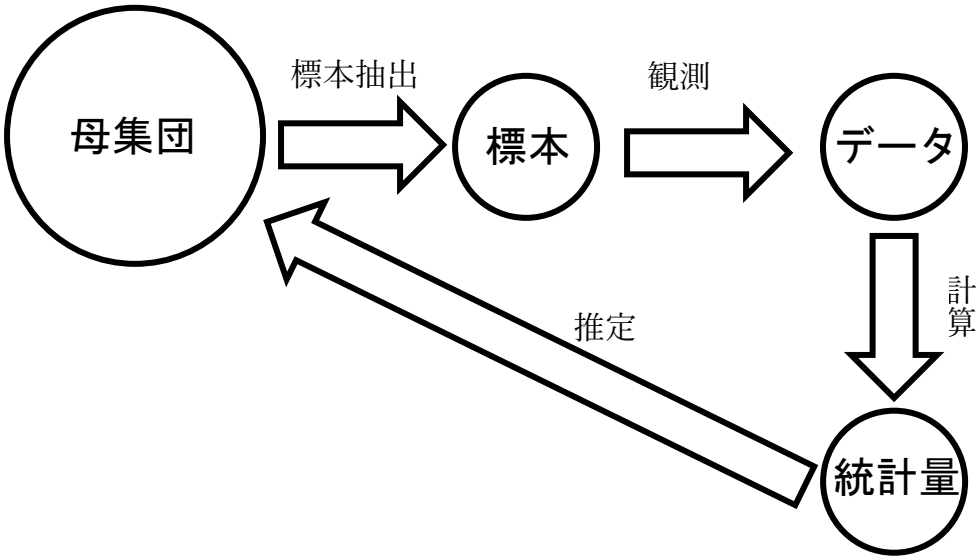
アンケートを実施してデータを収集する場合も、日本人全員（約１億２０００万人）にアンケートを実施するのは困難である。このように実際には全員に調査を実施することが現実的ではなかったり、コストがかかったりするため行えない場合がある。そこで全体から一部を取り出して調査を実施する方法が一般的である。調査対象全体を統計学では**母集団**という。母集団の一部を**標本（サンプル）**といい、母集団から標本を取り出すことを**標本抽出（サンプリング）**という。そして標本から母集団の性質を調べることを**標本調査**という。

統計学では標本に含まれる要素の数を、標本の大きさ（**サンプルサイズ**）と呼び、通常はnで表す。例えば右の図ではサンプルサイズ$n = 4$である。

例
母集団：第二高校の生徒（有限）
標本　：１年生の生徒

母集団：さいころを振って出た目（無限）
標本　：さいころを５回振って出た目





例：さいころの出る目は歪みがなければそれぞれ 1/6 の確率で出現する。このとき、さいころに歪みがないのかを検証したい。さいころは無限に振れるので、さいころを 100 回振って歪みの有無を検証する。この 100 回振ることが標本を抽出することである。出た目を観測したものがデータであり、平均値や標準偏差を計算したものが統計量である。その統計量など計算したものからさいころの歪みを推定する。

小皿で鍋全体の味を見るのが推定

10.3　統計量

得られたデータに対して何らかの計算をして得られたものを**統計量**という。統計量には、分布の大まかな位置を表す代表値である**平均値、中央値、最頻値**があり、データのばらつきを表す**分散、標準偏差**などがある。

代表値について覚えておくべきポイント

・平均値は極端に大きい（小さい）値に影響を受けやすいが、中央値、最頻値は影響を受けにくい

例：100 人の人の年収の平均値 509.3 万と中央値が 509.5 万の中に、年収 1 億円の人を追加すると、平均値が 603.3 万、中央値 509.6 万に変化する

・中央値の右側と左側には同じ数のデータが存在する

例：1 2 3 4 5 6 7 の中央値は 4。この 4 を境に右側にデータ数 3、左側にデータ数 3 存在

・代表値だけではデータの分布がどのようになっているのか分からない。

分散・標準偏差について覚えておくべきポイント

・分散も標準偏差もデータが平均値からどれだけばらついているかを表す指標

・標準偏差は分散の正の平方根をとったもの

・ばらつきが大きいと分散も標準偏差も大きくなる

例：テスト A：平均値 50、分散 10 とテスト B：平均値 50、分散 30 を比較すると、テスト B の方が平均値からのばらつきが大きい

10.4　データ分析の流れ

データ分析は目的ではなく、あくまで問題解決のための手段である。ある問題（現実と理想のギャップ）があり、その問題を解決するためにデータを収集、分析し、解決を図るのがデータ分析の流れである。この問題解決の流れは PPDAC サイクルと呼ばれている。

P(problem、問題)
P(plan、計画)
D(data、データ収集)
A(analysis、分析)
C(conclusion、結論)

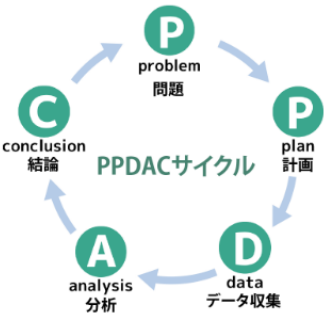


図 1：PPDAC サイクル

出典：DataStaRt　URL：https://www.stat.go.jp/dstart/

この章では Data、Analysis の部分について説明を進めていく。

10.5　データの収集

データを収集する方法は主に「実験・観察する」方法と「オープンデータを用いる」方法がある。オープンデータは Google などで検索をすると、様々な Web サイトに Excel や CSV 形式のデータで置いてある。オープンデータの Web サイトとして e-Stat、SSDSE(教育用標準データセット)などがある。そのほか警察庁や文部科学省などの Web サイトにも統計データとして配布されている場合もある。探究活動においてもデータを活用していくとよい。また、先ほどの「Data StaRt」にはデータ利活用の支援や先進事例等が掲載されているため、データの利活用や探究活動のテーマに悩んだ際には活用も可能。

図 3：SSDSE（教育用標準データセット）　**図 4：e-stat**　**図 2：Data StaRt**

実習

（1）Google ドライブに「datascience」という名前のフォルダを作成する
※半角小文字で入力すること。
フォルダ名やフォルダを作成する場所が異なるとエラーに繋がるので注意

（2）SSDSE にアクセスし、各自気になるデータを開いてみる
※Excel 形式でも CSV 形式でも可

10.6　データの分析（データの特徴を知る）

統計的な分析の手法は、データの種類やどのような分析をしたいかによって様々な手法が存在する。そのためまずはもっとも基本的な平均値や中央値、標準偏差等を用いた分析の手法を実践する。

例題 10-1

（１）classroom に配信された「10-1_平均値・中央値・標準偏差.ipynb」を開く

（２）B 組生徒の数学の点数が入力されている。この B 組生徒の数学の点数の平均値、中央値、標準偏差を求めなさい。

```
# 例題 10-1
# 以下のデータは B 組生徒の数学の点数である。
# 平均値、中央値、標準偏差を求めなさい。
import statistics as st
import io
import pandas as pd
b_data = pd.read_csv(io.StringIO('''
ID,点数
1,54
2,81
3,31
 ≈
8,71
9,41
10,92
'''), header=0)
# 平均値
b_mean =
# 中央値
b_median =
# 標準偏差
b_std =
print('平均値 =', b_mean)
print('中央値 =', b_median)
print('標準偏差 =', b_std)
```

演習 10-1

（１）例題 10-1 で使用したファイルに C 組生徒の数学の点数が入力されている。例題 10-1 を参考にして C 組生徒の数学の点数の平均値、中央値、標準偏差を求めなさい。

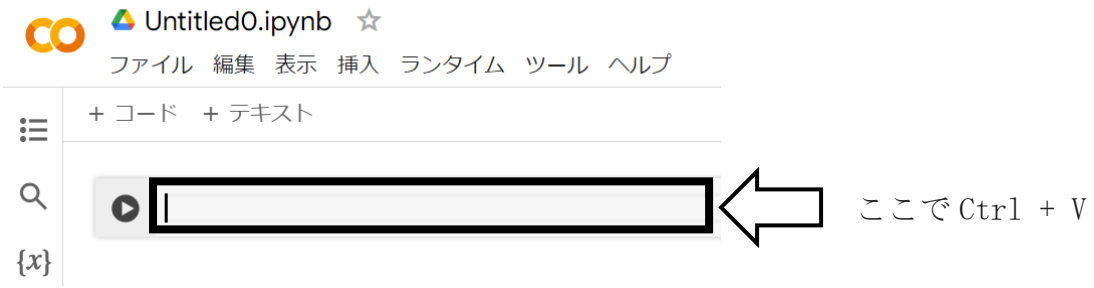
（２）例題 10-1 と演習 10-1 で得られた B 組と C 組の平均値、中央値、標準偏差を比較して分かることを述べなさい。

●便利テクニックについて（紹介）

　得点の数値などのデータを Python のコードに入力するのは非常に面倒くさい。そんなときにはスプレッドシートから必要なデータをコピー&ペーストするとよい。

手順 1：スプレッドシートの必要なセルを選択肢、コピーする（Ctrl + C）

手順 2：Google Colab のセルにペーストする（Ctrl + V）



手順 3（必要に応じて）：ヘッダー処理を行う

　デフォルトでは「header=None」となっているが、このままでは 1 行目の ID と点数もデータとしてカウントされてしまうため、1 行目をヘッダーとして読み込ませる場合には「header=0」もしくは「なにも入力しない」とすることで ID と点数をヘッダーとして読み込むことが可能。

```
import io
import pandas as pd
pd.read_csv(io.StringIO('''
ID,点数
1,54
2,81
3,31
4,65
5,65
6,56
7,65
8,71
9,41
10,92
'''), header=0)
```

	0	1
0	生徒ID	点数
1	1	54
2	2	81
3	3	31
4	4	65
5	5	65
6	6	56
7	7	65
8	8	71
9	9	41
10	10	92

図 6：header = None

	生徒ID	点数
0	1	54
1	2	81
2	3	31
3	4	65
4	5	65
5	6	56
6	7	65
7	8	71
8	9	41
9	10	92

図 5：header = 0

10.7　データの分析（2 つのデータの関係性）

2 つのデータに関係があるかどうかをみるには相関係数を求めるとよい。この相関係数は r で表される。以下相関係数についてのポイントである。

・-1 ≦ r ≦ 1
・r が正のとき正の相関、r が負のとき負の相関があるという。また、r が±1 に近いほど強い相関があるという。明確な基準は存在しないが、
・相関係数からは相関の強弱や有無しか分からない
　→変数 X と変数 Y に対して相関係数 r=0.8 のとき、**変数 X と変数 Y の間に強い相関がある**ことは分かるが、変数 X が原因で変数 Y が変化する（**因果関係**）は分からない
　例：都道府県別の佐藤姓の人数でと東北大学の合格者数のデータで相関係数を求めると r=0.8 であるので強い相関があることは分かるが、佐藤姓に変える（原因）と東北大学に合格するか（結果）は分からない。
・相関があるような結果が偶然得られる場合がある
・2 つのデータ以外に影響を及ぼしているデータが存在している場合がある
　→アイスの売り上げと水難事故件数に相関があるが、これはどちらも気温と相関があり、気温が 2 つのデータに影響を及ぼしていることが考えられる
・一般的な相関係数（ピアソンの積率相関係数）は直線的な関係を見るものである。
　→2 次関数のような関係となる場合、相関係数からは相関がないようにみえる
　※散布図を描画する学習の時に詳しく説明

例題 10-2

（1）classroom に配信された「10-2_相関係数.ipynb」のファイルを開く
（2）数学と英語の点数における相関の有無と強弱を調べなさい。

※これは架空のデータです

```
# 例題 10-2
# 数学の点数と英語の点数に相関があるか調べなさい。
import statistics as st
import io
import pandas as pd
b_data = pd.read_csv(io.StringIO('''
生徒 ID,数学,英語
1,39,40
2,46,48
  ≈
19,61,62
20,92,94
'''), header=0)
b_math = b_data['数学']
b_eng = b_data['英語']
print(                    )
```

演習 10-2

「10-2_相関係数.ipynb」のファイルに classroom に配信されたスプレッドシートのデータを貼り付け、相関の有無と強弱を調べなさい。

10.8　データの分析（データ可視化）

平均値などの統計量を計算するとデータがどのように分布しているのか想像しやすくなる。情報デザインにも通じるところだが、データを図やグラフにする（データ可視化）と短時間に多くの情報を伝えることが可能。自分で分析する場合も、他者にデータを示す場合にも活用できる。データから統計量を計算することと併せて、グラフにして可視化することでデータの傾向を読み解いていくことが大切である。

●グラフの種類（一部）
折れ線グラフ・・・時間とともに変化するデータに使用する。
棒グラフ・・・棒の高さで量の大小を比較する場合に使用する
円グラフ・・・各項目が全体に対する割合を見る場合に使用する
散布図・・・2 つの変数の間に相関があるかを見たいときに使用する
ヒストグラム・・・データの散らばり具合を見たいときに使用する
箱ひげ図・・・データの散らばり具合を見たいときに使用する
※ヒストグラムと違い、複数の散らばり具合を見る場合におすすめ

注意！
目的に応じて使用するグラフが異なる。どのような目的でデータを可視化するのかをしっかりと考えたうえで、適切なグラフを選択することが大切である

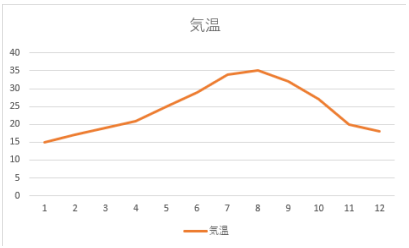


図 7:折れ線グラフ

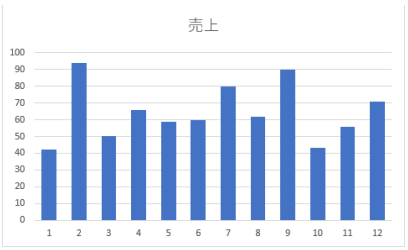


図 8:棒グラフ

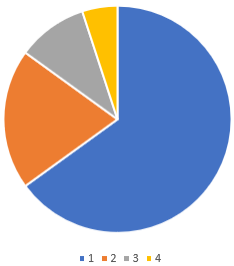


図 9:円グラフ

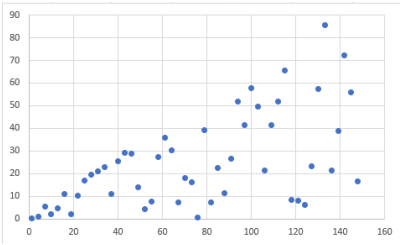


図 10:散布図

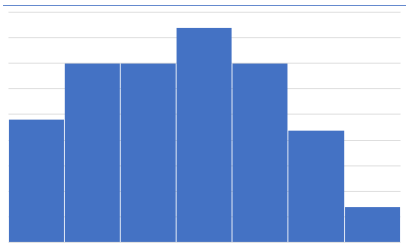


図 11:ヒストグラム

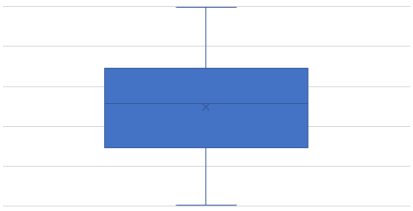


図 12:箱ひげ図

●グラフ描画のための準備

```
import matplotlib.pyplot as plt
```

matplotlib にある pyplot を使用するとグラフを簡単に描くことができる。長いので plt としてインポートしておく。

```
plt.グラフの種類
```

とすることで指定のグラフを描画することができる。

表 1:グラフの種類

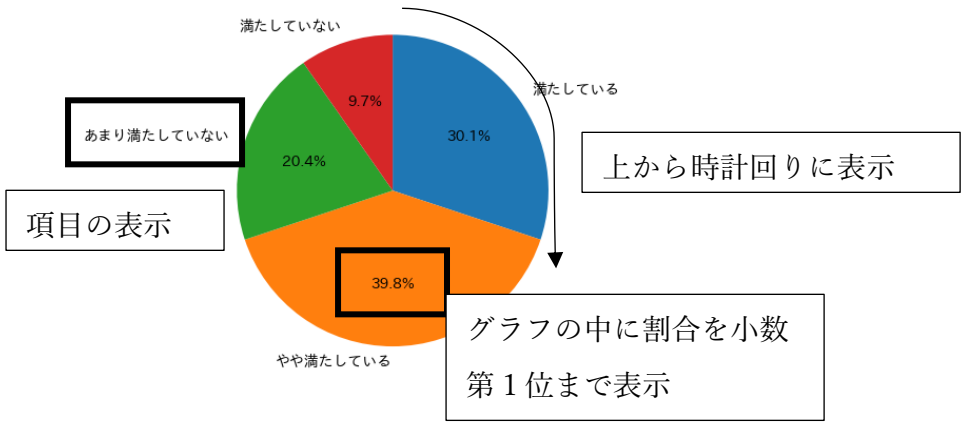
折れ線グラフ	plot
棒グラフ	bar
円グラフ	pie
散布図	scatter
ヒストグラム	hist
箱ひげ図	boxplot

例題 10-3

（1）classroom に配信されたスプレッドシート「グラフの描画」の「折れ線グラフ」のシートにある平均気温のデータを読み込み、折れ線グラフを表示しなさい。
グラフは横軸に「月」、縦軸に「平均気温」を表示し、グラフのタイトルは「平均気温」としなさい。

（2）「棒グラフ」のシートにある降水量の合計のデータを読み込み、棒グラフを表示しなさい。グラフは横軸に「市」、縦軸に「年間降水量(mm)」を表示し、それぞれ横軸と縦軸にラベルを表示しなさい。

（3）「円グラフ」のシートにあるアンケートのデータを読み込み、円グラフを下図のように表示しなさい。



（4）「散布図」のシートにある世帯人員と食料（合計）のデータを読み込み、散布図を作成しなさい。

（5）「ヒストグラム/箱ひげ図」のシートにある偏差値データを読み込み、ヒストグラムと箱ひげ図を作成し、上下に並べて表示しなさい。

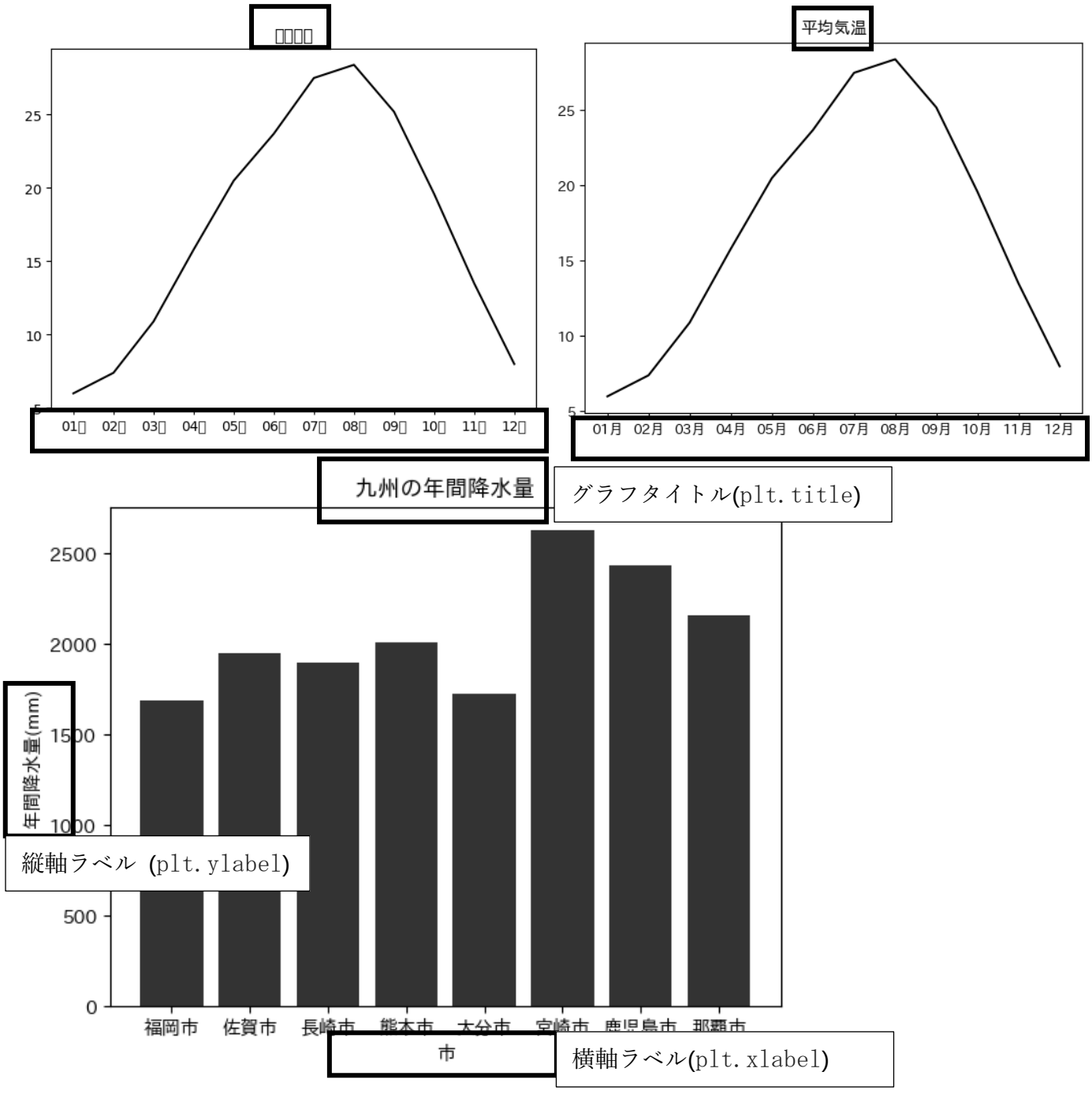
●グラフの書式を変更したり、タイトルを表示したりする方法

表 2

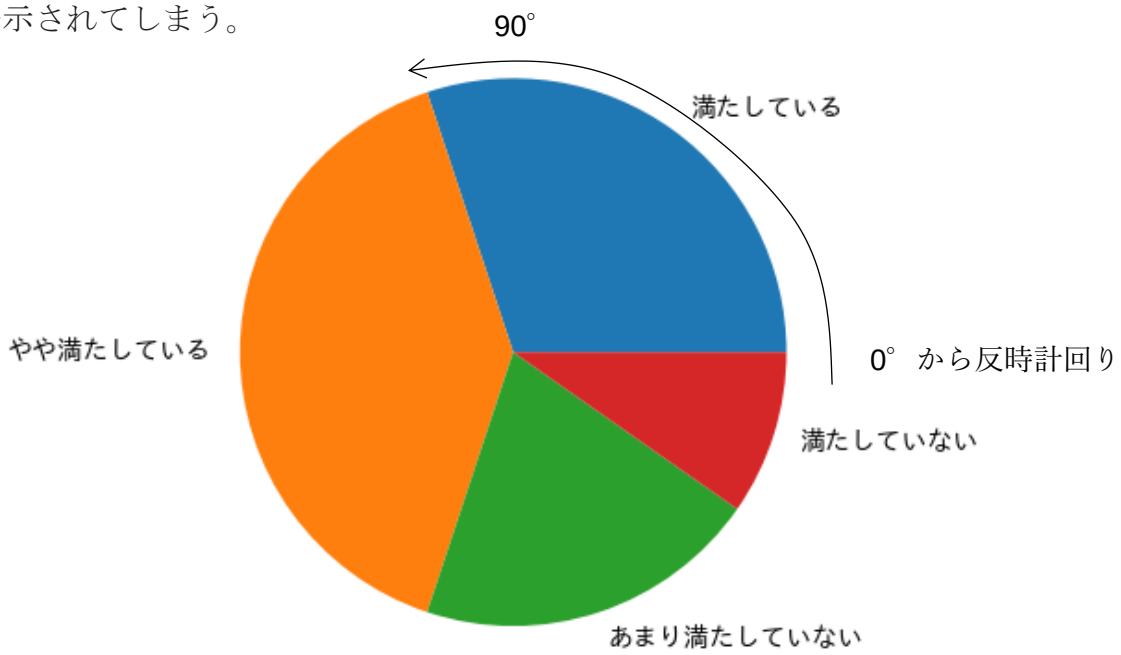
plt.title	グラフのタイトルを表示する
plt.xlabel	x 軸のラベルを表示する
plt.ylabel	y 軸のラベルを表示する

※基本的にタイトルやラベルなどに日本語を使用することはできないが、以下のコードを実行することで日本語を表示することができるようになる

```
# グラフタイトルに日本語を使用したい場合に実行
!pip install japanize-matplotlib
import japanize_matplotlib
```



円グラフは plt.pie(データ)で表示することができるが、デフォルトのままでは下図のように表示されてしまう。



そのため、plt.pie()の中に以下のように記述するとよい。
「counterclock=False, startangle=90」
counterclock=False とすると時計回りに表示することができる。startangle で項目の開始位置を設定できる。
さらに、円の中に割合の数字を入力する場合は「autopct='%.1f%%'」と記述する。
「%.1f%%」の「%.1」は小数点第 1 位まで表示、「f」は浮動小数点数で表記、「%%」は「%」を文字列として表示するという意味である。「%」が特別な意味を持っているので、%を文字列として表示したい場合には 2 つ並べることで「%」を表示させることができる。

ヒストグラムは「plt.hist」で描画することができるが、ヒストグラムの棒を「bin」といい、このビンの数を「bins = 」で指定することができる。デフォルトでは bins = 10 。ビンの数が異なると、ヒストグラムの印象が変わってくるため、適切なビンの数に設定する必要がある。

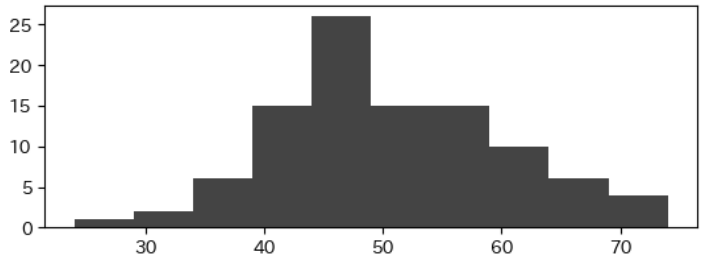


図 13:bins = 10

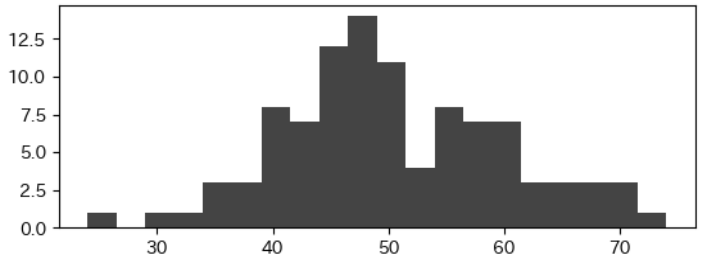
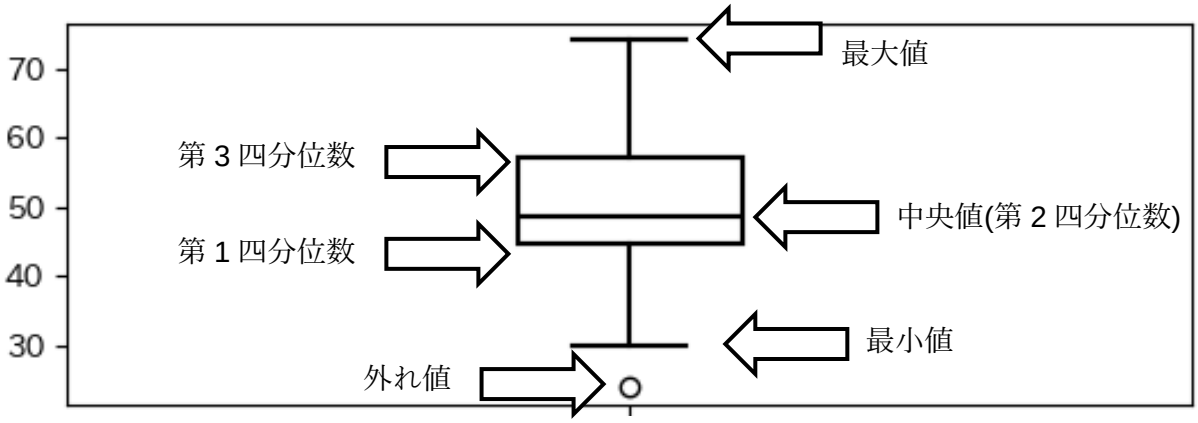
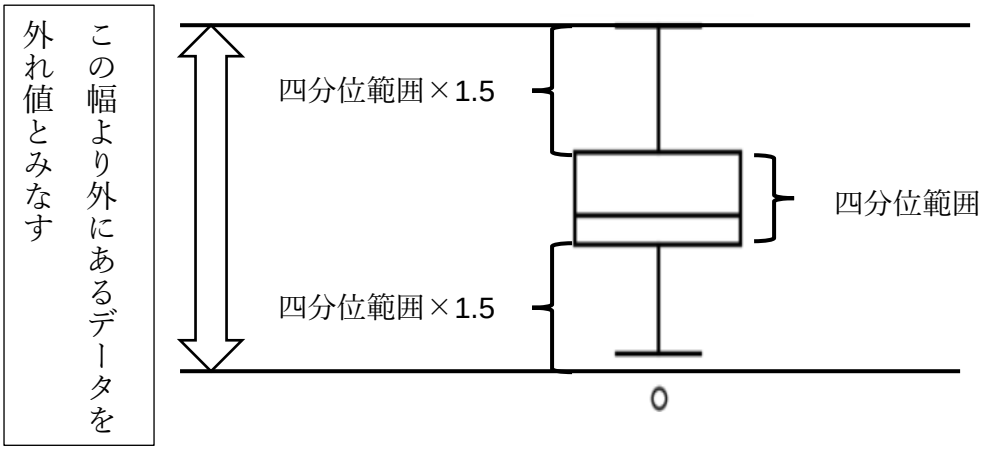


図 14:bins = 20



Python で箱ひげ図を描画すると外れ値（第 1 四分位数と第 3 四分位数のそれぞれから四分位範囲の 1.5 倍より外）の判定も自動で行い、〇で表示される。外れ値を除いた箱ひげ図を描画する場合は「plt.boxplot(ss_data, sym='')」としておくとよい。



演習 10-3

（1）classroom に配信された「10-3_グラフの描画.ipynb」を開き、以下の条件に合うグラフを描画しなさい。

条件：1 以上 100 以下の整数値を 100 個生成し、ビンの数を 20 に設定したヒストグラムを描画する。また、グラフの色を青以外に変更すること。

（2）classroom に配信されたスプレッドシート「グラフの描画」の「演習 10-3」のシートのデータを「10-3_グラフの描画.ipynb」に読み込み、データ 1 とデータ 2 の相関の有無を考察しなさい。

10.9　確率分布

問：2 枚の硬貨を同時に投げるとき、表と裏の出方は何通りあるか

この試行において表が出る硬貨の枚数を X とすると、X の取りうる値は（　　）、（　　）、（　　）であり、X がこれらの値をとる確率は以下の表のようになる

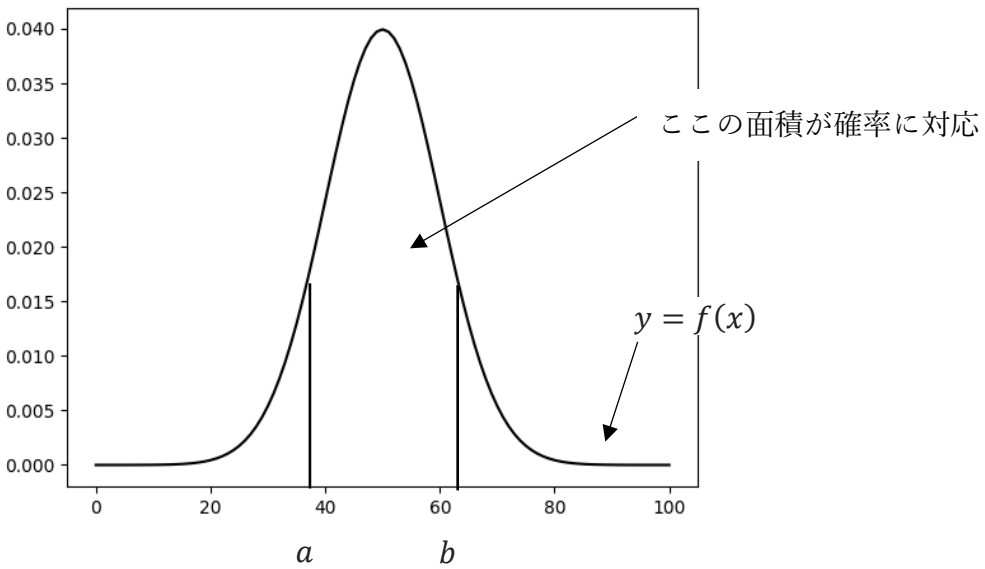
X				計
確率				1

例：2 枚の硬貨を投げる場合の確率変数 X

$P(X=0)=\frac{1}{4}$　$P(X=1)=\frac{2}{4}$　$P(X=2)=\frac{1}{4}$

$P(0 \leqq X \leqq 1)=\frac{1}{4}+\frac{2}{4}=\frac{3}{4}$　$P(1 \leqq X \leqq 2)=\frac{2}{4}+\frac{1}{4}=\frac{3}{4}$

この X のように試行の結果によってその値が定まり、各値に対応して確率が定まるような変数を確率変数という。
この確率変数 X の取りうる値とその確率を対応させたものを X の確率分布という。
サイコロ 1 個を振る試行では 1 から 6 の値をとるが、この 1 から 6 の値はとびとびの値（離散的）であり、このときの確率分布を**離散型確率分布**という。一方で、※身長や体重は小数点以下がずっと続くような値（連続的）であり、このときの確率分布を連続型確率分布という。この場合、値に幅を持たせた確率を考えることにする。
※とても精度の高い測定器を用いると理論上どこまでも値が続くので連続型として扱う



X に 1 つの曲線$y=f(x)$を対応させ、$a \leqq X \leqq b$となる確率$P(a \leqq X \leqq b)$が上の図の斜線部の面積で表されるようにする。このとき、$f(x)$を**確率密度関数**という。
通常このような関数の面積を手計算で求めるには、数学Ⅱ・Ⅲで学習する積分という手法を用いるが、この授業では Python で求める方法を紹介する。

10.10　二項分布

復習問題（反復試行の確率）

確率 $\frac{1}{2}$ で表が出て、確率 $\frac{1}{2}$ で裏が出るコインについて考える。このコインを 5 回投げたとき、表が 3 回出る確率を求めなさい。

確率 p の試行を n 回行ったとき、ちょうど k 回成功する確率は以下の式で求められる。

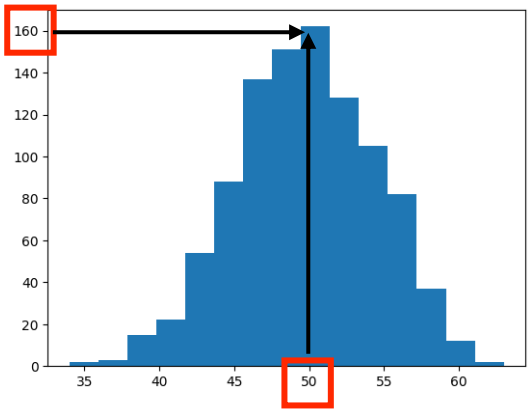
$$P(X=k) = {}_nC_k p^k (1-p)^{n-k}$$

この $P(X=k)$ は $P(X=0)$ から $P(X=n)$ まで存在（0 回成功する確率から n 回成功する確率まで存在）し、確率変数 X の確率分布は以下の表のようになる。

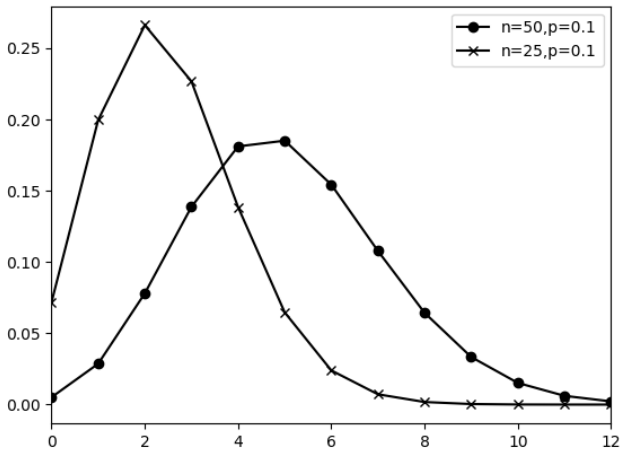
X	0	1	…	k	…	n	計
P	${}_nC_0(1-p)^n$	${}_nC_1 p(1-p)^{n-1}$		${}_nC_k p^k (1-p)^{n-k}$		${}_nC_n p^n$	1

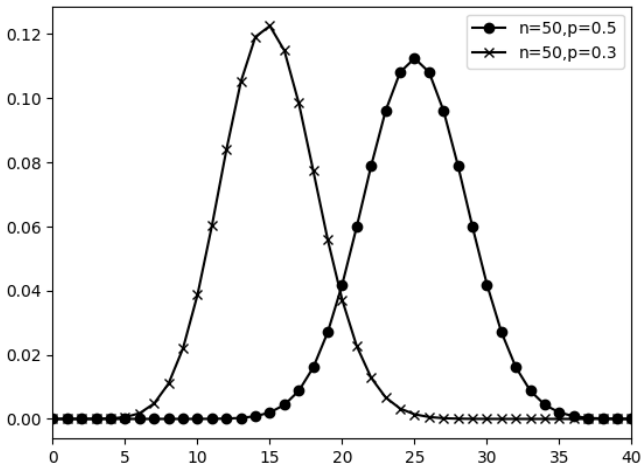
この表によって与えられる分布を二項分布(Binomial Distribution)といい、$B(n,p)$ で表す
また、このとき確率変数 X は二項分布に従うという。

表と裏が $\frac{1}{2}$ の確率でそれぞれ出現するコインを 100 回投げ、表が出た回数をカウントする操作を 1000 回繰り返した結果をヒストグラムで描画したのが右図である。表が出た回数が、100 回中 50 回付近がもっとも多く、50 から離れるほど出現回数が少なくなっている。



二項分布は試行回数 n と確率 p によって分布が異なる。その比較が下図である。左は確率を固定し、試行回数を変化させたものであり、右は試行回数を固定し、確率を変化させたものである。





10.11　正規分布

連続型確率変数の分布の代表的なものに正規分布（normal distribution）がある。世の中の様々なところで、正規分布で近似できる分布が出現する。例えば日本人の身長の分布（ただし、男女が混じっているような分布は正規分布とならない）は正規分布で近似できることが知られてる。また、正規分布は扱いやすく、研究も進んでいるため、統計学を学ぶうえで最も重要な分布である。正規分布の確率密度関数は次のように表される。

$$f(x)=\frac{1}{\sqrt{2\pi\sigma^2}}e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$　（μ は平均、σ は標準偏差）※式は覚えなくてよい

式は複雑だが、大事なのは正規分布の確率密度関数は平均μと標準偏差 σ の 2 つで定まるという点である。つまり、平均と標準偏差の 2 つで正規分布がどのような形を表すのかが決まるということである。この正規分布を$N(\mu,\sigma^2)$と表し、特に平均 0、標準偏差 1 の正規分布を（　　）という。

※ e はネイピア数と呼ばれる定数（2.718…となる無理数）数Ⅲで学習

例題 10-4

平均 50、標準偏差 10 の正規分布のグラフを描画しなさい。

```
# 例題 10-4
# 準備
import numpy as np
import scipy.stats as stats
import matplotlib.pyplot as plt

# 平均値と標準偏差を定義
mu = 50     #平均値
sigma = 10   #標準偏差

# 等間隔に点を生成する
#0～100 の間で等間隔な 101 個の点を生成する
x = np.linspace(0, 100, 101)

# 正規分布の確率密度関数に x の値を代入
y = stats.norm.pdf(x, mu, sigma)

# グラフを描画
plt.plot(x, y)
```

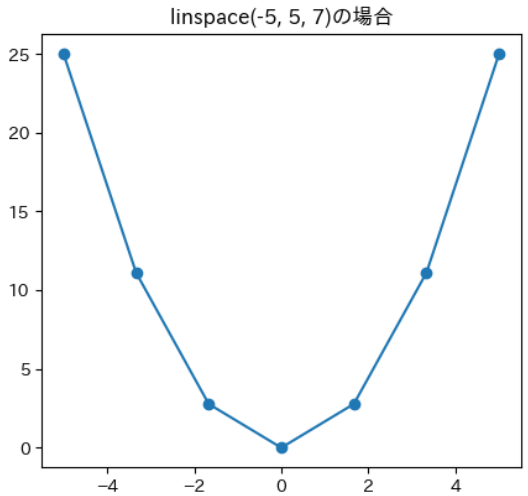
今回は 0～100 の間に等間隔の値を 101 個生成し x に代入したので、x の中身を見てみると

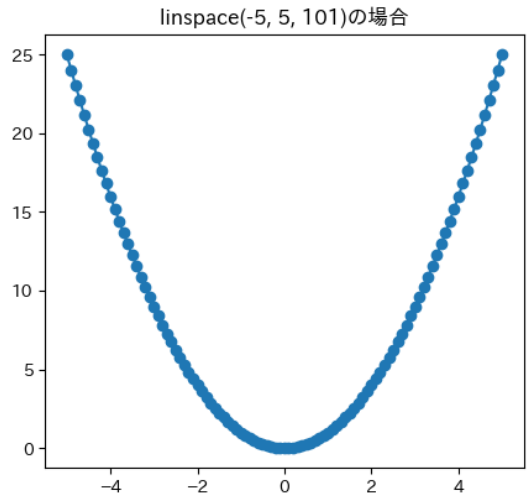
```
print(x)
[  0.   1.   2.   3.   4.   5.   6.   7.   8.   9.  10.  11.  12.  13.
  14.  15.  16.  17.  18.  19.  20.  21.  22.  23.  24.  25.  26.  27.
  28.  29.  30.  31.  32.  33.  34.  35.  36.  37.  38.  39.  40.  41.
  42.  43.  44.  45.  46.  47.  48.  49.  50.  51.  52.  53.  54.  55.
  56.  57.  58.  59.  60.  61.  62.  63.  64.  65.  66.  67.  68.  69.
  70.  71.  72.  73.  74.  75.  76.  77.  78.  79.  80.  81.  82.  83.
  84.  85.  86.  87.  88.  89.  90.  91.  92.  93.  94.  95.  96.  97.
  98.  99. 100.]
```

となり、x に 0 から 100 の数値を代入することができる。
また、norm.pdf(a, loc=0, scale=1)で平均 0(loc=0)、標準偏差 1(scale=1)の正規分布の確率密度関数$f(x)$ にa を代入した値$f(a)$ を求めることができる。pdf とは(Probability density function：確率密度関数)である。ここで例題 10-4 の y の中身を見てみると

```
print(y)
[1.48671951e-07 2.43896075e-07 3.96129909e-07 6.36982518e-07
 1.01408521e-06 1.59837411e-06 2.49424713e-06 3.85351967e-06
 5.89430678e-06 8.92616572e-06 1.33830226e-05 1.98655471e-05
                              ≈
 1.01408521e-06 6.36982518e-07 3.96129909e-07 2.43896075e-07
 1.48671951e-07]

※一部データを省略して表示している
```

この y の波線部は$x = 0$を$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$に代入した値、つまり$f(0)$である。
x と y には以下のような値が格納されている。
x = [0. 1. 2. 3. 4. 5. …]
y = [$f(0)$ $f(1)$ $f(2)$ $f(3)$ $f(4)$ $f(5)$ …]
この x と y を plt.plot(x, y)で座標平面上に点を取ることで、正規分布を描画することができる。plt.plot で 2 次関数のグラフや正規分布を描画する場合には x 軸の値を細かくとることで、曲線を表現することができる。以下の図は点の個数の違いによるグラフの見え方である。





演習 10-4
平均 50、標準偏差 10 の正規分布と平均 40、標準偏差 11 の正規分布の 2 つを描画し、平均と標準偏差の値で正規分布が定まることを確認しなさい。

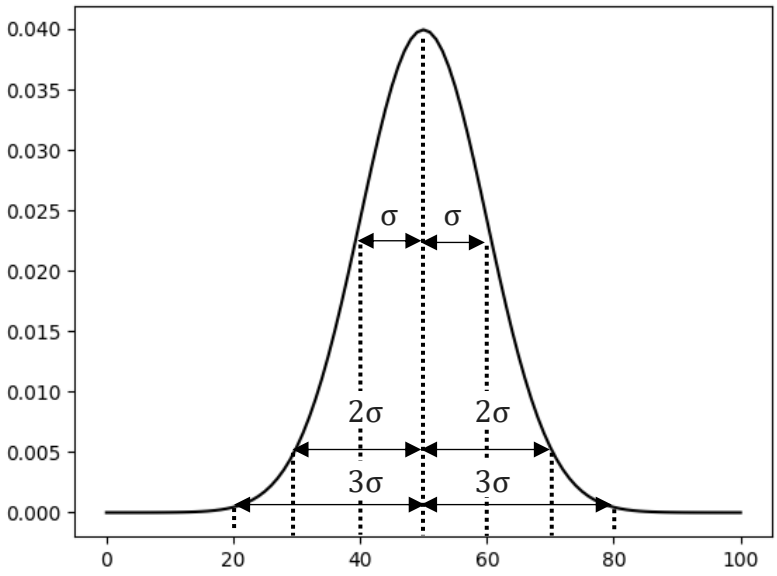
●正規分布の特徴

・平均 μ を中心とした釣り鐘型で、左右対称の分布

・$\mu-\sigma$から$\mu+\sigma$までの範囲の起こる確率が約（　　）％

・$\mu-2\sigma$から$\mu+2\sigma$までの範囲の起こる確率が約（　　）％

・$\mu-3\sigma$から$\mu+3\sigma$までの範囲の起こる確率が約（　　）％



図：平均 50、標準偏差 10 の正規分布

例えば平均$\mu = 167.6$(cm)、標準偏差$\sigma = 7.0$に従う正規分布（$N(167.6, 7.0^2)$)からランダムに1000 人抽出した場合、$\mu-\sigma = 160.6$から$\mu+\sigma = 174.6$の範囲に約 68.3%(683 人)、$\mu-2\sigma = 153.6$から$\mu+2\sigma = 181.6$の範囲に約 95.4%(954 人)、$\mu-3\sigma = 146.6$から$\mu+3\sigma = 188.6$の範囲に約 99.7%(997 人)が存在することになる。

●標準化

正規分布は上のように特徴が研究されておりとても扱いやすい。上のことから身長 182 cm以上の人は、この分布上では上位 5%に入ることが分かる。

例題 10-5

（1）国語の得点が 50 点で数学の得点が 40 点の生徒がいる。国語も数学も平均 40、標準偏差 10 の正規分布($N(40, 10^2)$)に従うとするとどちらの教科の方が、得点がいいと言えるだろうか。

（2）国語の得点が 50 点で数学の得点が 40 点の生徒がいる。国語は平均 60、標準偏差 10 の正規分布($N(60, 10^2)$)に、数学は平均 50、標準偏差 20 の正規分布($N(50, 20^2)$)に従うとするとどちらの教科の方が、得点がいいと言えるだろうか。

問題なのは異なる分布からサンプルを取り出しているので比較が難しいということ。これを解消するために分布を一つにしてしまおうという考え方が（　　）である。

$$z = \frac{(x-\mu)}{\sigma}$$

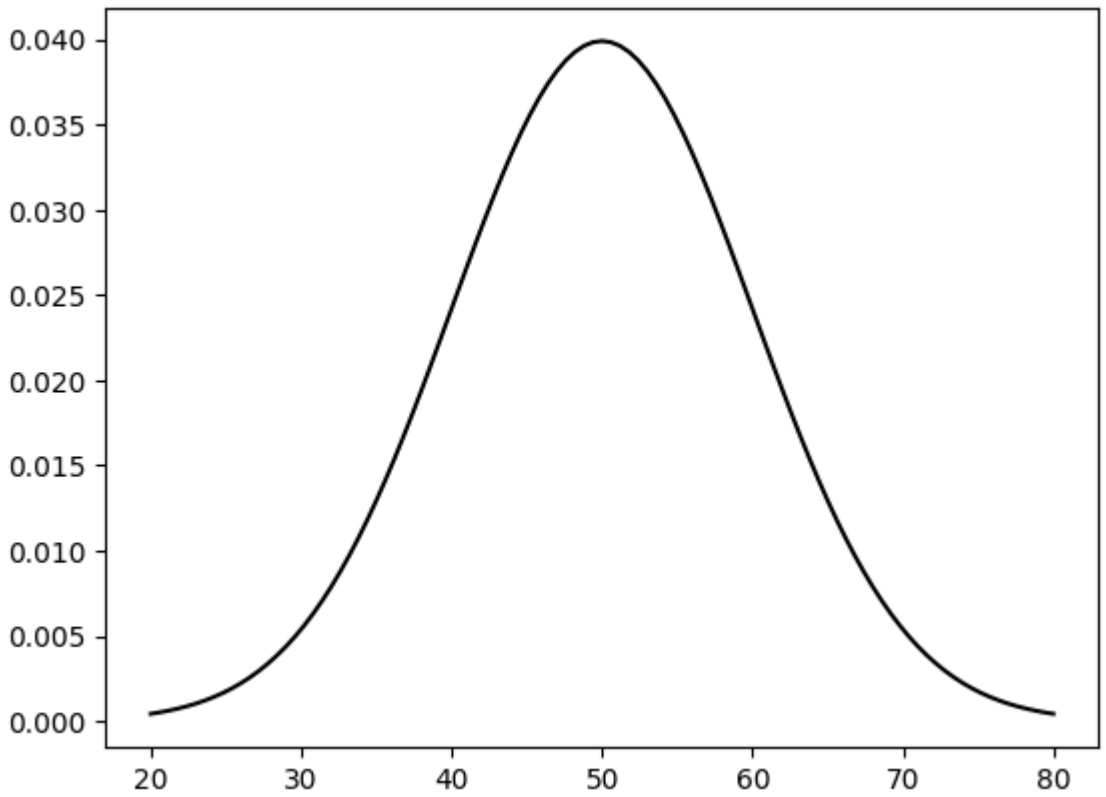
この z で変換をするとすべての正規分布を標準正規分布$(N(0,1^2))$に変換できる。この z 値を比較することでどちらが優れた点数なのか比べることが可能になる。さらに標準正規分布表を用いると上位何%に当たるのかも求められる。詳しい計算等は数学 B で学習。

演習 10-5

例題 10-5（２）の条件のもと、以下の式で求められる値を計算したところ、国語は 40、数学は 45 という値を得た。このとき国語と数学のどちらが集団の中で上位にいるか。

$$\frac{(x-\mu)}{\sigma} \times 10 + 50$$

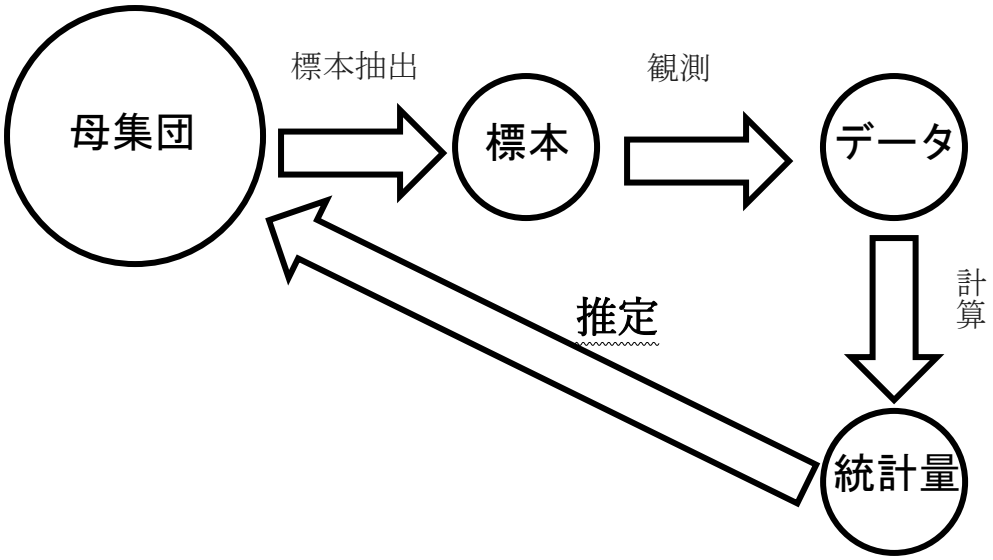
※平均 50、標準偏差 10 になるように標準化したもの→（　　）



11.　推定

11.1　推定とは

推定とは標本から母集団の特徴（平均など）を推測する方法であり、下図のような流れで実施する。また、推定には**点推定**と**区間推定**がある。点推定とは母集団の特徴（平均値など）を 1 つの値で推測することで、区間推定はある程度幅を持たせて推測すること。



母集団の特性を「母〇〇」と呼び、小文字のギリシャ文字（μ や σ など）で表す。例えば母集団の平均を**母平均**と呼び、 μ で表す。同様に、母集団の分散である**母分散**は σ^2、母集団の標準偏差である**母標準偏差**は σ で表す。一方で、標本から得られる統計量を「標本〇〇」と呼び、小文字のアルファベットで表す。標本の平均である**標本平均**は $\bar{x}$ 、標本の分散である**標本分散**は S^2 、標本の標準偏差である**標本標準偏差**は S で表す。

11.2　大数の法則

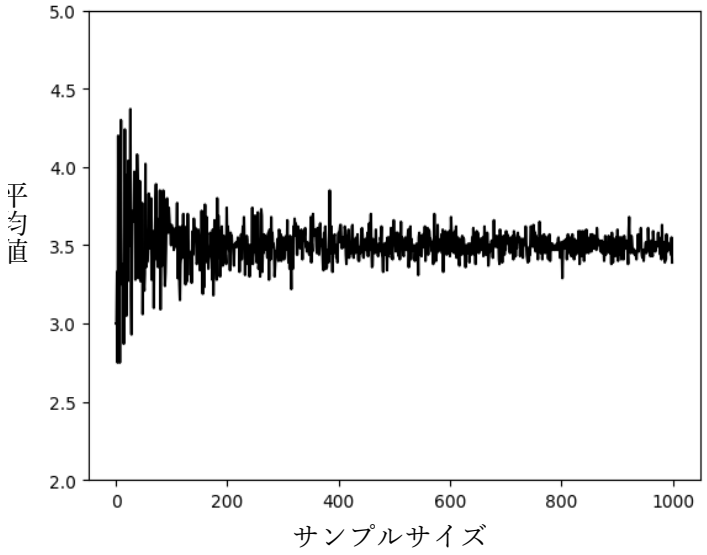
問い

サイコロを n 回振り、平均値を求める操作を 1000 回繰り返す。このとき、n を 1、10、100、…と大きくしていくと平均値はどう変化していくか。

重要

この n を**標本の大きさ（サンプルサイズ）**と呼ぶ。このサンプルサイズに似た言葉に、サンプル数というものがあるが、これは標本を抽出する操作を何回行ったのかという意味であるため、しっかりと区別する必要がある

右図を見ると、サンプルサイズが大きくなっていくにつれて、サイコロの出た目の平均値が 3.5 に近づいていることが分かる。つまり、サンプルサイズ 1 のときの出た目の平均値と、サンプルサイズ 100 のときの出た目の平均値を比べると、サンプルサイズが大きい方が、真の値に近づいていくのである。これを説明した統計学の定理が**大数の法則（たいすうのほうそく）**である。



図：サンプルサイズ n の標本抽出を 1000 回繰り返したシミュレーション

大数の法則（重要）

母平均 μ の母集団から標本を抽出する場合、サンプルサイズが大きくなるほど標本平均 $\bar{x}$ は母平均 μ に近づく。

11.3　点推定

母集団の特性が分からない状態であるので、それらを抽出した標本から推測しようと考える。しかし、標本はあくまで母集団から抽出した一部であるから真の値に完全には一致せず、真の値から大きい値か小さい値の方向へばらつきが生じる。平均値は、このばらつきが、大きい値だけに偏っていたり、小さい方だけに偏っていたりすることはないという性質を持つ。これを**不偏性**という。さらに大数の法則によりサンプルサイズが大きくなれば標本平均は母平均に近づくという性質を持つため、**標本平均を母平均とみなして推定する**ことができる（標本平均は母平均の推定量であるという）。一方で、母分散を推定する場合はそのまま標本分散を用いることはできない。標本分散は母分散よりも小さい方に偏っているため不偏性を持たないからである（標本分散は母分散の不偏推定量ではないという言い方をする）。そのため、標本分散の小さい方への偏りを修正するために不偏分散を用いて母分散を推定する。

$$\text{標本分散：} S^2 = \frac{(x_1-\bar{x})^2+(x_2-\bar{x})^2+\cdots+(x_n-\bar{x})^2}{n}$$

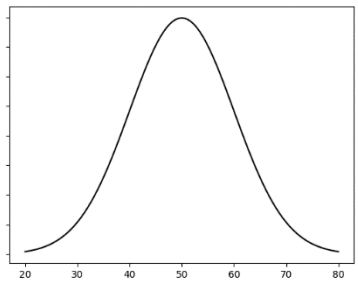
$$\text{不偏分散：} s^2 = \frac{(x_1-\bar{x})^2+(x_2-\bar{x})^2+\cdots+(x_n-\bar{x})^2}{n-1}$$

※標本分散の分母 n を $n-1$ に置き換えて過小評価を修正したものが不偏分散

平均 50、標準偏差 10 の正規分布に従う乱数を 15 個発生させる。得られた乱数の標本平均、標本分散、標本標準偏差を求める操作を 1000 回繰り返した平均が以下の通りである。

(標本平均)=49.9　　(標本分散)=91.8　　(標本標準偏差)=9.4

標本平均は母平均 50 に近いが、標本分散は母分散 100、標本標準偏差は母標準偏差 10 よりも過小評価されていることがわかる。



15 個抽出　標本平均、標本分散、標本標準偏差を求める

1000 回繰り返す

例題 11-1

（1）classroom に配信された「11-1_点推定.ipynb」を開き、例題 11-1 を実行しなさい。さらに、得られた標本平均、標本分散、標本標準偏差の値から母平均、母分散、母標準偏差を推定しなさい。

（2）（1）のデータの母平均、母分散、母標準偏差を表示し、（1）で推定した値との差を考えなさい。

※Python で求める必要はない。どれだけの精度で推定できているか考えること

（3）例題 11-1 のファイルを修正し、標本平均、不偏分散、*不偏標準偏差を求め、（1）で推定した値とどちらが高い精度で推定できているか考えなさい。

注意！

Python で不偏標準偏差を求める場合、不偏分散の平方根で計算を行うが、**不偏分散の平方根は母標準偏差の不偏推定量ではない。（若干小さい値が得られる）**ただし、授業のレベルを大きく上回るので、不偏分散の平方根を母標準偏差の推定量として話を進める。

演習 11-1

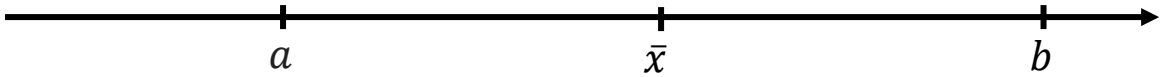
（1）平均 50、標準偏差 10 の正規分布に従う乱数を 15 個発生させ（サンプルサイズ $n = 15$ で標本を抽出し）、変数 sample に代入し sample の中身を表示しなさい。

（2）（1）で得られた sample の標本平均、不偏分散、不偏分散の平方根を求め、表示しなさい。

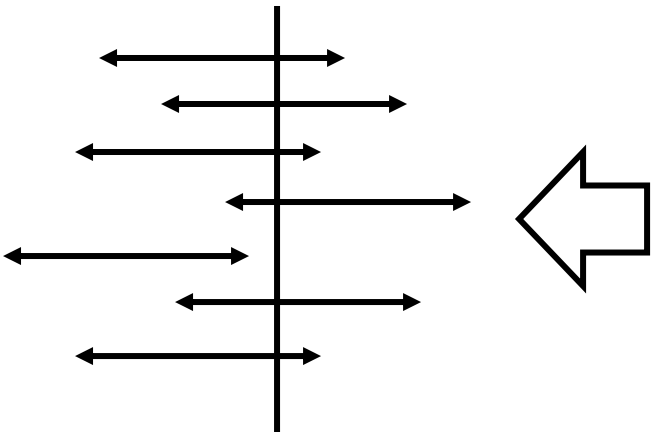
（3）（2）で得られた推定値から母平均、母分散、母標準偏差を推定しなさい。

11.4　区間推定

実際の調査では例題 11-1 のように標本抽出を 1000 回繰り返すことはほとんど不可能であり、演習 11-1 のように標本抽出を１回で母平均などを推定する必要がある。しかし、標本の取り方によっては母平均から離れた推定になることもある。このように母平均から標本の取り方によってばらつきが生じることを想定して、推定に幅をもたせる考え方が**区間推定**である。また、幅を持たせて得られた区間を**信頼区間**という。この信頼区間も標本から得られた値から推定したものであるから、標本を抽出するごとに信頼区間にばらつきが生じる。



上図の例で考えると、母平均 μ の９５％信頼区間は $a \leqq \bar{x} \leqq b$ という。この９５％信頼区間の意味は、標本抽出して標本平均を計算、信頼区間を求めるという操作を１００回繰り返したときに、９５回はその区間内に母平均を含むという意味である。この９５％に根拠はとくに存在せず、９０％だったり、９９％だったりする。９５％信頼区間がよく使われているので、この授業でも９５％信頼区間を求める演習を行っていく。



図：９５％信頼区間の意味

この母平均の区間推定には以下のパターンが存在する。

パターン１

正規母集団（母集団は正規分布に従う）で母分散既知（母分散の値が分かっている）

パターン２

正規母集団で母分散未知（母分散の値が分かっていない）

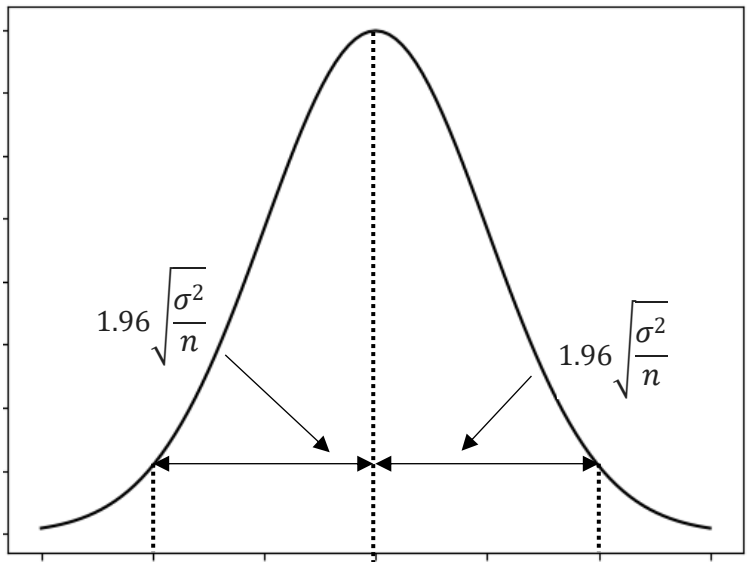
実際には母集団が正規分布とみなすことができないパターンもあるが、この授業では取り扱わない

パターン１：正規母集団で母分散既知の場合の母平均の区間推定

現実的には母平均が未知の状況で、母分散が既知の状況は考えられないが、パターン１が基本となり、パターン２とパターン３の区間推定を行っていくため、無駄なことではない。

正規分布($N(\mu,\sigma^2)$)に従う母集団から抽出した**標本平均の分布は $N\left(\mu,\frac{\sigma^2}{n}\right)$ の正規分布に従う**ことが分かっている。ここで重要なのは、標本分散がサンプルサイズ n によって変化するということである。

テキストP17（正規分布の特徴）では標準偏差の２倍(2σ)、標準偏差の３倍(3σ)にそれぞれ９５％、９９％が収まるとしているが、厳密には**標準偏差の 1.96 倍(1.96σ)、標準偏差の 2.58 倍(2.58σ)に９５％、９９％が収まる**。



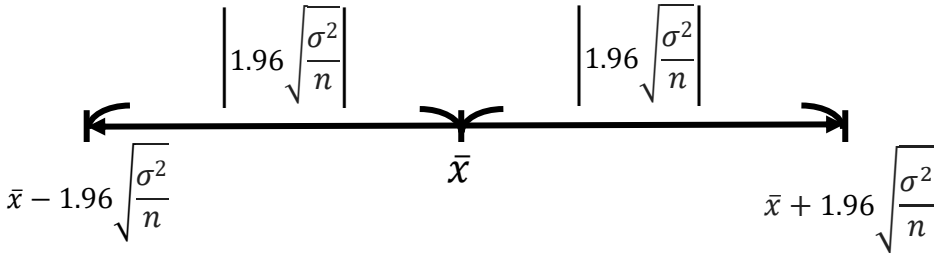
図：標本平均の分布

したがって、９５％信頼区間は標本平均から 1.96 倍の標本平均の標準偏差**（標準誤差）**で求めることができる。

標準誤差は $\sqrt{\frac{\sigma^2}{n}}$ $\left(\sigma^2\text{は母分散、}n\text{はサンプルサイズ}\right)$ であるので、９５％信頼区間は以下のようになる。

正規母集団で母分散既知($\boldsymbol{\sigma^2}$)の母平均μの９５％信頼区間（数Bでも学習する）

$$\bar{x}-1.96\sqrt{\frac{\sigma^2}{n}} \leqq \mu \leqq \bar{x}+1.96\sqrt{\frac{\sigma^2}{n}} \left(\bar{x}\text{は標本平均、}n\text{はサンプルサイズ}\right)$$



※９９％信頼区間の場合は 1.96 が 2.58 に変わる

例題 11-2

（１）母分散$\sigma^2 = 100$の正規母集団から$n = 100$の標本を抽出し、標本平均 $\bar{x}$ を求めたところ $\bar{x} = 50.6$ であった。このとき、母平均 μ の９５％信頼区間を求めなさい。

$\leqq \mu \leqq$

正規母集団で母分散既知の９５％信頼区間

scipy.stats.norm.interval(confidence, loc, scale)

confidence:信頼度（信頼度９５％のとき、0.95）
loc:標本平均 $\bar{x}$
scale:標準誤差$\sqrt{\frac{\sigma^2}{n}}$

（２）正規分布に従う母分散$\sigma^2 = 100$からサンプルサイズ $n = 10$ で標本を抽出し、９５％信頼区間を求める操作を 20 回繰り返した。母平均 $\mu = 50$ を含んでいる数を求めなさい。

20 個中　　個

演習 11-2

（１）演習 11-2(1)のセルには例題 11-2(1)と同じプログラムが入力されている。プログラムを１か所修正し９９％信頼区間を求め、９５％信頼区間の幅との違いを考えなさい。

$\leqq \mu \leqq$

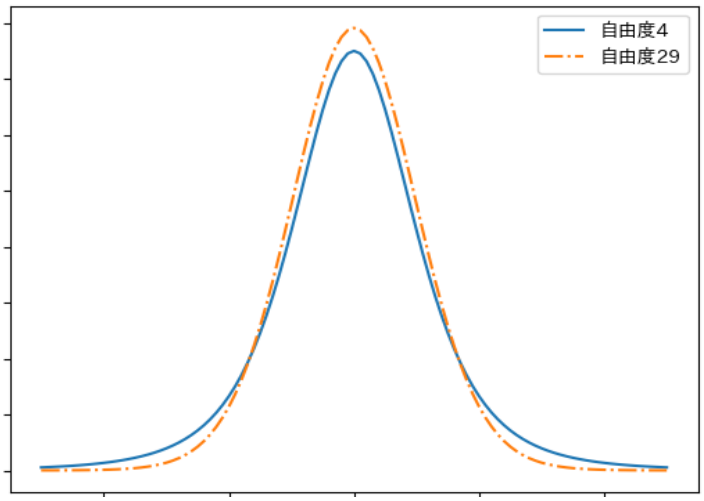
（２）classroom に配信されたスプレッドシート「区間推定」の「母分散既知」のシートには正規母集団に従う母集団から、サンプルサイズ $n = 20$ で標本を抽出したデータが入力されている。このデータから母平均 μ の９５％信頼区間を求めなさい。ただし、母分散 $\sigma^2 = 100$ とする。

$\leqq \mu \leqq$

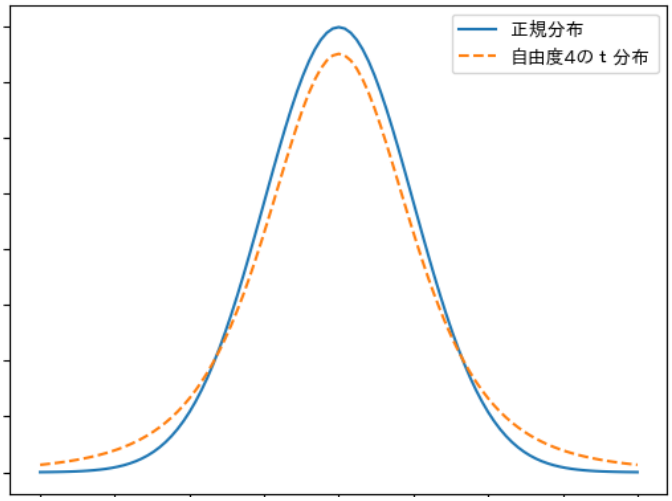
（３）信頼度は９５％のままで信頼区間の幅を半分にするためにはどのようにすればよいか考察しなさい。

パターン２：正規母集団で母分散未知の場合の区間推定

母分散既知の場合（パターン１）、正規分布$(N(\mu,\sigma^2))$に従う母集団から抽出した標本平均の分布は $N\left(\mu,\frac{\sigma^2}{n}\right)$ に従うことが分かっていたが、母分散 σ^2 が使えず不偏分散 s^2 で代用しなければいけない代償として、標本平均の分布は正規分布に似た t 分布という分布に従う。



図： t 分布



図： t 分布と正規分布の比較

ここで自由度とはサンプルサイズ−1 のことである。

t 分布まとめ
・t 分布は自由度で形が変わる分布
・自由度が十分に大きいとき（サンプルサイズ$n \geqq 30$）、 t 分布は正規分布で近似できる
・正規母集団で母分散未知の母集団から抽出した標本平均が従う分布

⇨
・サンプルサイズが 30 以上（大標本）であれば不偏分散を母分散とみなして、パターン１（母分散既知）で区間推定が可能
・サンプルサイズが小さい場合（小標本）であれば t 分布で区間推定

正規母集団で母分散未知の９５％信頼区間
scipy.stats.t.interval(confidence, df, loc, scale)

confidence:信頼度（信頼度９５％のとき、0.95）
df:自由度（サンプルサイズ−1）
loc:標本平均 $\bar{x}$
scale:標準誤差 $\sqrt{\frac{s^2}{n}}$ (s^2は不偏分散)

例題 11-3

（1）正規母集団からサンプルサイズ $n = 30$ で標本を抽出した。抽出した標本から母平均 μ の９５％信頼区間を求めなさい。ただし、母分散は分からないものとする。

$\leqq \mu \leqq$

（2）正規母集団からサンプルサイズ $n = 5$ で標本を抽出した。抽出した標本から母平均 μ の９５％信頼区間を求めなさい。ただし、母分散は分からないものとする。

$\leqq \mu \leqq$

（3）正規母集団からサンプルサイズ $n = 5$ と $n = 30$ で標本を抽出した。抽出した標本から母平均 μ の９５％信頼区間を求めなさい。また、不偏分散を母分散とみなしたとき($\sigma^2 = s^2$としたとき)、母平均 μ の９５％信頼区間と比較しなさい。

サンプルサイズ$n = 5$

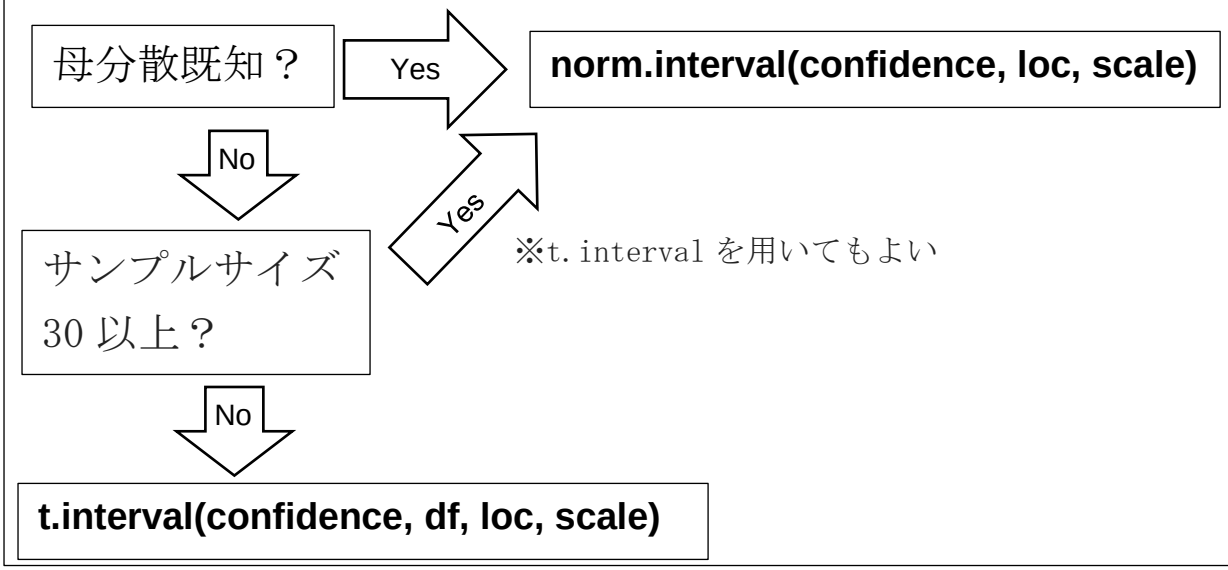
不偏分散を母分散とみなした場合	$\leqq \mu \leqq$
t 分布に従うと仮定した場合	$\leqq \mu \leqq$

サンプルサイズ$n = 30$

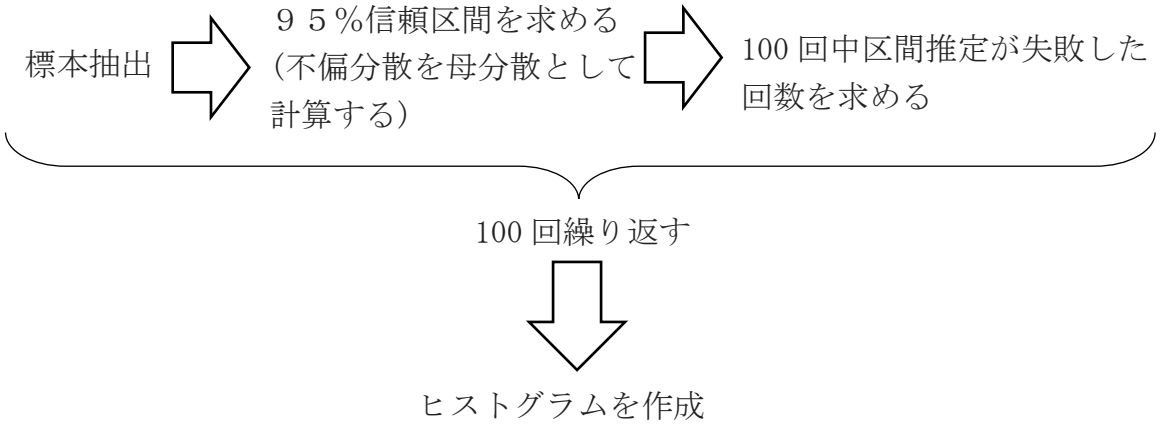
不偏分散を母分散とみなした場合	$\leqq \mu \leqq$
t 分布に従うと仮定した場合	$\leqq \mu \leqq$

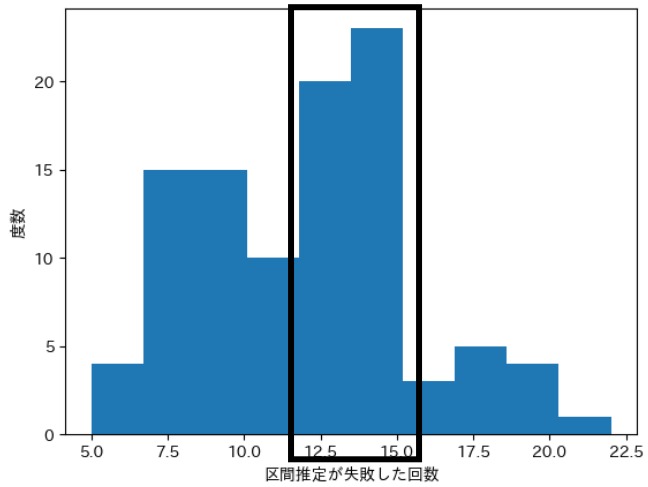
以上の結果から小標本の場合に不偏分散を母分散とみなしてパターン１のように区間推定することは区間幅に差が大きいため、信頼できる結果を得ることができない。

区間推定のパターンまとめ（正規母集団から抽出）

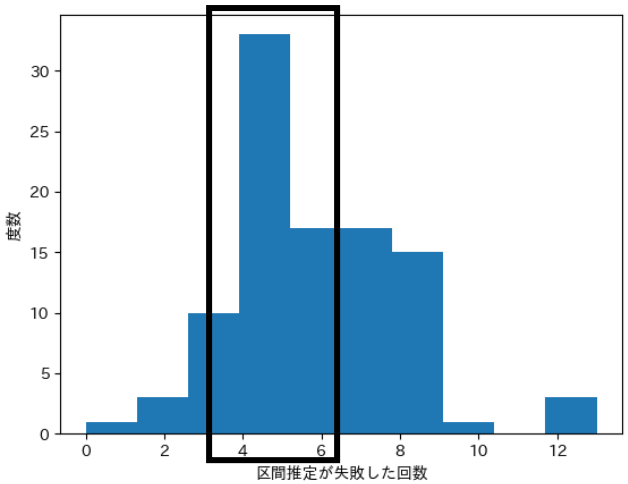


以下の図は標本抽出から９５％信頼区間を求める操作を 100 回繰り返し、そのうち区間推定が失敗した回数を計算する操作を 100 回繰り返し、ヒストグラムにしたものである。





図：小標本$(n = 5)$のとき



図：大標本$(n = 30)$のとき

９５％信頼区間を求めたので、理論上 100 回中 5 回程度は区間推定が失敗する（母平均を捉えない）ことがある。大標本（右図）の場合は理論値付近がよく起こっているのに対して、小標本（左図）の場合、区間推定が理論値よりも多く失敗していることが分かる。したがって小標本の場合、不偏分散を母分散とみなして norm.interval で区間推定を行っても、正しく信頼区間を得ることができないということを示している。

演習 11-3

classroom に配信されたスプレッドシート「区間推定」の「母分散未知」のシートには正規母集団から、サンプルサイズ $n = 5$ で標本を抽出したデータが入力されている。このデータから母平均 μ の９５％信頼区間を求めなさい。

11.5　二項分布の正規分布近似

問 1

表か裏が出るコインがある。このコインが不正なコインであるかどうか知りたい。不正なコインでないとすると、表と裏はそれぞれ 50%の確率で起こるものとする。

このコインを 4 回投げたときすべて表がでた。この状況からこのコインは不正なコインであると言えるか。下の選択肢に〇をつけ周囲の人と意見交換をしてみましょう。

不正なコインである・不正なコインではない

問 2

表と裏が出るコインがある。このコインが不正なコインであるかどうか知りたい。不正なコインでないとすると、表と裏はそれぞれ 50%の確率で起こるものとする。

このコインを 100 回投げたとき 61 回表がでた。この状況からこのコインは不正なコインであると言えるか。

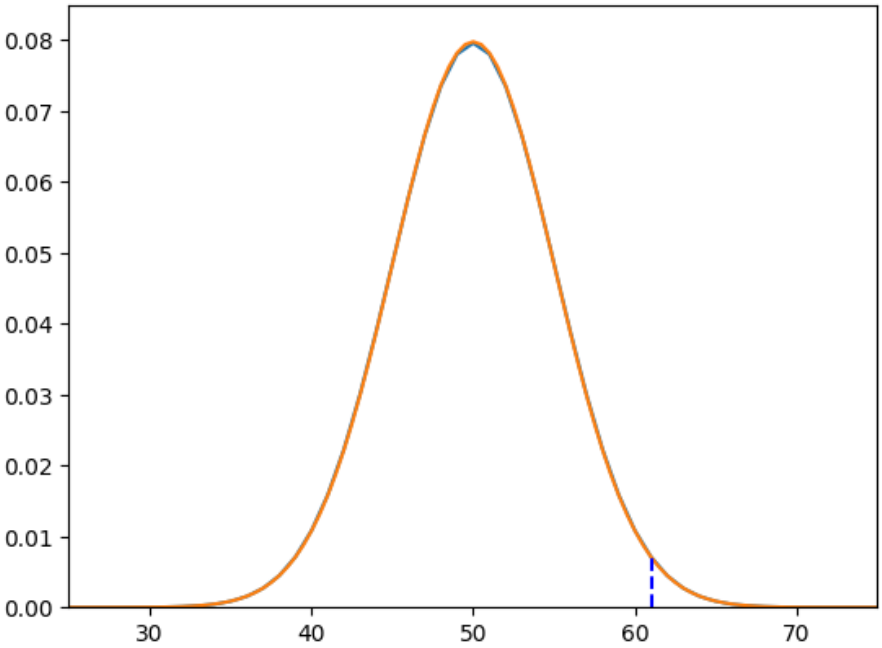
問 1 の状況であれば$\left(\frac{1}{2}\right)^4 = \frac{1}{16}$という計算を行うだけであるが、問 2 の状況であると

$$_{100}C_{61}\left(\frac{1}{2}\right)^{61}\left(\frac{1}{2}\right)^{39}$$

を計算して確率を計算しなければならない。非常に面倒な計算が必要。このときに活躍するのが、二項分布の正規分布近似である。

確率 p で成功する試行を n 回行ったときの成功回数 X が従う分布が二項分布$(B(n,p))$であった。

下図は $B(n,p)$ と$N(np,np(1-p))$ を重ねて描画した図である。二項分布と正規分布がほとんど重なっていることが分かる。つまり、今回の状況では、二項分布で計算することと、正規分布で計算することに差がない（二項分布を正規分布で近似できる）ということである。



二項分布の正規分布近似
二項分布 $B(n,p)$ は試行回数 n が十分大きければ、正規分布 $N(np,np(1-p))$ で近似できる。

試行回数 n が十分大きければとあるが、n が$np > 5, np(1-p) > 5$ を満たすとき、正規分布で近似できるとされている。

Python では以下のように入力すれば確率を簡単に求めることができる

二項分布の確率計算
確率 p で成功する試行を n 回繰り返したときに、X 回成功する確率
`scipy.stats.binom.pmf(X, n, p)`

例：確率 0.5 で表が出るコインを 100 回投げたときに、61 回表がでる確率
`scipy.stats.binom.pmf(61, 100, 0.5)`
実行結果：0.0071107322699265375

正規分布で計算した場合
`scipy.stats.norm.pdf(61, 100 * 0.5, numpy.sqrt(100 * 0.5 * (1 - 0.5)))`
実行結果：0.007094918569246285

11.6　仮説検定（母分散既知の仮説検定）

仮説検定（単に検定ともいう）は確率論（これまで学習した正規分布の性質等）を用いて結論を導く方法である。

仮説検定の手順

手順 1：自分が証明したい仮説（対立仮説）と逆の仮説（帰無仮説）を設定し、その仮説が成り立つとして進める 。

手順 2：帰無仮説が間違っていると判断する基準（有意水準）を決める

手順 3：データの種類等に応じて検定統計量を計算する

手順 4：検定統計量から仮説の状況が起こる確率（p値）を計算する

手順 5：手順 4 で求めた確率がめったに起きない確率であったとき「立てた仮説（帰無仮説）が間違いだったのでは？」と判断する（帰無仮説を棄却する）。
そうでない場合は判断を※保留にする

※仮説が正しいという判断はしない

11.5 二項分布の正規分布近似の問 1 を用いて仮説検定のイメージを説明

表か裏が出るコインがある。このコインが不正なコインであるかどうか知りたい。不正なコインでないとすると、表と裏はそれぞれ 50%の確率で起こるものとする。

このコインを 4 回投げたときすべて表がでた。この状況からこのコインは不正なコインであると言えるか。

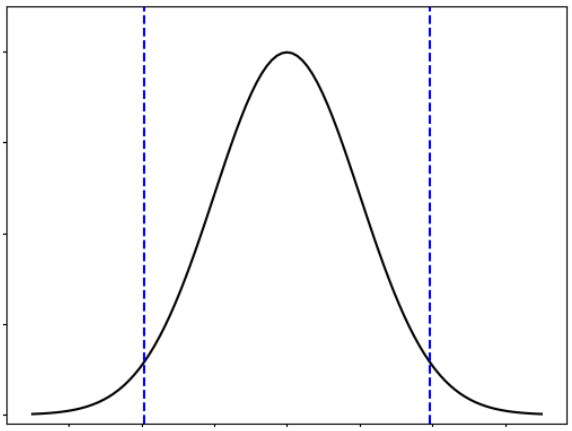
帰無仮説：

対立仮説：

帰無仮説が成り立つという仮定のもとで、4 回すべて表がでる確率は

である。

この確率は滅多に起こらないこととして帰無仮説を棄却する判断材料となる確率が有意水準である。



例題 11-4

ある雑誌の報道によると、都市 A に住む令和 3 年の独身男性の食費の月額平均は 44611 円であるという。しかし、この金額に疑問を持った太郎君は、10 人を無作為に抽出し、次の標本を得た。

41951, 51356, 54441, 44520, 45904, 54008, 45756, 36956, 47456, 49999（単位は円）

太郎君の得たこの標本をもとに、有意水準 5%で独身男性の食費の月額平均が 44611 円かどうか検定をしてみる。ただし、独身男性の食費の月額平均は正規分布に従うとし、その分散 5000^2 は既知とする。

帰無仮説：

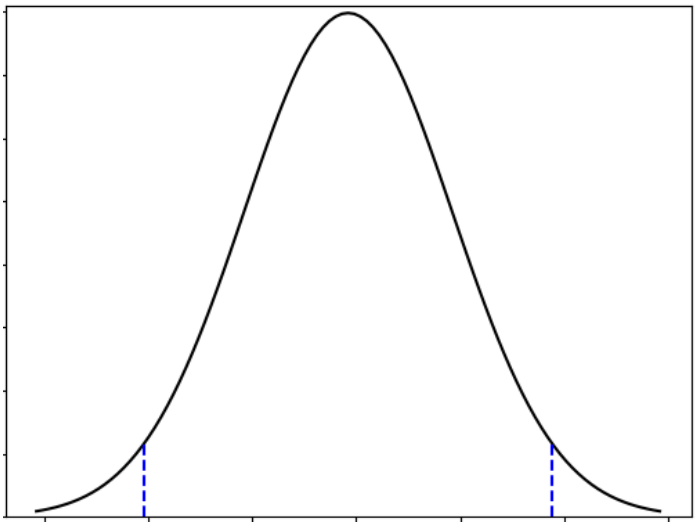
対立仮説：

（1）太郎君の得た標本から標本平均を求めなさい。

```
# 例題 11-4(1)
# 準備
import numpy as np

# 母分散の設定
sigma = 5000

# 標本抽出
sample = [41951, 51356, 54441, 44520, 45904,
          54008, 45756, 36956, 47456, 49999]
ss = len(sample)    #サンプルサイズ

# 標本平均
m = ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
print('標本平均 =', m)
```

結果：

求めた標本平均mがどこに来るかで、帰無仮説が棄却されるか、保留されるかが変わってくる。したがって標本平均の位置を確かめてみる。



（2）有意水準 5%で検定を行った結果を求めなさい。

帰無仮説のもとで標本平均の分布は $N\left(44611, \frac{5000^2}{10}\right)$ の正規分布に従う

※正規分布$(N(\mu, \sigma^2))$に従う母集団から抽出した標本平均の分布は $N\left(\mu, \frac{\sigma^2}{n}\right)$ の正規分布に従う

ここで正規母集団で母分散既知$(\boldsymbol{\sigma^2})$の母平均μの９５％信頼区間は

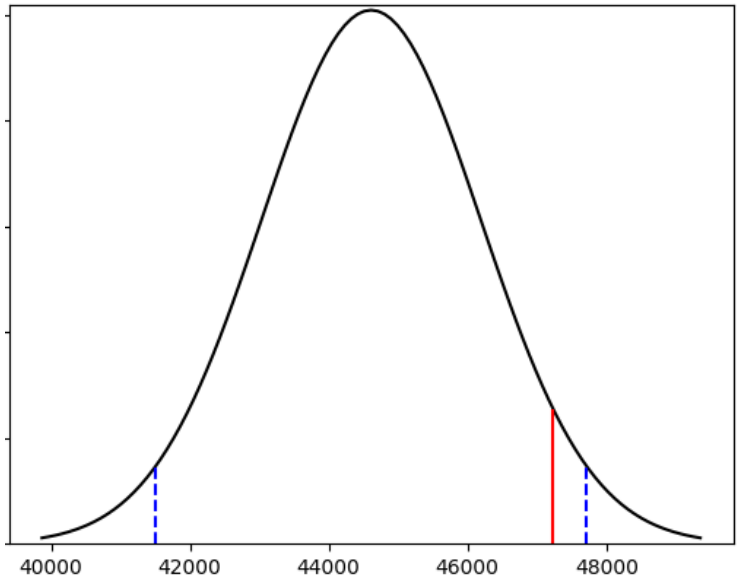
$$\bar{x} - 1.96\sqrt{\frac{\sigma^2}{n}} \leqq \mu \leqq \bar{x} + 1.96\sqrt{\frac{\sigma^2}{n}} \left(\bar{x}\text{は標本平均、}n\text{はサンプルサイズ}\right)$$

であったので、標本平均mが

$$m \leqq \bar{x} - 1.96\sqrt{\frac{5000^2}{10}}, \qquad \bar{x} + 1.96\sqrt{\frac{5000^2}{10}} \leqq m$$

を満たすとき、帰無仮説は棄却される

今回の例題では$m = 47234.7$ であったので、m は棄却域に入らない。よって帰無仮説は棄却されず、都市 A に住む、令和 3 年の独身男性の食費の月額平均は 44611 円でないとは言えない。



数学 B では標準化　$z = \frac{(x-\mu)}{\sigma}$ で変換を行い、棄却域に入るかどうかを検討する。

演習 11-4

（１）例題 11-4 の状況で、標本平均の分布は$N\left(44611, \frac{5000^2}{10}\right)$ の正規分布に従うことがわかっている。このとき、$z = \frac{(x-\mu)}{\sigma}$ で標準化を行ったときの正規分布の平均と分散を求めなさい。

平均：　　分散：

（２）標本平均 $m = 47234.7$ を標準化した値（Z値）を求めなさい。

（３）標準正規分布における有意水準 5%の棄却域は ± 1.96 であることを利用して、（２）で求めたZ値から仮説検定の結果を求めなさい。

11.7　仮説検定（母分散未知の仮説検定：1 標本の t 検定）

母分散未知で小標本の検定では、t 検定という方法で検定を行う。平均 μ 、母分散 σ^2 に従う正規母集団から標本を抽出し、標本平均を求める。その標本平均を用いて算出した統計量 t は、自由度 $n-1$ の t 分布に従うことを利用して検定を行う。

$$t = \frac{\bar{x} - \mu}{\sqrt{\frac{s^2}{n}}} \quad \left(\bar{x}\text{は標本平均、}\mu\text{は母平均、}s^2\text{は不偏分散、}n\text{はサンプルサイズ}\right)$$

例題 11-5

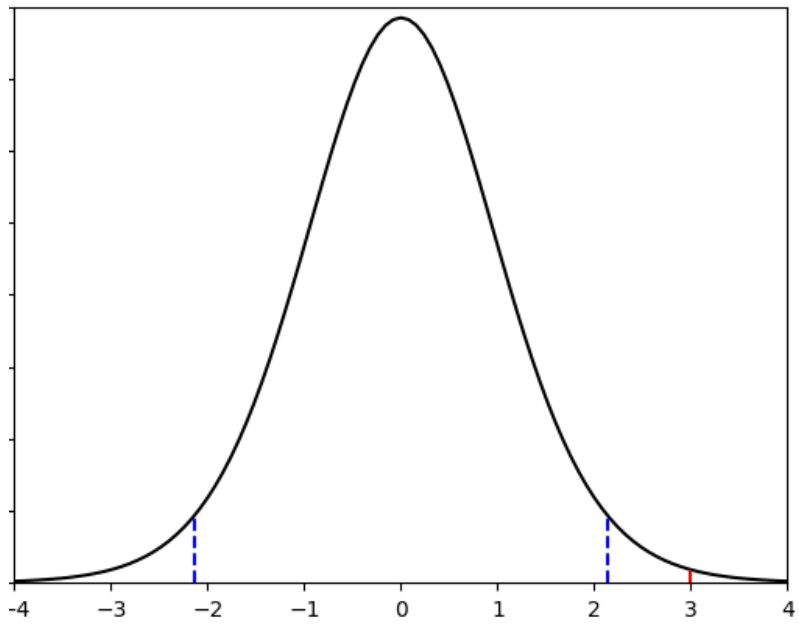
ある工場で部品を製造している。製造された部品からランダムに 16 個を選び長さを測定したところ、平均値は 7.3、不偏分散は 0.16 であった。この工場で製造している部品の長さは 7.0cm といえるかどうかを有意水準 5%で仮説検定を行いなさい。ただし、部品の長さは正規分布に従うとする。

帰無仮説：

対立仮説：

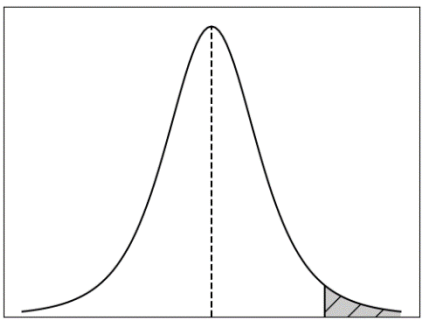
$$t = \frac{7.3 - 7}{\sqrt{\frac{0.16}{16}}}$$
$$= \frac{0.3}{0.1}$$
$$= 3$$

自由度 15 の t 分布における両側 5%点は 2.131 であり、$t > 2.313$ より統計量 t は棄却域に含まれる。したがって、帰無仮説は棄却され、工場で製造している部品の長さは 7.0cm とは言えないと結論付けることができる。



表：t 分布のパーセント点

$\nu \diagdown \alpha$	0.10	0.05	0.025	0.01	0.005
1	3.078	6.314	12.706	31.821	63.656
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169
11	1.363	1.796	2.201	2.718	3.106
12	1.356	1.782	2.179	2.681	3.055
13	1.35	1.771	2.160	2.650	3.012
14	1.345	1.761	2.145	2.624	2.977
15	1.341	1.753	2.131	2.602	2.947
16	1.337	1.746	2.120	2.583	2.921
17	1.333	1.740	2.110	2.567	2.898
18	1.330	1.734	2.101	2.552	2.878
19	1.328	1.729	2.093	2.539	2.861
20	1.325	1.725	2.086	2.528	2.845
21	1.323	1.721	2.080	2.518	2.831
22	1.321	1.717	2.074	2.508	2.819
23	1.319	1.714	2.069	2.500	2.807
24	1.318	1.711	2.064	2.492	2.797
25	1.316	1.708	2.060	2.485	2.789
26	1.315	1.706	2.056	2.479	2.779
27	1.314	1.703	2.052	2.473	2.771
28	1.313	1.701	2.048	2.467	2.763
29	1.311	1.699	2.045	2.462	2.756
30	1.310	1.697	2.042	2.457	2.750
40	1.303	1.684	2.021	2.423	2.704
60	1.296	1.671	2.000	2.390	2.660



ν は自由度、α は左図の網掛け部分（確率）

演習 11-5

（1）例題 11-5 と同様の状況において、Python で統計量 t を計算し、検定結果を出力するプログラムを作成しなさい。

（2）A 社ではある機械で工業製品を作っている。この機械で作られた製品の重量は平均が 15g になるように設定されている。この機械で作られた製品の中からランダムに 20 個を取り出して重量を測定したところ以下のようになった。このサンプルから、平均は 15g ではないといえるか、有意水準 5%で検定を行いなさい。ただし、重量のデータは classroom に配信されたスプレッドシート「1 標本の t 検定」から読み込みなさい。

15.12, 16.32, 13.70, 12.14, 13.75, 13.45, 14.56, 13.65, 17.35, 14.69,
15.08, 16.05, 14.06, 11.66, 15.16, 12.18, 14.84, 14.02, 15.24, 12.95

現実的にはデータを収集してスプレッドシート等に入力する。そのデータから平均、不偏分散、サンプルサイズを求め、t 値を計算し、その t 値が棄却域に入っているかを考える必要がある。しかし、この手順を踏むのは大変。そこで、Python には t 検定を簡単に行えるものが準備されている。

scipy.stats.ttest_1samp(x, popmean)

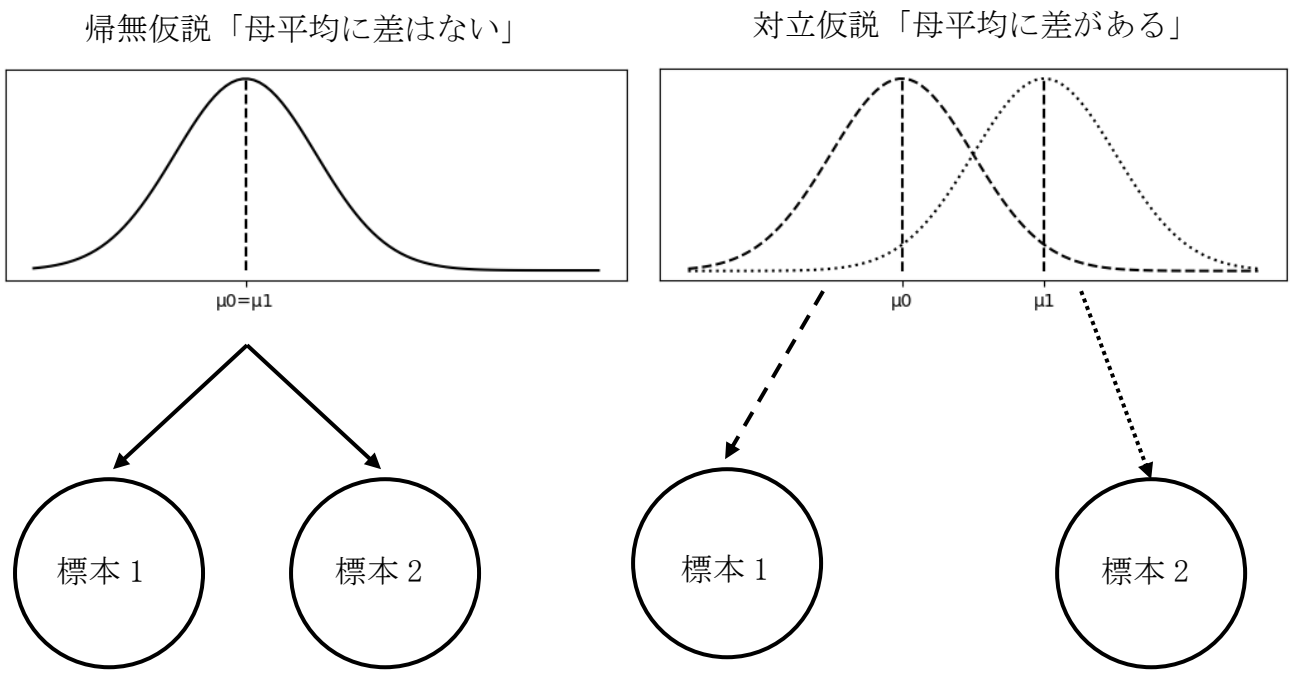
x:データを入力

popmean：検定したい母平均

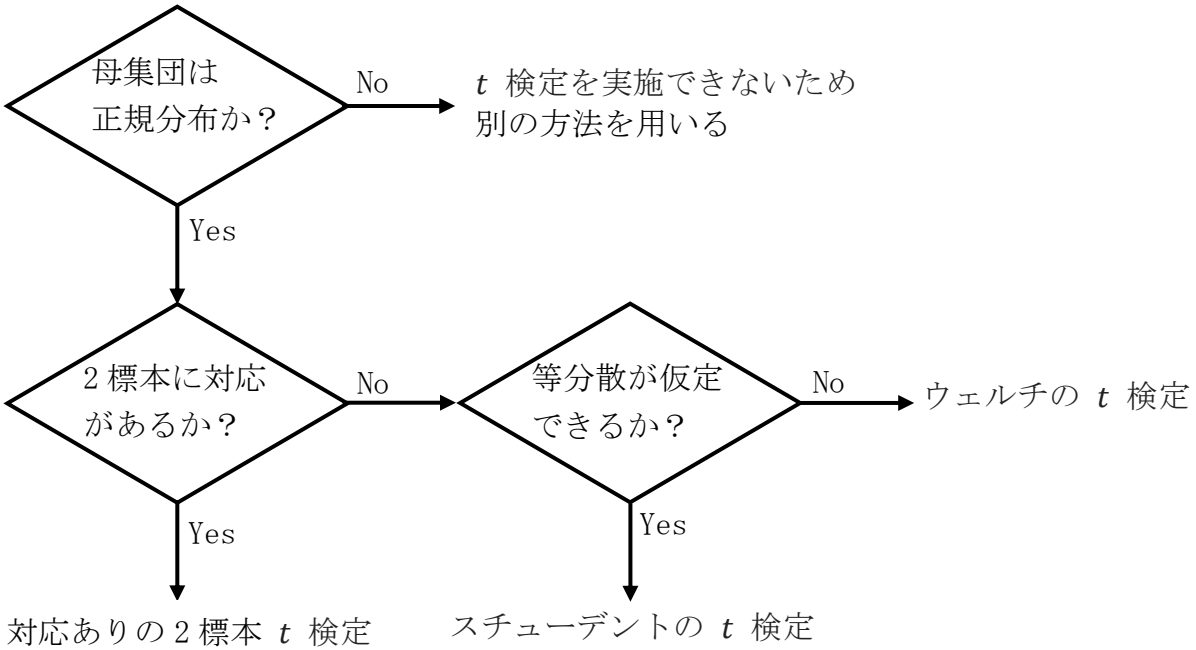
11.8　仮説検定（2 標本の t 検定）

11.7 では標本が 1 種類だけであった。ところが実際には実験前のデータと実験後のデータを比較したり、2 組の標本を見て、母平均に差があるのかどうかを検定したりする場合がある。その場合には 1 標本の t 検定ではなく、2 標本の t 検定と呼ばれる方法を実施する。

2 標本の t 検定での帰無仮説は「母平均に差はない」となる。2 標本を抽出した母平均に差はないということは、2 標本は同じ母平均の正規分布から抽出されたものということができる。



2 標本の t 検定を行う場合に注意しなければいけないのは、どのような状況なのかをはっきりさせることである。2 標本の t 検定では状況に応じて検定統計量が異なるため、間違えた結果を引き起こすことが想定される。以下に状況に応じて使用する方法をまとめた。



対応あり　　対応なし

実験前　　実験後

標本 1	標本 2	標本 1	標本 2
A	A	A	F
B	B	B	G
C	C	C	H
D	D	D	I
E	E	E	J

2 つの標本で同じ個体を用いる場合は対応あり、別の個体を用いる場合は対応なし。対応ありのデータに対して、対応なしの検定を用いるのはよいが、対応なしのデータに対応ありの検定を用いることはできないことに注意が必要。

等分散性が仮定できるとは、抽出した 2 つの標本の分散が同じであること。分散が等しいかどうかを検定することもできるが、等分散であるかどうかにかかわらず、等分散性を仮定しないウェルチの t 検定を用いる場合が多いようである。等分散性が仮定できるデータにウェルチの t 検定を用いることはできるが、等分散性が仮定できないデータにスチューデントの t 検定を用いることはできない。

例題 11-6

以下の表は B 組の生徒 45 人と C 組の生徒 40 人の数学の点数をまとめたものである。この標本からＢ組とＣ組の平均に差があるといえるか、有意水準 5%で検定しなさい。ただし、母集団は正規分布に従うものとし、等分散性については仮定しないものとする。得点のデータは classroom に配信してあるスプレッドシート「2 標本の t 検定」からデータを読み込みなさい。

番号	1	2	…	40	41	42	43	44	45
B 組	62	81	…	77	73	61	74	54	35
C 組	71	54	…	63					

このときの検定統計量 t は以下の式で与えられる。

$$t = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{{s_1}^2}{n_1} + \frac{{s_2}^2}{n_2}}}$$

この検定統計量 t は自由度 ν の t 分布に従うが、通常のように自由度が $(n-1)$ にはならない。少し複雑な計算を行うことで自由度 ν を求めることができるが、Python で求める方法を紹介する。

scipy.stats.ttest_ind (x, y, equal_var=True)

x, y:データを入力

equal_var：デフォルトは等分散を仮定する(True)、仮定しない場合は(False)

演習 11-6

ある薬の実験において、A 群には薬 1 を、B 群には薬 2 を一定期間飲んでもらった。以下の表はその結果をまとめたものである。この結果から薬 1 に血圧を変化させる効果があるといえるか。有意水準 5%で検定を行いなさい。ただし、等分散性は仮定しないウェルチの t 検定

A 群	B 群
112.7	127.1
109.4	121.8
123.2	120.3
119.3	130.8
113.9	123.9
118.8	109.6
127.6	126.9
126.9	123.4
121.9	128.2
110.2	140.1