ＡＩ基礎セミナー

第６回
数学の基礎（３回目：統計学）

改訂履歴

日付	担当者	内容
2021/05/08	M.Takeda	Git 公開
2022/04/03	M.Takeda	「(3.2.3) 二項分布」の説明を更新
2022/10/10	M.Takeda	「(3.3.2) 対数正規分布」を追加
		「(3.3.4) 連続一様分布」ｸﾞﾗﾌ・ﾘｽﾄ差替え

目次

(1) はじめに

・「第６回　数学の基礎（３回目)」は、「数学の基礎」シリーズの最終回となります。

・数学の基礎の３回目は、統計学を中心に扱い、機械学習で必要となる最小限の基本的な事項について把握することを目標とし、Python での実装をみながら概観します。

・統計学は非常に範囲が広く奥行きも深い学問です。
本セミナーでは扱う範囲を以下のものに絞り込むことにします。
  ・基本的な統計指標
  ・確率変数と確率分布
  ・統計学の基本定理
  ・相関分析
  ・仮説検定

・分類手法の「K-means法」、「混合ガウスモデル (Gaussian Mixture Model, GMM)」や、回帰モデルの「線形回帰モデル」といった手法は、後の機械学習の回で改めて扱うこととします。
（線形回帰モデルは「数学の基礎（１回目：代数学)」でも少し触れています）

・確認問題は、本文を要約したような問題となっております。
解いてみて、理解度を確認してみてください。

(2) 基本的な統計指標

この章では、基本的な統計指標を記します。

(2.1) 平均値、中央値、四分位数

データの代表値として、平均値、中央値、四分位数について見てみましょう。

・「平均値（ﾍｲｷﾝﾁ、mean value)」は、観測値の総和を観測値の個数で割ったものです。

$$\text{平均値} = (1/n)\ \sum_{i=1}^{n} x_i \qquad \cdots(\text{式}2.1\text{-}1)$$

n：観測値の個数

x_i：i番目の観測値（i=1～n）

・「中央値（ﾁｭｳｵｳﾁ、median)」とは、観測値を小さい順に並べたとき中央に位置する値です。
データの分布が対称である場合は、中央値は平均値に等しくなります。
分布が対称でなくても、中央値と平均値が等しくなる事もあります。

・中央値は平均値と類似した目的で使いますが、全体の傾向を表す代表値として適切である場合が多いです。
（例えば、ある国の年収に着目した時、貧富の差が激しい国では、ほんの一握りの富裕層が
平均年収を釣り上げてしまい、平均値より中央値の方が年収の実態を反映した指標となります。）

・データを小さい順に並び替えたときに、データの数で四等分した時の区切り値を
「四分位数（ｼﾌﾞﾝｲｽｳ、quartile)」と言います。四等分すると三つの区切りの値が得られ、
小さいほうから「25パーセンタイル（第一四分位数）」、「50パーセンタイル（中央値）」、
「75パーセンタイル（第三四分位数）」とよびます。
第一四分位数・第三四分位数の差は、「四分位範囲（ｼﾌﾞﾝｲﾊﾝｲ、interquartile range, IQR)」といい、
分布のばらつきの代表値となっています。

（例2.1）2019/1/3 の鹿児島の気温（気象庁ｱﾒﾀﾞｽより）

時刻	気温
時	℃
1	2.9
2	3.3
3	3.7
4	3.3
5	3.0
6	3.2
7	3.5
8	3.6
9	6.1
10	9.6
11	10.6
12	12.0
13	12.8
14	12.7
15	13.1
16	12.6
17	11.1
18	9.8
19	8.9
20	7.4
21	6.7
22	6.0
23	5.6

順位	気温
昇順に並べ替え	
1	2.9
2	3.0
3	3.2
4	3.3
5	3.3
6	**3.5**
7	3.6
8	3.7
9	5.6
10	6.0
11	6.1
12	**6.7**
13	7.4
14	8.9
15	9.6
16	9.8
17	10.6
18	**11.1**
19	12.0
20	12.6
21	12.7
22	12.8
23	13.1
（平均）	7.5

←25パーセンタイル（第一四分位数）

←中央値 ≒平均値（7.5℃）

←50パーセンタイル（第二四分位数）

←75パーセンタイル（第三四分位数）

四分位範囲 ＝11.1 - 3.5 = 7.6

(2.2) 階級分けと度数分布

データの集計方法として、階級分けと度数分布について見てみましょう。

・データをある「階級幅 (ｶｲｷｭｳﾊﾊﾞ, class width)」で区切って、
幾つかの「階級 (ｶｲｷｭｳ, class)」に分けます。
その階級に属する個数を「度数 (ﾄﾞｽｳ, frequency)」と呼び、
それをデータの全個数で割ったものを「相対度数 (ｿｳﾀｲﾄﾞｽｳ、relative frequency)」と呼びます。
各階級を代表する値を「階級値 (ｶｲｷｭｳﾁ, class value)」と呼び、階級の中央値に相当します。
各階級毎の度数の分布は「度数分布 (ﾄﾞｽｳﾌﾞﾝﾌﾟ, frequency distribution)」と言い、
階級値の順に度数を表化したものが「度数分布表 (ﾄﾞｽｳﾌﾞﾝﾌﾟﾋｮｳ)」です。

・例として、ある中学校の男子生徒 50人の身長について測定した結果を取り上げます (例2.2)。

(例2.2) ある中学校の男子生徒 50人の身長についての度数分布と、平均、偏差、分散、標準偏差

身長範囲	階級値	度数	相対度数	各生徒の身長	偏差	偏差の二乗
142.5 ～ 147.5	145	2	0.04	143.5	−16.5	272.5
				145.5	−14.5	210.5
147.5 ～ 152.5	150	5	0.10	148.0	−12.0	144.2
				150.0	−10.0	100.2
				151.0	−9.0	81.1
				151.5	−8.5	72.4
				152.0	−8.0	64.1
152.5 ～ 157.5	155	9	0.18	152.5	−7.5	56.4
				153.0	−7.0	49.1
				153.5	−6.5	42.4
				154.0	−6.0	36.1
				154.5	−5.5	30.3
				155.0	−5.0	25.1
				155.5	−4.5	20.3
				156.0	−4.0	16.1
				156.5	−3.5	12.3
157.5 ～ 162.5	160	15	0.30	157.5	−2.5	6.3
				157.8	−2.2	4.9
				158.0	−2.0	4.0
				158.3	−1.7	2.9
				158.5	−1.5	2.3
				158.7	−1.3	1.7
				159.0	−1.0	1.0
				159.2	−0.8	0.7
				159.4	−0.6	0.4
				159.5	−0.5	0.3
				160.3	0.3	0.1
				161.0	1.0	1.0
				161.1	1.1	1.2
				161.4	1.4	1.9
				162.4	2.4	5.7
162.5 ～ 167.5	165	12	0.24	162.6	2.6	6.7
				162.8	2.8	7.8
				163.2	3.2	10.2
				163.7	3.7	13.6
				164.1	4.1	16.7
				164.3	4.3	18.4
				165.4	5.4	29.1
				165.6	5.6	31.3
				166.3	6.3	39.6
				166.4	6.4	40.9
				167.3	7.3	53.2
				167.4	7.4	54.6
167.5 ～ 172.5	170	5	0.10	167.8	7.8	60.7
				168.0	8.0	63.9
				169.0	9.0	80.9
				170.4	10.4	108.0
				172.0	12.0	143.8
172.5 ～ 177.5	175	2	0.04	173.0	13.0	168.8
				177.0	17.0	288.7

(項目)	標本数	合計	平均	偏差総和	分散	標準偏差
(集計)	50	1	160.0	0.0	50.1	7.1

(参考資料)

「数学公式事典」 黒田孝朗・須田貞之著 文研出版 1978年

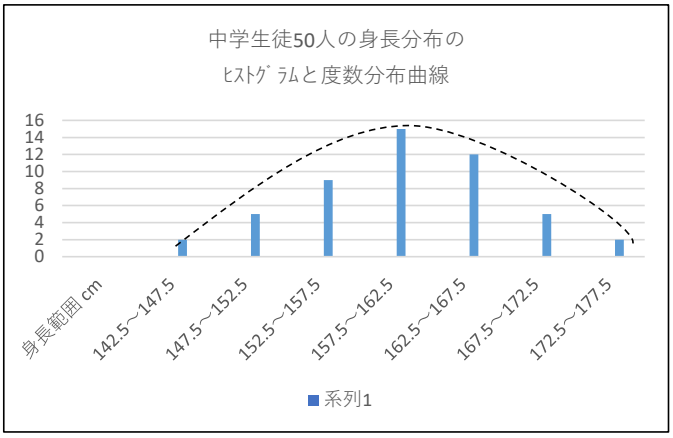
・着目する階級までの度数を累積したものを「累積度数（ﾙｲｾｷﾄﾞｽｳ）」と呼びます。
着目する階級までの相対度数を累積したものを「累積相対度数（ﾙｲｾｷｿｳﾀｲﾄﾞｽｳ）」と呼びます。

(例2.3)（例2.2)の男子生徒 50人の身長についての度数分布表と、相対度数・累積度数

身長範囲 cm	階級値 x	度数 f	相対度数	累積度数	累積相対度数
142.5～147.5	145	2	0.04	2	0.04
147.5～152.5	150	5	0.10	7	0.14
152.5～157.5	155	9	0.18	16	0.32
157.5～162.5	160	15	0.30	31	0.62
162.5～167.5	165	12	0.24	43	0.86
167.5～172.5	170	5	0.10	48	0.96
172.5～177.5	175	2	0.04	50	1.00
	（合計）	50	1.00		

・度数分布表から、横軸上に階級値を、
縦軸に度数を目盛り、横軸を階級に従って区分し、
それを底辺とする長方形を、面積が度数に
比例するように描いた柱状ｸﾞﾗﾌが
「ﾋｽﾄｸﾞﾗﾑ（histogram)」です（右図）。

・度数分布で、標本数を大きくし、階級の区切りを
非常に細かくして行くと、
ﾋｽﾄｸﾞﾗﾑの上辺をなす階段状の線は
滑らかな曲線に近づきます。これを「度数分布曲線
(Frequency distribution curve)」といいます。



【出典・参考】

中央値⇒ https://ja.wikipedia.org/wiki/中央値
平均値⇒ https://ja.wikipedia.org/wiki/平均
四分位数⇒ https://bellcurve.jp/statistics/course/19277.html
度数分布⇒ https://ja.wikipedia.org/wiki/度数分布
度数分布、相対度数⇒ https://bellcurve.jp/statistics/course/1625.html

(2.3) 偏差と分散、標準偏差

データのばらつきの指標として、偏差、分散、標準偏差について見てみましょう。

・各データと平均値μとの差を「偏差（ﾍﾝｻ, deviation)」と言いますが、これの総和は常に0になるため、
データのばらつき具合の指標として、偏差をそのままで扱うのには問題があります。

偏差の総和

$$\begin{aligned} &= \sum_{i=1}^{n} (x_i - \mu) \\ &= \sum_{i=1}^{n} (x_i) - n*\mu \\ &= \sum_{i=1}^{n} (x_i) - n*(1/n)\sum_{i=1}^{n} (x_i) \\ &= 0 \end{aligned}$$

・偏差の代わりに、偏差の二乗和の平均を取ったものを「分散（ﾌﾞﾝｻﾝ, variance)」と呼び、
これをデータのばらつき具合の指標として使い、σ^2（ｼｸﾞﾏの2乗)または V（ﾌﾞｲ)で記述します。
分散はデータがどの程度平均値の周りにばらついているかの指標になります。

分散

$$\sigma^2 = (1/n) \sum_{i=1}^{n} (x_i - \mu)^2 \qquad \cdots(式2.3\text{-}1)$$

・分散の平方根を「標準偏差（ﾋｮｳｼﾞｭﾝﾍﾝｻ, standard deviation)」と呼び、σ(ｼｸﾞﾏ)で記述します。
二乗和を取った分散とは異なり、元のデータと同じ次元（単位）になっています。

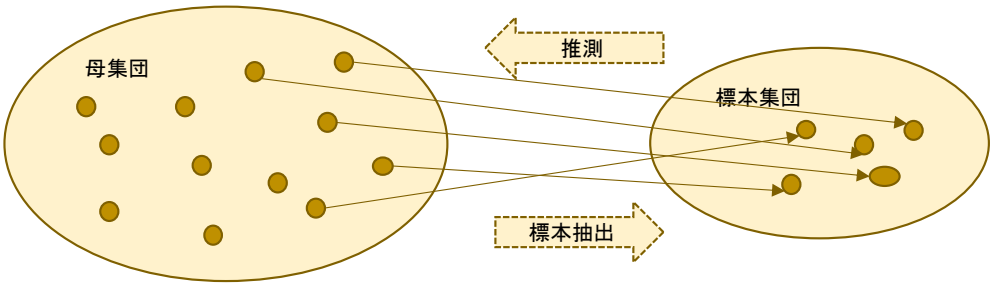
標準偏差

$$\sigma = \sqrt{(1/n) \sum_{i=1}^{n} (x_i - \mu)^2} \qquad \cdots(式2.3\text{-}2)$$

(2.4) 母集団と標本集団

・「母集団（ﾎﾞｼｭｳﾀﾞﾝ、population)」とは、調査対象となる集合全体を言います。

・母集団全体を調査の対象とすることが出来ない時、母集団の中から一定の数のデータを抜き出します。
抜き出した集団を「標本あるいは標本集団（ﾋｮｳﾎﾝｼｭｳﾀﾞﾝ、sample)」、
標本集団を選ぶことを「標本抽出（ﾋｮｳﾎﾝﾁｭｳｼｭﾂ、sampling)」、
抜き出したデータの個数を「標本の大きさ（sample size)」と呼びます。



・データの集合が母集団であるとき、
データの集合の平均を「母平均（ﾎﾞﾍｲｷﾝ、population mean)」（μ（ﾐｭｰ））、
データの集合の標準偏差を「母標準偏差（ﾎﾞﾋｮｳｼﾞｭﾝﾍﾝｻ, population standard deviation)」（σ（ｼｸﾞﾏ））、
母標準偏差の二乗を「母分散（ﾎﾞﾌﾞﾝｻﾝ, population variance)」（σ^2）と呼びます。

母集団の母数

母平均　$\mu = (1/N) \sum_{i=1}^{N} x_i$　N：母集団の個数　…(式2.4-1)

母分散　$\sigma^2 = (1/N) \sum_{i=1}^{N} (x_i - \mu)^2$　…(式2.4-2)

・データの集合が標本集団であるとき、
データの集合の平均を「標本平均（ﾋｮｳﾎﾝﾍｲｷﾝ、sample mean)」（x_{mean}）、
データの集合の標準偏差を「標本標準偏差（ﾋｮｳﾎﾝﾋｮｳｼﾞｭﾝﾍﾝｻ, sample standard deviation)」（s（ｴｽ））、
標本標準偏差の二乗を「標本分散（ﾋｮｳﾎﾝﾌﾞﾝｻﾝ, sample variance)」（s^2（ｴｽのﾆｼﾞｮｳ））と呼びます。

標本集団の統計量

標本平均　$x_{mean} = (1/n) \sum_{i=1}^{n} x_i$　n：標本の大きさ　…(式2.4-3)

標本分散　$s^2 = (1/n) \sum_{i=1}^{n} (x_i - x_{mean})^2$　…(式2.4-4)

・n を標本の大きさとして、
標本分散 s^2 に n/(n-1) を掛けたものを「不偏分散（ﾌﾍﾝﾌﾞﾝｻﾝ, unbiased estimate of variance)」
（u^2）と呼びます。不偏分散は標本平均が標本全体を代表しているものとみて、
データの個数を１減らして分散を評価したもので、標本分散に代わる指標です。

標本集団の統計量

不偏分散　$u^2 = (1/(n-1)) \sum_{i=1}^{n} (x_i - x_{mean})^2 = (n/(n-1))\, s^2$　…(式2.4-5)

・「推測統計学（ｽｲｿｸﾄｳｹｲｶﾞｸ、inferential statistics)」では、母集団の母数（母平均と母分散）を、
それから無作為抽出（random sampling）した標本集団の統計量（標本平均と不偏分散）から推定します。

【出典・参考】

母集団と標本⇒ https://www.weblio.jp/content/母集団と標本

母集団と標本⇒ https://bellcurve.jp/statistics/course/8003.html

標本分散と不偏分散⇒ https://bellcurve.jp/statistics/course/8614.html

(3) 確率変数と確率分布

(3.1) 確率変数と確率分布

・「確率変数（ｶｸﾘﾂﾍﾝｽｳ、random variable)」とは、ある確率で値を取る変数のことです。
確率変数を X、その確率を P(X) で表現します。

・確率変数がとる値とその出現確率の対応を「確率分布（ｶｸﾘﾂﾌﾞﾝﾌﾟ、probability distribution)」と言います。確率分布には、確率変数が離散的である「離散型確率分布」と、連続的である「連続型確率分布」があります。

(例3.1) さいころを投げて出る目 X は {1, 2, 3, 4, 5, 6} の何れかで、
それぞれの目が出る確率 P(X) は 1/6 であることから、
さいころを投げて出る目 X は離散型の確率変数となります。
確率変数 X の確率分布は離散型であり、以下の様になります：

出る目 X	1	2	3	4	5	6	(合計)
確率 P(X)	1/6	1/6	1/6	1/6	1/6	1/6	1

(例3.2) コインを投げて出る目 X は {表=1,裏=0} の何れかで、
それぞれの目が出る確率 P(X) は 1/2 であることから、
コインを投げて出る目 X は離散型の確率変数となります。
確率変数 X の確率分布は離散型であり、以下の様になります：

出る目 X	表(=1)	裏(=0)	(合計)
確率 P(X)	1/2	1/2	1

(例3.3) 「(例2.2) ある中学校の男子生徒 50人の身長についての度数分布・・・」で取り上げた
身長 X は連続型の値ですが、身長範囲で階級分けした階級値 Y は、
その相対度数を出現確率として扱うことが出来、離散型の確率変数であるといえます。
確率変数 Y の確率分布は離散型となり、以下の様になります：

身長範囲 cm	階級値 Y	度数 f	相対度数	確率 P(Y)
142.5～147.5	145	2	0.04	0.04
147.5～152.5	150	5	0.10	0.10
152.5～157.5	155	9	0.18	0.18
157.5～162.5	160	15	0.30	0.30
162.5～167.5	165	12	0.24	0.24
167.5～172.5	170	5	0.10	0.10
172.5～177.5	175	2	0.04	0.04
	(合計)	50	1.00	1.00

【出典・参考】

統計学の時間⇒ https://bellcurve.jp/statistics/course/
確率変数⇒ https://bellcurve.jp/statistics/course/6596.html
確率分布⇒ https://bellcurve.jp/statistics/glossary/800.html
指数分布⇒ https://bellcurve.jp/statistics/course/8009.html
正規分布⇒ https://bellcurve.jp/statistics/course/7797.html
二項分布⇒ https://bellcurve.jp/statistics/course/6979.html
二項分布⇒ https://ja.wikipedia.org/wiki/二項分布
ﾎﾟｱｿﾝ分布⇒ https://bellcurve.jp/statistics/course/6984.html
確率密度関数⇒ https://ja.wikipedia.org/wiki/確率密度関数
連続一様分布⇒ https://ja.wikipedia.org/wiki/連続一様分布
離散一様分布⇒ https://ja.wikipedia.org/wiki/離散一様分布

(3.2) 離散型確率分布

・確率変数が離散的である（飛び飛びの値を取る）場合の確率分布が
「離散型確率分布（ﾘｻﾝｶﾞﾀｶｸﾘﾂﾌﾞﾝﾌﾟ、discrete probability distribution)」です。

・離散型確率変数の各値 x_i に対してその発生確率$P(x_i)$ を表す(離散型)確率分布の関数を、
「確率質量関数（ｶｸﾘﾂｼﾂﾘｮｳｶﾝｽｳ、probability mass function)」と言います。
離散型確率分布では、各離散値 x_i に対する発生確率の総和は１となります。

$$\sum_i P(x_i) = 1 \quad (0 \leqq P(x_i) \leqq 1)$$

・以下に、離散型確率分布の幾つかを紹介します。

(3.2.1) 離散一様分布

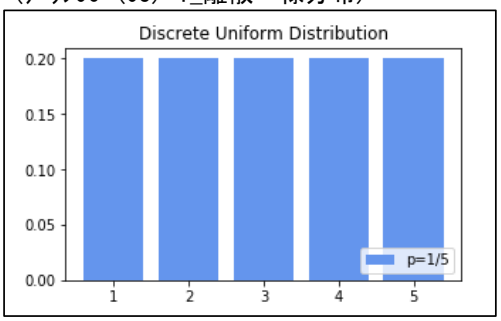
・既出の(例3.1)のサイコロの出る目や、(例3.2)コインの出る目のように、
離散的な確率変数 X の全ての値 k について、発生確率 P(X=k) が等しい場合の確率分布を、
「離散一様分布 (ﾘｻﾝｲﾁﾖｳﾌﾞﾝﾌﾟ、discrete uniform distribution)」と言います。

・離散一様分布の確率質量関数 P(X=k) は、以下の式で表されます。

離散一様分布の確率質量関数

$$P(X=k) = H \qquad (H = 1 / \text{確率変数の総数}) \qquad \cdots(式3.2\text{-}1)$$

(ｸﾞﾗﾌ06-(03)-1_離散一様分布)



(ﾘｽﾄ06-(03)-1_離散一様分布)

```
#*************************************************
# ﾘｽﾄ06-(03)-1_離散一様分布
#*************************************************
import numpy as np
import matplotlib.pyplot as plt
from mpl_toolkits.mplot3d import Axes3D
%matplotlib inline

# 確率質量関数 P(k)
def fUniformDist(n):
    return 1/n

# パラメータに応じた分布計算
n = 5
Xlist = range(n)
Llist = []
Plist = []
for x in Xlist:
    Llist.append(x + 1)
    Plist.append( fUniformDist(n) )

# グラフ描画
plt.figure(figsize=(5, 3))
plt.title('Discrete Uniform Distribution')
plt.bar(Xlist, Plist, tick_label=Llist, align="center", ¥
        facecolor="cornflowerblue", label='p=1/5')
plt.legend(loc='lower right')
plt.show()
```

(3.2.2) ベルヌーイ分布

・「コインを投げたときに出るのは表か裏」のように、何かを行ったときに起こる結果が
事象 A, B の２つ(A：成功する、B：失敗する)しかない試行のことを
「ベルヌーイ試行 (Bernoulli trial)」といいます。

・１回のベルヌーイ試行で成功する回数（0, 1 の何れか）を確率変数 X としたとき、
X が従う確率分布が「ベルヌーイ分布 (ﾍﾞﾙﾇｰｲﾌﾞﾝﾌﾟ、Bernoulli distribution)」です。

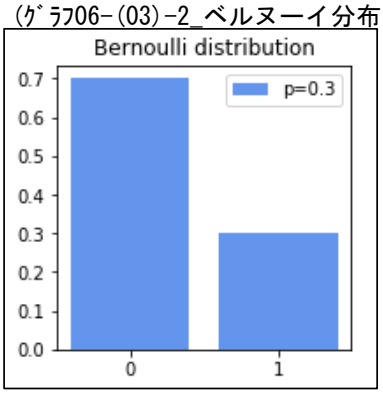
・１回のベルヌーイ試行で成功する確率を p とすると、確率変数 X が k となる
ベルヌーイ分布の確率質量関数 P(X=k) は以下の式で表されます。

ベルヌーイ分布の確率質量関数

$$P(X=k) = p^k (1-p)^{1-k} \quad \text{for } k \in \{0,1\} \qquad \cdots(式3.2\text{-}2)$$

・この式は、次節「(3.2.3) 二項分布」の式で、n=1 としたものに等しいです。
確率分布は、P(1)＝p、P(0)＝1-p　となります。

(ｸﾞﾗﾌ06-(03)-2_ベルヌーイ分布 (P=0.3))



(ﾘｽﾄ06-(03)-2_ベルヌーイ分布)

```
#****************************************************
# ﾘｽﾄ06-(03)-2_ベルヌーイ分布
#****************************************************
import numpy as np
import matplotlib.pyplot as plt
from mpl_toolkits.mplot3d import Axes3D
%matplotlib inline

# ベルヌーイ分布関数
def fBernoulli(p,k):
    return ((p)**k) * ((1-p)**(1-k))

# パラメータに応じた分布計算
n = 2
p = 0.3
Xlist = range(n)
Llist = []
Plist = []
for x in Xlist:
    Llist.append(x)
    Plist.append( fBernoulli(p, x) )

# グラフ描画
plt.figure(figsize=(3, 3))
plt.title('Bernoulli distribution')
plt.bar(Xlist, Plist, tick_label=Llist, align="center", facecolor="cornflowerblue", label='p=0.3')
plt.legend(loc='upper right')
plt.show()
```

(3.2.3) 二項分布

・n 回のベルヌーイ試行において成功する回数を確率変数 X （∈{0,1,..,n}）としたとき、
X が従う確率分布が「二項分布（ﾆｺｳﾌﾞﾝﾌﾟ、binominal distribution)」です。

・ベルヌーイ分布に従う一つの確率変数を X_i（取り得る値は 0 または 1)としたとき、
n個の確率変数 X_i（i=1～n）の和 X が従う確率分布が二項分布になる、ということもできます。

$$X = \Sigma_{i=1}^{n} X_i \qquad X \in \{0,1,..,n\}、X_i \in \{0,1\}、i=1～n$$

・ｎ回のベルヌーイ試行で成功する回数 X が k となる確率質量関数 P(X=k) は、
１回のベルヌーイ試行で成功する確率を p とすると、以下の式で表されます。

二項分布の確率質量関数

$$P(X=k) = {}_nC_k \, p^k \, (1-p)^{n-k} \qquad \text{for } k \in \{0,1,..,n\} \qquad \cdots(式3.2\text{-}3)$$

（解説）

${}_nC_k$:

n個から k個取る組合せの数を「${}_nC_k$（ｴﾇｼｰｹｰ、n choose x)」で表します。
(C は Combination の略) :

$${}_nC_k = n!／k!(n-k)!$$
$$= n(n-1)\cdots(n-k+1)／k(k-1)\cdots2\cdot1 \qquad \text{for } k \in \{0,1,..,n\}$$

$p^k \, (1-p)^{n-k}$:

n 回のベルヌーイ試行で、成功したのが何回目の組合せかということを考慮せずに、
単に、成功回数＝k（失敗回数＝n-k）となる割合は、以下の式となります :

$$p^k \, (1-p)^{n-k} \qquad \text{for } k \in \{0,1,..,n\}$$

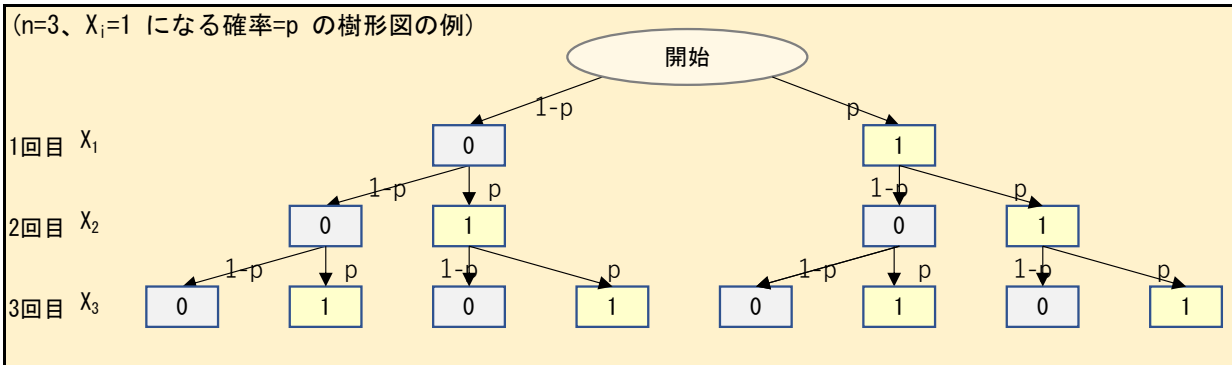
n 回のベルヌーイ試行で、成功回数＝k(失敗回数＝n-k）となる組合せの数が ${}_nC_k$ であり、
(式3.2-3)は、組合せ数を考慮した確率になっています。
右辺は pと(1-p) の二項展開式の「一般項」に相当し、「${}_nC_k$」は「二項係数」といいます。

$$1 = (p + (1-p))^n = \Sigma_{k=0}^{n} {}_nC_k \, p^k (1-p)^{n-k}$$

（この展開式を「二項定理」と言います）

・二項分布は「樹形図」で表現すると分かり易いです。

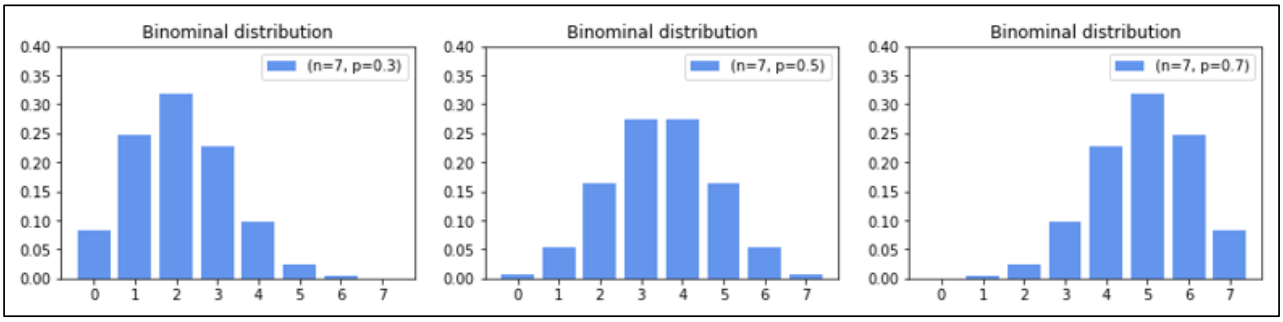
(n=3、X_i=1 になる確率=p の樹形図の例)



合計(k) :	0	1	1	2	1	2	2	3
確率 :	$(1-p)^3$	$p(1-p)^2$	$p(1-p)^2$	$p^2(1-p)$	$p(1-p)^2$	$p^2(1-p)$	$p^2(1-p)$	p^3
ﾊﾟﾀｰﾝNo. :	1	2	3	4	5	6	7	8

n=3	k	組合せﾊﾟﾀｰﾝ	組合せ数	個々のﾊﾟﾀｰﾝの発生確率
	k=0	ﾊﾟﾀｰﾝNo. 1	$1={}_3C_0$	$(1-p)^3$
	k=1	ﾊﾟﾀｰﾝNo. 2, 3, 5	$3={}_3C_1$	$p(1-p)^2$
	k=2	ﾊﾟﾀｰﾝNo. 4, 6, 7	$3={}_3C_2$	$p^2(1-p)$
	k=3	ﾊﾟﾀｰﾝNo. 8	$1={}_3C_3$	p^3

(ｸﾞﾗﾌ06-(03)-3_二項分布(n=7,P=0.3/0.5/0.7))



(ﾘｽﾄ06-(03)-3_二項分布(n=7,P=0.3/0.5/0.7))

```
#****************************************************
# ﾘｽﾄ06-(03)-3_二項分布(n=7,P=0.3/0.5/0.7)
#****************************************************
import numpy as np
import math
import matplotlib.pyplot as plt
from mpl_toolkits.mplot3d import Axes3D
%matplotlib inline

# 二項分布の確率質量関数
def fBinominal(n, p, k):
    return fCombinaton(n,k) * ((p)**k) * ((1-p)**(n-k))

# 組合せ nCr
def fCombinaton(n, r):
    return  math.factorial(n) // ¥
           (math.factorial(n - r) * math.factorial(r))

# パラメータに応じた分布計算
n = 7
p1 = 0.3
p2 = 0.5
p3 = 0.7
Xlist = range(n+1)
Llist = range(n+1)
Plist1 = []
Plist2 = []
Plist3 = []
for x in Xlist:
    Plist1.append( fBinominal(n, p1, x) )
for x in Xlist:
    Plist2.append( fBinominal(n, p2, x) )
for x in Xlist:
    Plist3.append( fBinominal(n, p3, x) )

# グラフ描画
plt.figure(figsize=(15, 3))

plt.subplot(1, 3, 1)
plt.title('Binominal distribution')
plt.bar(Xlist, Plist1, tick_label=Llist, align="center", ¥
        facecolor="cornflowerblue", label='(n=7, p=0.3)')
plt.legend(loc='upper right')
plt.ylim(0.0, 0.4)

plt.subplot(1, 3, 2)
plt.title('Binominal distribution')
plt.bar(Xlist, Plist2, tick_label=Llist, align="center", ¥
        facecolor="cornflowerblue", label='(n=7, p=0.5)')
```

```
plt.legend(loc='upper right')
plt.ylim(0.0, 0.4)

plt.subplot(1, 3, 3)
plt.title('Binominal distribution')
plt.bar(Xlist, Plist3, tick_label=Llist, align="center", ¥
        facecolor="cornflowerblue", label='(n=7, p=0.7)')
plt.legend(loc='upper right')
plt.ylim(0.0, 0.4)

plt.show()
```

(3.2.4) ポアソン分布

・試行回数 n が非常に大きく、事象の発生が極めてまれ（発生確率 p が非常に小さい）で、
その事象が単位時間あたりに起こる回数を確率変数 X としたとき、
X が従う確率分布が「ポアソン分布（ﾎﾟｱｿﾝﾌﾞﾝﾌﾟ、Poisson distribution)」です。

・ポアソン分布は、二項分布において、n→∞、p→0、n*p＝λ(一定)、とした極限での分布に相当します。

・単位時間あたりにある事象が平均してλ回起こる場合に、確率変数 X=k となる
ポアソン分布の確率質量関数 P(X=k) は、以下の式で表されます。

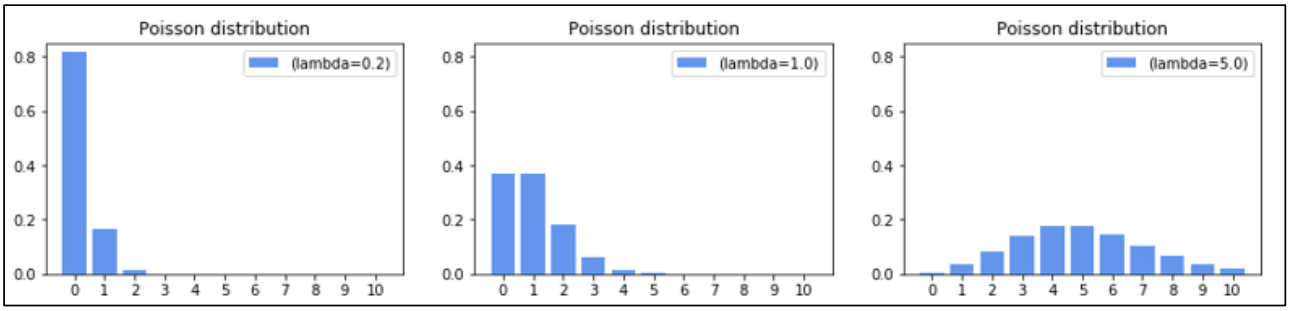
ポアソン分布の確率質量関数

$$P(X=k) = \exp(-\lambda) * \lambda^{k} / k!$$

λ：単位時間当たりの平均発生回数

…(式3.2-4)

(ｸﾞﾗﾌ06-(03)-4_ポアソン分布(λ=0.2/1.0/5.0))



(ﾘｽﾄ06-(03)-4_ポアソン分布(λ=0.2/1.0/5.0))

```
#*************************************************
# ﾘｽﾄ06-(03)-4_ポアソン分布(λ=0.2/1.0/5.0)
#*************************************************
import numpy as np
import math
import matplotlib.pyplot as plt
from mpl_toolkits.mplot3d import Axes3D
%matplotlib inline

# 確率質量関数
def fPoissonDist(lamdval, k):
    return  np.exp(-lamdval) *  (lamdval ** k) /  math.factorial(k)

# パラメータに応じた分布計算
n = 10
lambda1 = 0.2
lambda2 = 1.0
lambda3 = 5.0
Xlist = range(n+1)
Llist = range(n+1)
Plist1 = []
Plist2 = []
Plist3 = []
for x in Xlist:
    Plist1.append( fPoissonDist(lambda1, x) )
for x in Xlist:
    Plist2.append( fPoissonDist(lambda2, x) )
for x in Xlist:
    Plist3.append( fPoissonDist(lambda3, x) )

# グラフ描画
plt.figure(figsize=(15, 3))
```

```
plt.subplot(1, 3, 1)
plt.title('Poisson distribution')
plt.bar(Xlist, Plist1, tick_label=Llist, align="center", ¥
        facecolor="cornflowerblue", label='(lambda=0.2)')
plt.legend(loc='upper right')
plt.ylim(0.0, 0.85)

plt.subplot(1, 3, 2)
plt.title('Poisson distribution')
plt.bar(Xlist, Plist2, tick_label=Llist, align="center", ¥
        facecolor="cornflowerblue", label='(lambda=1.0)')
plt.legend(loc='upper right')
plt.ylim(0.0, 0.85)

plt.subplot(1, 3, 3)
plt.title('Poisson distribution')
plt.bar(Xlist, Plist3, tick_label=Llist, align="center", ¥
        facecolor="cornflowerblue", label='(lambda=5.0)')
plt.legend(loc='upper right')
plt.ylim(0.0, 0.85)

plt.show()
```

(3.2.5) 幾何分布

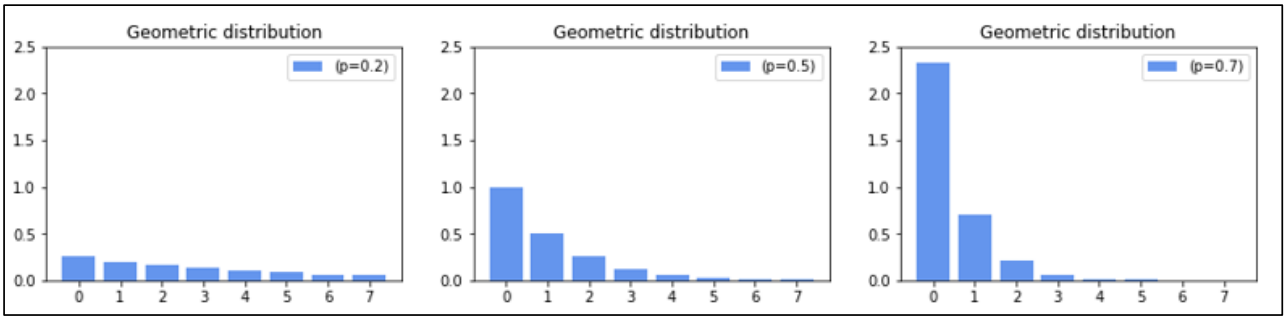
・ベルヌーイ試行を何回か繰り返すときに、初めて成功するまでの回数（1,2,...）を確率変数 X としたとき、X が従う確率分布が「幾何分布（ｷｶﾞﾝﾌﾟ、geometric distribution)」です。

・１回のベルヌーイ試行で成功する確率を p として、確率変数 X=k となる幾何分布の確率質量関数 P(X=k) は以下の式で表されます。

幾何分布の確率質量関数

$$P(X=k) = p\ (1-p)^{k-1} \qquad \cdots(式3.2\text{-}5)$$

(ｸﾞﾗﾌ06-(03)-5_幾何分布(p=0.2/0.5/0.7))



(ﾘｽﾄ06-(03)-5_幾何分布(p=0.2/0.5/0.7))

```
#****************************************************
# ﾘｽﾄ06-(03)-5_幾何分布(p=0.2/0.5/0.7)
#****************************************************
import numpy as np
import math
import matplotlib.pyplot as plt
from mpl_toolkits.mplot3d import Axes3D
%matplotlib inline

# 確率質量関数
def fGeometric(p, k):
    return  p * ((1-p)**(k-1))

# パラメータに応じた分布計算
n = 7
p1 = 0.2
p2 = 0.5
p3 = 0.7
Xlist = range(n+1)
Llist = range(n+1)
Plist1 = []
Plist2 = []
Plist3 = []
for x in Xlist:
    Plist1.append( fGeometric(p1, x) )
for x in Xlist:
    Plist2.append( fGeometric(p2, x) )
for x in Xlist:
    Plist3.append( fGeometric(p3, x) )

# グラフ描画
plt.figure(figsize=(15, 3))

plt.subplot(1, 3, 1)
plt.title('Geometric distribution')
plt.bar(Xlist, Plist1, tick_label=Llist, align="center", ¥
        facecolor="cornflowerblue", label='(p=0.2)')
```

```
plt.legend(loc='upper right')
plt.ylim(0, 2.5)

plt.subplot(1, 3, 2)
plt.title('Geometric distribution')
plt.bar(Xlist, Plist2, tick_label=Llist, align="center", ¥
        facecolor="cornflowerblue", label='(p=0.5)')
plt.legend(loc='upper right')
plt.ylim(0, 2.5)

plt.subplot(1, 3, 3)
plt.title('Geometric distribution')
plt.bar(Xlist, Plist3, tick_label=Llist, align="center", ¥
        facecolor="cornflowerblue", label='(p=0.7)')
plt.legend(loc='upper right')
plt.ylim(0, 2.5)

plt.show()
```

(3.3) 連続型確率分布

・確率変数が連続的である（連続的な値を取る）場合の確率分布を
「連続型確率分布（ﾚﾝｿﾞｸｶﾞﾀｶｸﾘﾂﾌﾞﾝﾌﾟ, continuous probability distribution)」と言います。

・連続型確率変数の各値 x に対してその発生確率f(x) を表す(連続型)確率分布の関数が、
「確率密度関数（ｶｸﾘﾂﾐｯﾄﾞｶﾝｽｳ, probability density function)」です。

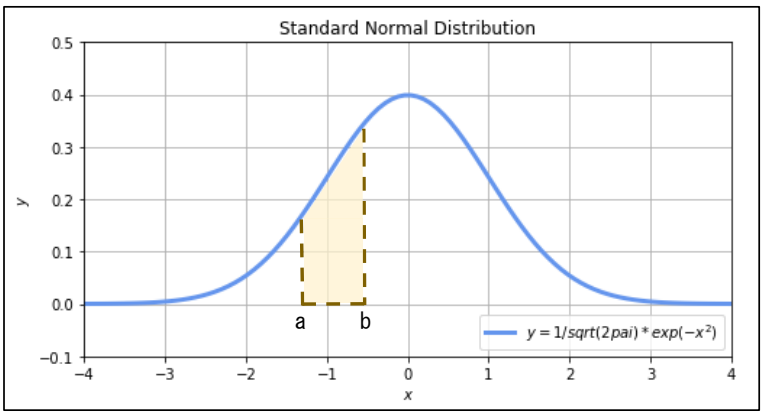
・連続型確率分布では、確率変数の定義域全体に渡る確率密度関数 f(x)の積分は１となります。
（「確率変数の定義域全体に渡る積分は１」というのは、
定義域全体に渡る確率の総和が１になる、つまり全体で100%になる、という意味です。）

$$\int_{-\infty}^{+\infty} f(x)\ dx = 1 \quad (0 \leqq f(x) \leqq 1)$$

・連続型確率分布では、確率変数 X の値の範囲 a≦X≦b に対して、その確率を計算します。

$$P(a \leqq X \leqq b) = \int_{a}^{b} f(x)\ dx$$

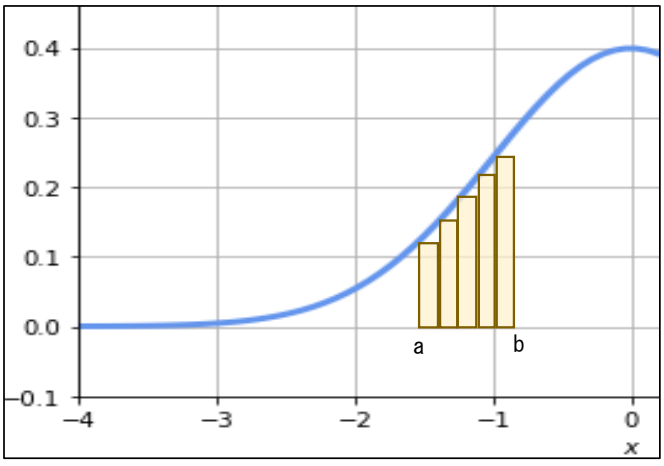
（これは右図の例では、座標軸と
点線とグラフで囲まれた
範囲の面積に相当します）



※ 説明では
「積分記号∫(ｲﾝﾃｸﾞﾗﾙ, integral)」
が出てきます。
直感的には x の関数 f(x) について、
区間 a≦x≦b の x 軸と関数 f(x)
で囲まれた図形の面積であると理解
してください。但し f(x)<0 の部分は
マイナスの面積（引き算）となります。

積分 $\int_{a}^{b} f(x)\ dx$
は、定義域 x の区間 a≦x≦b を N-1 個の区間（x_1=a, x_N=b, 区間幅=Δx_i, i=1～N-1）に分割し、
各区間での関数の値 $f(x_i)$ に区間幅Δx_i を掛けたものの総和を取り、
（つまり $x=x_i$ での関数値 $f(x_i)$ に幅Δx_i を掛け矩形に近似した面積を算出し、その総和をとる）
区間の数を多く(N→∞)、区間幅を小さく($\Delta x_i \to 0$)していった時の極限値が、この積分値です。

$$\int_{a}^{b} f(x)\ dx = \lim_{\Delta x_i \to 0,\, N \to \infty} \sum_{i=1}^{N-1} f(x_i) * \Delta x_i \qquad (x_1=a,\ x_N=b)$$



・以下に、連続型確率分布の幾つかを紹介します。

(3.3.1) 正規分布

・平均値の付近に集積するような、釣り鐘型のデータ分布を表した連続的な確率変数を X としたとき、
X が従う確率分布の一つが「正規分布 (ｾｲｷﾌﾞﾝﾌﾟ、normal distribution)」です。
正規分布は「ガウス分布 (Gaussian distribution)」とも言います。

・X が従う正規分布は、$N(\mu,\ \sigma^2)$ で表現します。
確率変数 X から次式により、μ=0、σ=1 となる標準化確率変数 Z へ変換することができます。

$$Z \quad = (X - \mu) / \sigma$$

この Z が従う確率分布は「標準正規分布」と言い、N(0, 1) となります。
標準正規分布に従う確率変数 Z については、「(3.6) 標準化確率変数」を参照のこと。

・平均値をμ、標準偏差をσとしたときに、正規分布の確率密度関数 f(x) は次式で表されます。

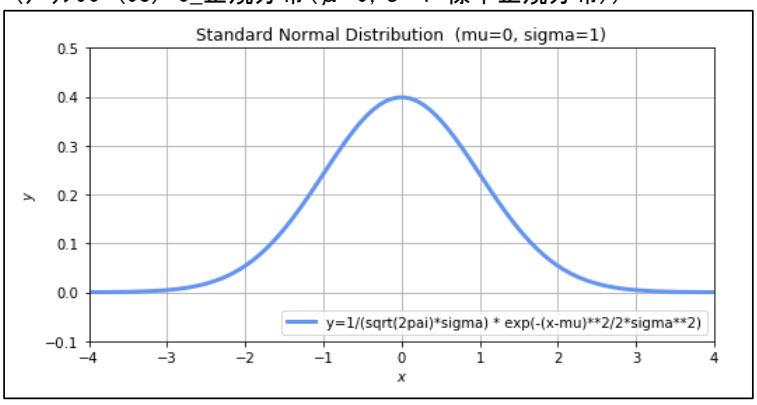
「正規分布 $N(\mu,\ \sigma^2)$」の確率密度関数

$$f(x) = 1/(\sqrt{2\pi}\sigma)\ \exp(-(x-\mu)^2/2\sigma^2) \qquad (-\infty<x<\infty) \qquad \cdots(式3.3\text{-}1)$$

「標準正規分布 N(0, 1)」の確率密度関数

$$f(x) = 1/(\sqrt{2\pi})\ \exp(-x^2/2) \qquad (-\infty<x<\infty) \qquad \cdots(式3.3\text{-}1')$$

(ｸﾞﾗﾌ06-(03)-6_正規分布(μ=0, σ=1 標準正規分布))



(ﾘｽﾄ06-(03)-6_正規分布(μ=0, σ=1 標準正規分布))

```
#*************************************************
# ﾘｽﾄ06-(03)-6_正規分布(μ=0, σ=1 標準正規分布)
#*************************************************
# 標準正規分布
import numpy as np
import matplotlib.pyplot as plt
from mpl_toolkits.mplot3d import Axes3D
%matplotlib inline

# ガウス分布関数
def gauss(x, mu, sigma):
    return (1 / (np.sqrt(2 * np.pi) * sigma))¥
           * np.exp(-(x - mu)**2 / (2 * sigma**2))

# データとグリッド
xn = 100
xlist = np.linspace(-4, 4, xn)
mu = 0
sigma = 1

# グラフ描画
plt.figure(figsize=(8,4))
plt.plot( xlist, gauss(xlist, mu, sigma), 'cornflowerblue', linewidth=3,¥
         label='y=1/(sqrt(2pai)*sigma) * exp(-(x-mu)**2/2*sigma**2)' )
plt.title('Standard Normal Distribution  (mu=0, sigma=1)')
plt.legend(loc='lower right')
```

```
plt.ylim(-0.1, 0.5)
plt.xlim(-4, 4)
plt.xlabel('$x$')
plt.ylabel('$y$')
plt.grid(True)
plt.show()
```

(3.3.2) 対数正規分布

・確率変数 X (>0) の対数 ln(X) (Xの自然対数) が
「正規分布 (ｾｲｷﾌﾞﾝﾌﾟ、normal distribution)」に従う場合、
X が従う確率分布が「対数正規分布 (log normal distribution)」です。

・対数正規分布は、グラフを見てわかるように x の小さい部分に分布が集中し、
x の大きい方に裾野を長く延ばすような分布になっており、
国民の所得分布の近似モデルなどとして使用されます。

・上記のとおり、確率変数 X (>0) の自然対数 ln(X) の平均値をμ、標準偏差をσとしたときに、
新たな Y=ln(X) という確率変数 Y は、正規分布 N(μ, σ^2) に従います。
X が従う対数正規分布は、Λ(μ, σ^2) で表現します。

・特に、確率変数 X (>0) の対数 ln(X) が従う確率分布が標準正規分布 N(0, 1) の場合、
X が従う分布を「標準対数正規分布」と言い、Λ(0, 1)で表現します。

・対数正規分布の確率密度関数 f(x) は次式で表されます。

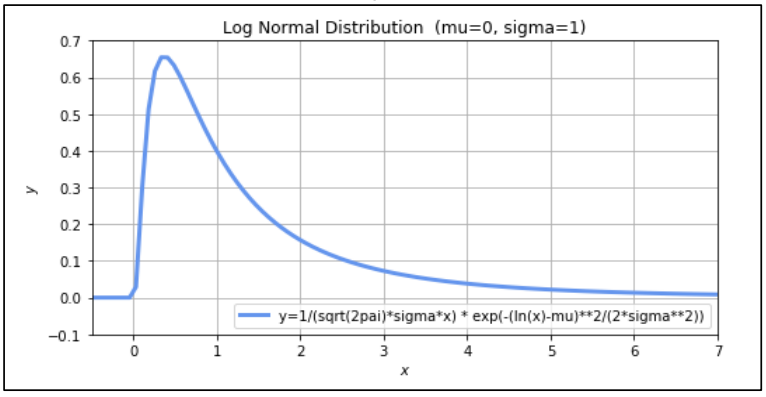
対数正規分布の確率密度関数

$$f(x) = 1/(\sqrt{2\pi}\sigma x)\ \exp(-(\ln(x)-\mu)^2/2\sigma^2) \quad (0<x<\infty) \quad \cdots(式3.3\text{-}2)$$
$$= 0 \quad (x \leqq 0)$$

対数正規分布の諸統計量

自然対数 ln(X) の平均値	:	μ
自然対数 ln(X) の標準偏差	:	σ
最頻値	:	$e^{\mu-\sigma**2}$
中央値	:	e^{μ}
平均値	:	$e^{\mu+\sigma**2/2}$
分散	:	$e^{2\mu+\sigma**2}\ (e^{\sigma**2}-1)$

(ｸﾞﾗﾌ06-(03)-7_対数正規分布(μ=0, σ=1 対数正規分布))



(ﾘｽﾄ06-(03)-7_対数正規分布(μ=0, σ=1 対数正規分布))

```
#************************************************
# ﾘｽﾄ06-(03)-7_対数正規分布(μ=0, σ=1)
#************************************************
# 対数正規分布
import numpy as np
import matplotlib.pyplot as plt
from mpl_toolkits.mplot3d import Axes3D
%matplotlib inline

# 対数正規分布関数
def logNormalDistribution(xlist, mu, sigma):
    ylist = []
    for x in xlist:
        if x <= 0:
            ylist.append(0.0)
        else:
            ylist.append(
                1 / (np.sqrt(2 * np.pi) * sigma * x)¥
                * np.exp(-(np.log(x) - mu)**2 / (2 * sigma**2)) )
    return ylist

# データとグリッド
xn = 100
xlist = np.linspace(-0.5, 7, xn)
mu = 0
sigma = 1

# グラフ描画
plt.figure(figsize=(8,4))
plt.plot( xlist, logNormalDistribution(xlist, mu, sigma), 'cornflowerblue', linewidth=3,¥
         label='y=1/(sqrt(2pai)*sigma*x) * exp(-(ln(x)-mu)**2/(2*sigma**2))' )
plt.title('Log Normal Distribution  (mu=0, sigma=1)')
plt.legend(loc='lower right')
plt.ylim(-0.1, 0.7)
plt.xlim(-0.5, 7)
plt.xlabel('$x$')
plt.ylabel('$y$')
plt.grid(True)
plt.show()
```

【出典・参考】

対数正規分布⇒ https://sci-fx.net/math-log-norm-dist/

対数正規分布の諸統計量⇒ https://ja.wikipedia.org/w/index.php?curid=2320607

(3.3.3) 指数分布

・機械が故障してから次に故障するまでの期間や、災害が起こってから次に起こるまでの期間のように、
着目している事象が次に起こるまでの期間を確率変数 X としたとき、
X が従う確率分布が「指数分布 (ｼｽｳﾌﾞﾝﾌﾟ、exponential distribution)」です。

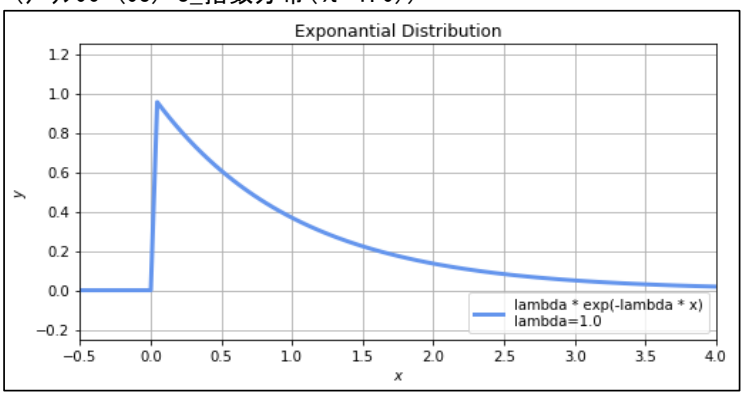
・ある期間に平均してλ(ラムダ) 回起こる事象について、次に同事象が起こるまでの期間が x となる
指数分布の確率密度関数 f(x) は次の式で表されます。

指数分布の確率密度関数

$$f(x) = \lambda \exp(-\lambda x) \quad \text{for } x \geqq 0 \quad \lambda \text{は単位期間内の事象の平均発生回数}$$

$$= 0 \quad \text{for } x < 0 \qquad \cdots(\text{式3.3-3})$$

(ｸﾞﾗﾌ06-(03)-8_指数分布(λ=1.0))



(ﾘｽﾄ06-(03)-8_指数分布(λ=1.0))

```
#****************************************************
# ﾘｽﾄ06-(03)-8_指数分布(λ=1.0)
#****************************************************
import numpy as np
import matplotlib.pyplot as plt
from mpl_toolkits.mplot3d import Axes3D
%matplotlib inline

# 指数分布関数
def shisuBunpu(xarray, lambdaval):
    ylist = []
    for x in xarray :
        if x <= 0 :
            ylist.append(0)
        else:
            ylist.append(lambdaval * np.exp(-lambdaval * x))
    return ylist

# データとグリッド
xn = 100
xlist = np.linspace(-0.5, 4, xn)
lambdaval1 = 1.0
ylist1 = shisuBunpu(xlist, lambdaval1)

# グラフ描画
plt.figure(figsize=(8,4))
plt.plot( xlist, ylist1, 'cornflowerblue', linewidth=3, label='lambda * exp(-lambda * x)¥nlambda=1.0' )
plt.title('Exponantial Distribution')
plt.legend(loc='lower right')
plt.ylim(-0.25, 1.25)
plt.xlim(-0.5, 4)
plt.xlabel('$x$')
plt.ylabel('$y$')
plt.grid(True)
plt.show()
```

(3.3.4) 連続一様分布

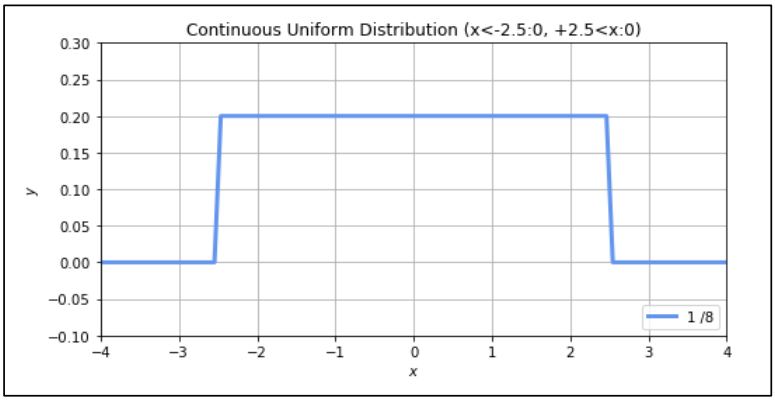
・定義域内の任意の値に対し発生確率が等しい連続的な確率変数を X としたとき、
X が従う確率分布が「連続一様分布 (ﾚﾝｿﾞｸｲﾁﾖｳﾌﾞﾝﾌﾟ、continuous uniform distribution)」です。

・確率変数 X の範囲が a≦ X ≦b (つまり[a, b]) の時、
連続一様分布の確率密度関数 f(x) は次式で表されます。

連続一様分布の確率密度関数

$$f(x) = 1/(b-a) \quad \text{for } a \leqq x \leqq b$$
$$= 0 \quad \text{for } x < a,\ b < x \qquad \cdots(式3.3\text{-}4)$$

(ｸﾞﾗﾌ06-(03)-9_連続一様分布 (非0区間[-2.5,2.5]))



(ﾘｽﾄ06-(03)-9_連続一様分布 (非0区間[-2.5,2.5]))

```
#****************************************************
# ﾘｽﾄ06-(03)-9_連続一様分布
#****************************************************
import numpy as np
import matplotlib.pyplot as plt
from mpl_toolkits.mplot3d import Axes3D
%matplotlib inline

# 確率密度関数
def fContUniformDist(xList, rangeStart, rangeEnd):
    yList = []
    hConst = 1 / (rangeEnd - rangeStart)
    for x in xList:
        if (x < rangeStart) or (rangeEnd < x):
            yList.append(0.0)
        else:
            yList.append(hConst)
    return yList

# データとグリッド
Xn = 100
xList = np.linspace(-4, 4, Xn)

# グラフ描画
plt.figure(figsize=(8,4))
plt.plot( xList, fContUniformDist(xList, -2.5, 2.5), 'cornflowerblue', linewidth=3, label='1/5' )
plt.title('Continuous Uniform Distribution (x<-2.5:0, +2.5<x:0)')
plt.legend(loc='lower right')
plt.ylim(-0.1, 0.3)
plt.xlim(-4, 4)
plt.xlabel('$x$')
plt.ylabel('$y$')
plt.grid(True)
plt.show()
```

(3.4) 平均と確率変数の期待値

・平均については「(2) 基本的な統計指標」で述べていますが、
ここでは、確率変数の観点から焼き直しして述べます。

・「平均 (ﾍｲｷﾝ、mean value)」は分布の中心の尺度であり、
観測値の総和を観測値の個数(観測度数)で割ったもので、μ(ﾐｭｰ)で表します。

平均

$$\mu = \sum_{i=1}^{N} x_i / N$$

x_i：各観測値（i=1,2,…,N）、N：観測度数

…(式3.4-1)

・確率変数 X の取り得る値にその発生確率 P(X) を掛けた値の総和を「期待値 (ｷﾀｲﾁ、expected value)」
と言い、E(X) で表します。
これは、確率変数のすべての値 X に確率 P(X) の重みをつけた加重平均となっています。

期待値

・離散型確率分布の期待値

$$E(X) = \sum_{i=1}^{N} x_i * P(x_i)$$

$P(x_i)$：確率質量関数　…(式3.4-2)

・連続型確率分布の期待値

$$E(X) = \int_{-\infty}^{+\infty} x * f(x)\, dx$$

$f(x)$：確率密度関数　…(式3.4-3)

・平均値が観測値全体の和を観測度数で割った値を指すのに対し、
期待値は、１回の観測で期待される値のことを指します。

・離散型確率分布の場合、数式的には、
平均μは、離散一様分布(式3.2-1)で発生確率 $P(x_i)$ を 1/N で置き換えた期待値に等しくなります。

$$E(X) = \sum_{i=1}^{N} x_i * P(x_i) = \sum_{i=1}^{N} x_i * (1/N) = \mu$$

・以下に期待値についての公式を幾つか掲載します。

期待値についての公式

(1) 定数 a の期待値

$$E(a) = a$$

a:定数値　…(式3.4-4)

(2) 確率変数 X の定数倍の期待値

$$E(bX) = b*E(X)$$

b:定数値　…(式3.4-5)

(3) 確率変数 X と定数 a との和の期待値

$$E(a + X) = a + E(X)$$

a:定数値　…(式3.4-6)

(4) 確率変数 X の定数倍と定数 a との和の期待値

$$E(a + bX) = a + b*E(X)$$

a,b:定数値　…(式3.4-7)

(5) 確率変数 X についての２次式の期待値

$$E(a + bX + cX^2) = a + b*E(X) + c*E(X^2)$$

a,b,c:定数値　…(式3.4-8)

【出典・参考】
平均⇒ https://ja.wikipedia.org/wiki/平均
平均⇒「初等統計学」ﾊﾞﾀﾁｬﾘﾔ、ｼﾞｮﾝｿﾝ著 箕谷千鳳彦訳 東京図書 1980年

(3.5) 確率変数の分散

・分散についても「(2) 基本的な統計指標」で述べていますが、
ここでは、確率変数の観点から焼き直しして述べます。

・「分散 (ﾌﾞﾝｻﾝ、variance)」は分布のばらつきの尺度です。
分散は観測値の偏差の二乗和を観測値の個数(観測度数)で割ったもので、
σ^2 (ｼｸﾞﾏの2乗)または V(X) (ﾌﾞｲ)で記述します。
分散の平方根を「標準偏差 (ﾋｮｳｼﾞｭﾝﾍﾝｻ, standard deviation)」と呼び、σ (ｼｸﾞﾏ)で記述します。

分散

$$\sigma^2 = (1/N) \sum_{i=1}^{N} (x_i - \mu)^2 \quad \cdots(式3.5\text{-}1)$$

x_i : 各観測値 (i=1,2,…,N)
N : 観測度数
μ : 平均値

標準偏差

$$\sigma = \sqrt{(1/N) \sum_{i=1}^{N} (x_i - \mu)^2} \quad \cdots(式3.5\text{-}2)$$

・確率変数 X の「分散 (ﾌﾞﾝｻﾝ, variance)」は、
確率変数 X の分布が平均値(期待値) μ からどれだけ散らばっているかを示すもので、
確率変数 X の観測値 x_i のμからの偏差 $(x_i - \mu)$ の二乗和の平均を取ったものです。
これは、偏差の二乗の期待値となっています。

分散

・離散型確率分布の分散

$$V(X) = \sum_{i=1}^{N} (x_i - \mu)^2 * P(x_i) \quad \cdots(式3.5\text{-}3)$$

$P(x_i)$: 確率質量関数

$$= E((X - \mu)^2) \quad \cdots(式3.5\text{-}4)$$

μ : 平均値

・連続型確率分布の分散

$$V(X) = \int_{-\infty}^{+\infty} (x - \mu)^2 * f(x)\, dx \quad \cdots(式3.5\text{-}5)$$

f(x) : 確率密度関数
μ : 平均値

※ (式3.5-1)は、(式3.5-3)で、離散一様分布に従う場合 ($P(x_i) = 1/N$) の分散に相当します。

・以下に分散・標準偏差についての公式を幾つか掲載します。

分散についての公式

(1) 定数 a の分散・標準偏差

$$V(a) = 0 \quad \cdots(式3.5\text{-}6)$$

$$\sigma(a) = 0 \quad \cdots(式3.5\text{-}7)$$

a:定数値

(2) 確率変数 X の定数倍の分散・標準偏差

$$V(bX) = b^2 * V(X) \quad \cdots(式3.5\text{-}8)$$

$$\sigma(bX) = |b| * \sigma(X) \quad \cdots(式3.5\text{-}9)$$

b:定数値

(3) 確率変数 X と定数 a との和の分散・標準偏差

$$V(a + X) = V(X) \quad \cdots(式3.5\text{-}10)$$

$$\sigma(a + X) = \sigma(X) \quad \cdots(式3.5\text{-}11)$$

a:定数値

(4) 確率変数 X の定数倍と定数 a との和の分散・標準偏差

$$V(a + bX) = b^2 * V(X) \quad \cdots(式3.5\text{-}12)$$

$$\sigma(a + bX) = |b| * \sigma(X) \quad \cdots(式3.5\text{-}13)$$

a,b:定数値

【出典・参考】
分散⇒ https://ja.wikipedia.org/wiki/分散_(確率論)
分散⇒ http://kabblog.net/2257/
分散⇒「初等統計学」ﾊﾞﾀﾁｬﾘﾔ、ｼﾞｮﾝｿﾝ著 箕谷千鳳彦訳 東京図書 1980年

(3.6) 標準化確率変数

・既に述べた平均と分散についての公式を用いて、
平均μ、標準偏差σの確率変数 X を、
平均０、標準偏差１の確率変数 Z へ変換することが出来ます。

・この Z を「標準化確率変数（ﾋｮｳｼﾞｭﾝｶｶｸﾘﾂﾍﾝｽｳ, standardized random variable)」と言い、
確率変数 X に対し次の変換により Z を与えます：

標準化確率変数

$$Z = (X - \mu) / \sigma$$

μ ：平均値　　…(式3.6-1)
σ ：標準偏差

（標準化されることの検証）

$$\begin{aligned}
\text{期待値}\quad E(Z) &= E((X-\mu)/\sigma) \\
&= E(X/\sigma) - E(\mu/\sigma) \\
&= 1/\sigma * E(X) - \mu/\sigma \\
&= 1/\sigma * \mu - \mu/\sigma \\
&= 0 \\
\text{分散}\quad V(Z) &= V((X-\mu)/\sigma) \\
&= 1/\sigma^2 * V(X-\mu) \\
&= 1/\sigma^2 * V(X) \\
&= 1/\sigma^2 * \sigma^2 \\
&= 1
\end{aligned}$$

・平均値がμ、標準偏差がσの正規分布に従う確率変数 X に対し、標準化確率変数 Z が従う確率分布は
「標準正規分布（ﾋｮｳｼﾞｭﾝｾｲｷﾌﾞﾝﾌﾟ, Standard Normal Distribution)」と言い、
次のような確率密度関数になります：

確率変数X が従う確率密度関数(正規分布)

$$f(x) = 1/(\sqrt{2\pi}\sigma)\ \exp(-(x-\mu)^2/2\sigma^2)$$

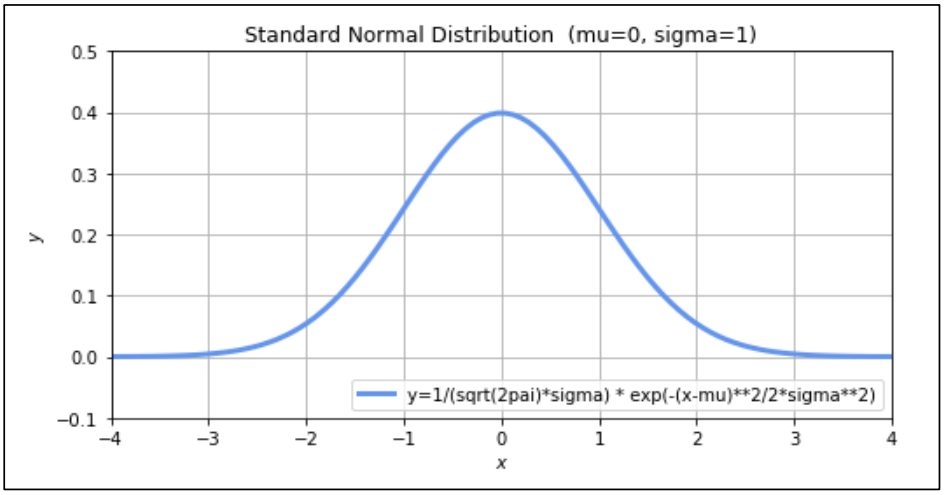
$(-\infty<x<\infty)$　　…(式3.6-2)
μ ：平均値
σ ：標準偏差

標準化確率変数 Z が従う確率密度関数(標準正規分布)

$$f(z) = 1/(\sqrt{2\pi})\ \exp(-z^2/2)$$

$(-\infty<z<\infty)$　　…(式3.6-3)

（標準正規分布）



【出典・参考】
標準化確率変数⇒「初等統計学」ﾊﾞﾀﾁｬﾘﾔ、ｼﾞｮﾝｿﾝ著 箕谷千鳳彦訳 東京図書 1980年
標準化確率変数⇒ https://k-san.link/standardized/
標準正規分布表⇒ https://www.koka.ac.jp/morigiwa/sjs/standard_normal_distribution.htm

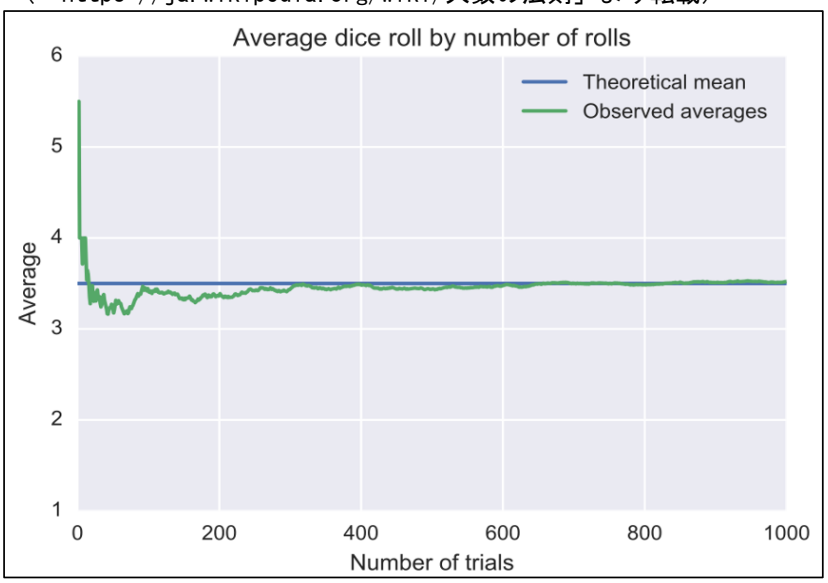
(4) 基本定理

確率論・統計学における基本定理を幾つか記します。

(4.1) 大数の法則

・事象Ａの発生確率をｐとします。
事象Ａについてのｎ回の独立試行で、事象Ａが起こった回数をｒとします。
試行回数ｎが非常に大きくなると、相対度数　ｒ／ｎ　は、事象の発生確率ｐに限りなく近づきます。
これを「大数の法則（ﾀｲｽｳﾉﾎｳｿｸ、Law of Large Numbers、LLN）」と呼び、
確率論・統計学における基本定理の一つとなっています。

（例）大数の法則の例
「サイコロを何度も繰り返し振り続けると、出た目の標本平均は平均に収束する」という例
（「https://ja.wikipedia.org/wiki/大数の法則」より転載）



(4.2) 中心極限定理

・平均μ、標準偏差σを持つ**正規分布の母集団**からの、大きさｎの無作為標本において、
標本平均　x_{mean} の分布は、平均μ、標準偏差$\sigma/\sqrt{n}$　の正規分布になります。
特に、標本平均　x_{mean} の標準化確率変数 Z は正規分布に従います：

$$Z = (x_{mean} - \mu) / (\sigma/\sqrt{n})$$

としたとき、 Z の確率分布 ＝N（0, 1）

・一方、平均μ、標準偏差σを持つ**任意の母集団**からの、大きさｎの無作為標本において、
標本平均　x_{mean} の分布は、各標本の大きさｎが大きい時、
近似的に平均μ、標準偏差$\sigma/\sqrt{n}$　の正規分布になります。
特に、標本平均　x_{mean} の標準化確率変数 Z は、近似的に正規分布に従います：

$$Z = (x_{mean} - \mu) / (\sigma/\sqrt{n})$$

としたとき、 Z の確率分布 ≒ N（0, 1）

これを「中心極限定理（ﾁｭｳｼﾝｷｮｸｹﾞﾝﾃｲﾘ、central limit theorem、CLT）」と呼び、
確率論・統計学における基本定理の一つとなっています。

【出典・参考】
大数の法則⇒ https://ja.wikipedia.org/wiki/大数の法則
大数の法則⇒ 「数学公式事典」 黒田孝朗・須田貞之著 文研出版 1978年
中心極限定理⇒ https://ja.wikipedia.org/wiki/中心極限定理
中心極限定理⇒「初等統計学」ﾊﾞﾀﾁｬﾘﾔ、ｼﾞｮﾝｿﾝ著 箕谷千鳳彦訳 東京図書 1980年

(5) 相関分析

複数の確率変数間の関係性を知るために、複数の確率変数を同時に観測します。
以下では、２つの確率変数 X, Y とその関係性の指標について記述します。

(5.1) 散布図と相関関係

・２つの要素 X, Y からなる一組のデータが得られたときに、
２つの要素の関係を見るために、X を横軸に、Y を縦軸に取り、各データを該当する位置にプロットしたグラフを「散布図 (ｻﾝﾌﾟｽﾞ, scatter plot)」といいます。
データを散布図で表すと、X が変化したときに、Y がどのように変化するかが確認できます。

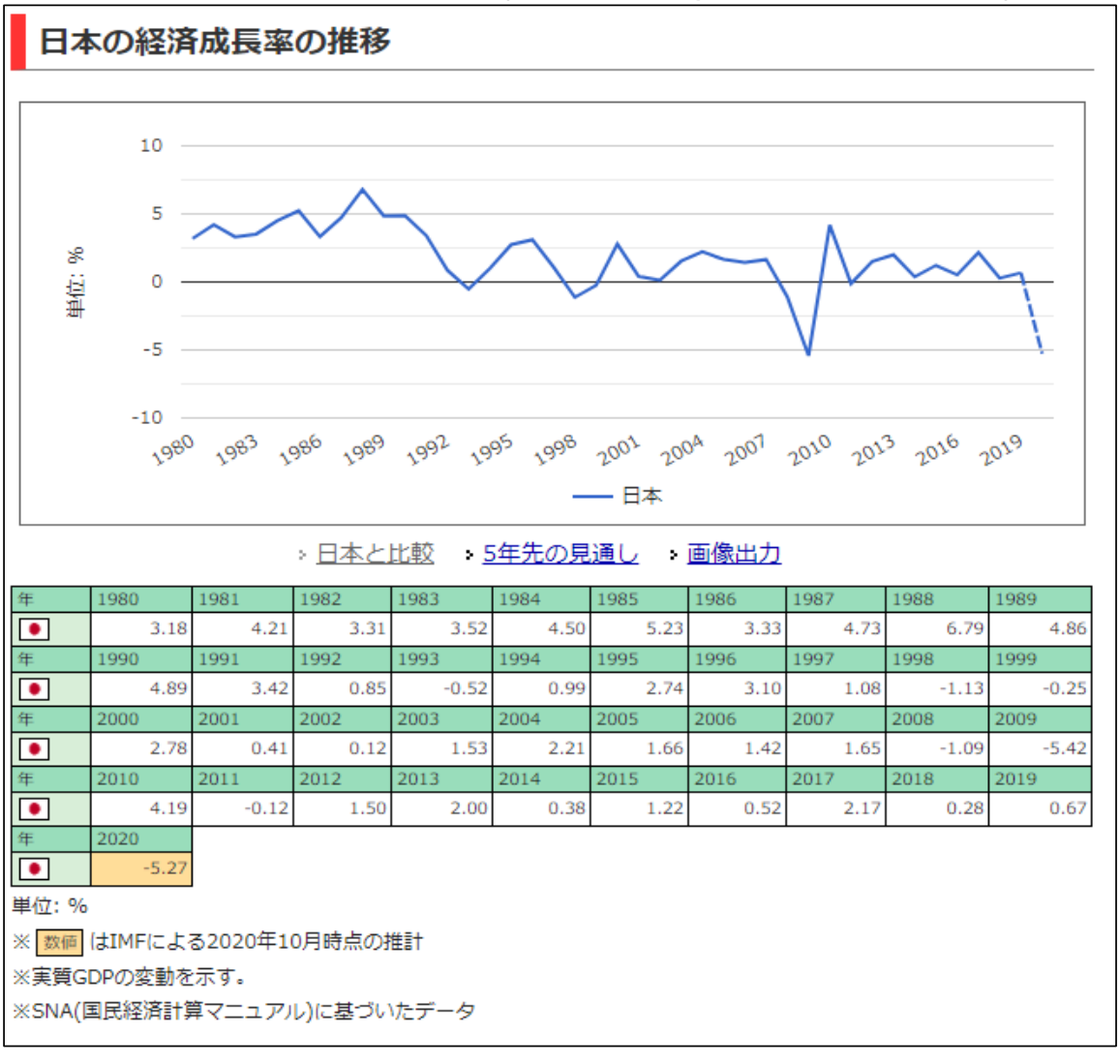
・２つの要素の間に何らかの関係がある場合、
これらのデータ間には「相関関係 (ｿｳｶﾝｶﾝｹｲ, correlation)」があるといいます。
一方が増加すれば他方も増加する時、「正の相関関係 (positive correlation)」があるといい、
一方が増加すれば他方が減少する時、「負の相関関係 (negative correlation)」があると言います。

(例5.1) 完全失業率(全国)と経済成長率の散布図と時系列

・「先進諸国の失業率と実質経済成長率は強い負の相関関係にある」という記事がありましたので、
日本の 2000年～2019年について、総務省が公開している「完全失業率(全国)」と、
ＩＭＦが公開している「実質経済成長率」で、表と散布図・時系列を描いてみます。

実質経済成長率

(出典)「日本の経済成長率の推移」⇒ https://ecodb.net/country/JP/imf_growth.html
⇒ https://www.imf.org/en/Countries/JPN#countrydata



年	1980	1981	1982	1983	1984	1985	1986	1987	1988	1989
●	3.18	4.21	3.31	3.52	4.50	5.23	3.33	4.73	6.79	4.86
年	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999
●	4.89	3.42	0.85	-0.52	0.99	2.74	3.10	1.08	-1.13	-0.25
年	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
●	2.78	0.41	0.12	1.53	2.21	1.66	1.42	1.65	-1.09	-5.42
年	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019
●	4.19	-0.12	1.50	2.00	0.38	1.22	0.52	2.17	0.28	0.67
年	2020									
●	-5.27									

単位: %
※ 数値 はIMFによる2020年10月時点の推計
※実質GDPの変動を示す。
※SNA(国民経済計算マニュアル)に基づいたデータ

完全失業率(全国)

(出典)「e-Stat 統計で見る日本」⇒ https://www.e-stat.go.jp/dbview?sid=0003008332

「完全失業率」とは「15歳以上の働く意欲のある人（労働力人口）のうち、
仕事を探しても仕事に就くことのできない人（完全失業者）の割合」を言い、
総務省の労働力調査により毎月発表されるものです。

表章項目 率 【%】　産業 全産業　就業状態 完全失業者　地域 全国

	総数	男	女
2000年	4.7	4.9	4.5
2001年	5.0	5.2	4.7
2002年	5.4	5.5	5.1
2003年	5.3	5.5	4.9
2004年	4.7	4.9	4.4
2005年	4.4	4.6	4.2
2006年	4.1	4.3	3.9
2007年	3.9	3.9	3.7
2008年	4.0	4.1	3.8
2009年	5.1	5.3	4.8
2010年	5.1	5.4	4.6
2011年	…	…	…
2012年	4.3	4.6	4.0
2013年	4.0	4.3	3.7
2014年	3.6	3.7	3.4
2015年	3.4	3.6	3.1
2016年	3.1	3.3	2.8
2017年	2.8	3.0	2.7
2018年	2.4	2.6	2.2
2019年	2.4	2.5	2.2
2020年	2.8	3.0	2.5

・データ数が20個と少なく、大雑把な判断にならざるを得ないですが、
散布図からは、「失業率と実質経済成長率は強い負の相関関係にある」とは判断が付きかねます。

年	完全失業率 %	実質経済成長率 %
2000	4.7	2.78
2001	5.0	0.41
2002	5.4	0.12
2003	5.3	1.53
2004	4.7	2.21
2005	4.4	1.66
2006	4.1	1.42
2007	3.9	1.65
2008	4.0	-1.09
2009	5.1	-5.42
2010	5.1	4.19
2011	4.5	-0.12
2012	4.3	1.50
2013	4.0	2.00
2014	3.6	0.38
2015	3.4	1.22
2016	3.1	0.52
2017	2.8	2.17
2018	2.4	0.28
2019	2.4	0.67

・時系列を見ると、実質経済成長率の下がった 2001～2002年頃に、
完全失業率が 5%前後にやや上がっています。

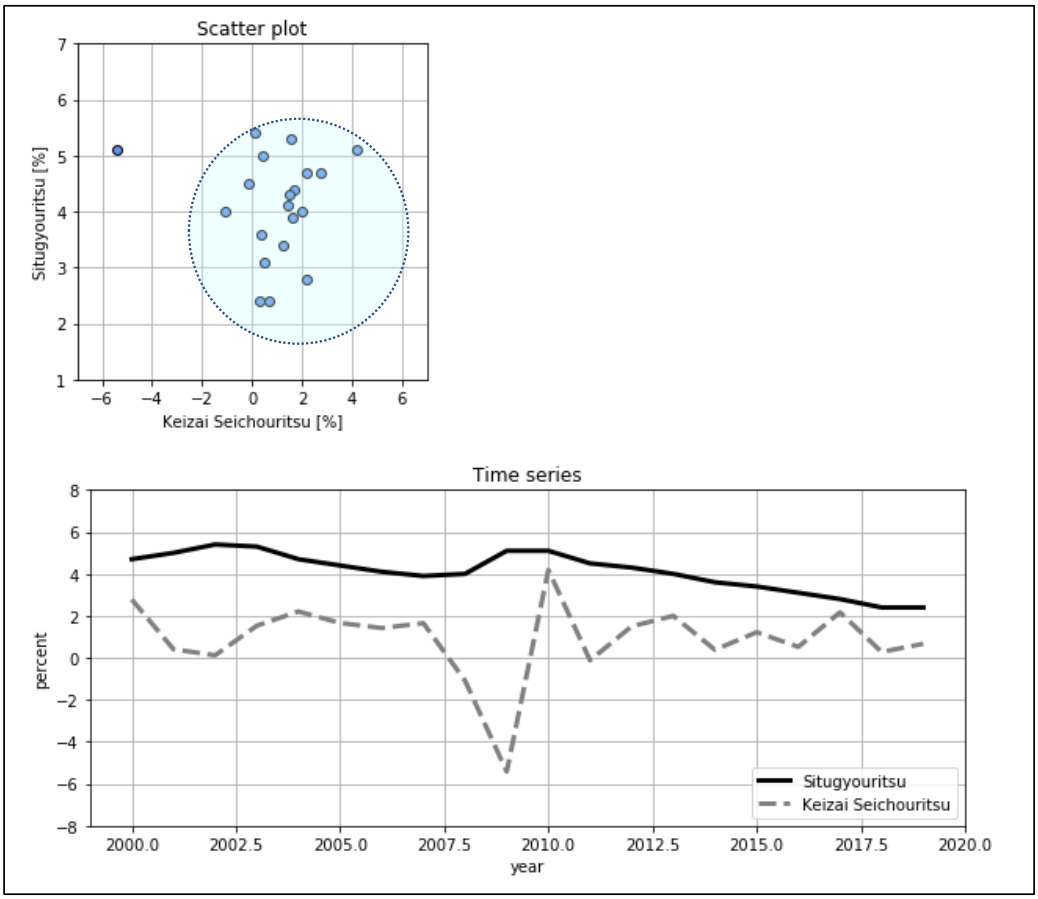
・同様に 2008～2009年頃のリーマンショックで実質経済成長率の
下がった時にも、完全失業率が 5%前後にやや上がっています。

・逆に2010年頃はリーマンショックの反動の立ち直りで、
実質経済成長率が上がっていて、完全失業率がやや下がっています。

・上記の傾向からも、概ね
「失業率と実質経済成長率は負の相関関係にある」と言えそうですが、
データ数も少なく、この期間の経済成長率はずっと低い状態にあり、
散布図を見た限りでは、「負の相関」と判断することはできません。

・ちなみに相関係数を(5.2)節で計算しますが約 -0.05 で、
負ではあるもののOに近く、「負の相関」というより
「相関が無い」と言ってもおかしくない値であることが分かります。

（※ 2011年は東日本最震災の影響でデータが整っておらず、
2011年03月～08月を除いた平均値です。
他の年は01月～12月の平均値です。
（出典：総務省統計局提供労働力調査　長期時系列データ））



(ﾘｽﾄ06-(05)-1_散布図と相関関係)

```
#************************************************
# ﾘｽﾄ06-(05)-1_散布図と相関関係
#************************************************
import numpy as np
import matplotlib.pyplot as plt

# 年
NEN_list = [ 2000,2001,2002,2003,2004,2005,2006,2007,2008,2009,
             2010,2011,2012,2013,2014,2015,2016,2017,2018,2019 ]

# 経済成長率(日本)
X_list = [ 2.78, 0.41, 0.12, 1.53, 2.21, 1.66, 1.42, 1.65, -1.09, -5.42,
           4.19, -0.12, 1.50, 2.00, 0.38, 1.22, 0.52, 2.17, 0.28, 0.67 ]
X_min = -7
X_max = 7

# 完全失業率(全国)
Y_list = [ 4.7, 5.0, 5.4, 5.3, 4.7, 4.4, 4.1, 3.9, 4.0, 5.1,
           5.1, 4.5, 4.3, 4.0, 3.6, 3.4, 3.1, 2.8, 2.4, 2.4 ]
Y_min = 1
Y_max = 7

# 散布図表示 ------------------------------
plt.figure(figsize=(4,4))
plt.plot(X_list, Y_list, marker='o', linestyle='None', markeredgecolor='black', color='cornflowerblue')
plt.xlim(X_min, X_max)
plt.ylim(Y_min, Y_max)
plt.title('Scatter plot')
plt.xlabel( 'Keizai Seichouritsu [%]' )
plt.ylabel( 'Situgyouritsu [%]' )
plt.grid(True)
plt.show()

# 時系列表示 ------------------------------
plt.figure(figsize=(10,4))
plt.plot( NEN_list, Y_list, color='black', linewidth=3, label='Situgyouritsu' )
plt.plot( NEN_list, X_list, color='gray', linewidth=3, linestyle='--', label='Keizai Seichouritsu' )
plt.title('Time series')
plt.legend(loc='lower right')
plt.ylim(X_min-1, X_max+1)
plt.xlim(NEN_list[0]-1, NEN_list[19]+1)
plt.xlabel('year')
plt.ylabel('percent')
plt.grid(True)
plt.show()
```

(5.2) ２つの確率変数の確率分布

・１つの確率変数とその確率分布については、既に「(3) 確率変数と確率分布」で見てきました。
これを２つの確率変数 X、Y に拡張し、その確率分布について見ます。

・確率変数 X がとる値を $\{ x_1, x_2, \cdots, x_k \}$ の k 個とし、
確率変数 Y がとる値を $\{ y_1, y_2, \cdots, y_l \}$ の l 個とすると、
確率変数の組合せ (X,Y) は、(x_i, y_j) i=1～k, j=1～l の k×l 個あります。

・離散型の確率分布で、$f(x_i, y_j)$ を $(X=x_i, Y=y_j)$ の値を同時に取る確率分布関数とします。
これを「結合確率分布 (ｹﾂｺﾞｳｶｸﾘﾂﾌﾞﾝﾌﾟ, joint propability distribution)」と言います。
（同時確率分布、同時分布とも言います）

結合確率分布関数

$$f(x_i, y_j) = P(X=x_i, Y=y_j) \qquad i=1\sim k,\ j=1\sim l \qquad \cdots(式5.2\text{-}1)$$

これにより、確率変数 X, Y の結合確率分布は以下の様に２次元の表で表現できます：

X ＼ Y		Y の値				Xの周辺分布
		y_1	y_2	…	y_l	
Xの値	x_1	$f(x_1, y_1)$	$f(x_1, y_2)$	…	$f(x_1, y_l)$	$f_x(X=x_1)$
	x_2	$f(x_2, y_1)$	$f(x_2, y_2)$	…	$f(x_2, y_l)$	$f_x(X=x_2)$
	…	…	…	…	…	…
	x_k	$f(x_k, y_1)$	$f(x_k, y_2)$	…	$f(x_k, y_l)$	$f_x(X=x_k)$
Yの周辺分布		$f_y(Y=y_1)$	$f_y(Y=y_2)$	…	$f_y(Y=y_l)$	(合計) 1

・確率変数 X(または Y) のみに着目した確率分布 $f_x(X)$ (または $f_y(Y)$) は、
X(または Y)の「周辺分布 (ｼｭｳﾍﾝﾌﾞﾝﾌﾟ, marginal distribution)」と言い、次式で求められます：

X の周辺分布 f_x、Yの周辺分布 f_y

$$f_x(X=x_i) = \sum_{j=1}^{l} f(x_i, y_j) \qquad i=1\sim k \qquad \cdots(式5.2\text{-}2)$$

$$f_y(Y=y_j) = \sum_{i=1}^{k} f(x_i, y_j) \qquad j=1\sim l \qquad \cdots(式5.2\text{-}3)$$

・確率変数 X(または Y) のみに着目した平均と分散・標準偏差等の統計指標は、
結合確率分布へ遡ることなく、X(または Y)の周辺分布のみから計算することが出来ます。

確率変数 X、Y の期待値

$$E_x(X) = \mu_x = \sum_{i=1}^{k} x_i * f_x(X=x_i) \qquad \cdots(式5.2\text{-}4)$$

$$E_y(Y) = \mu_y = \sum_{j=1}^{l} y_j * f_y(Y=y_j) \qquad \cdots(式5.2\text{-}5)$$

確率変数 X、Y の分散

$$V_x(X) = \sigma_x^2 = \sum_{i=1}^{k} (x_i-\mu_x)^2 * f_x(X=x_i) \qquad \cdots(式5.2\text{-}6)$$

$$= E_x\left((X-\mu_x)^2 \right)$$

$$V_y(Y) = \sigma_y^2 = \sum_{j=1}^{l} (y_j-\mu_y)^2 * f_y(Y=y_j) \qquad \cdots(式5.2\text{-}7)$$

$$= E_y\left((Y-\mu_y)^2 \right)$$

$f_x(x_i)$：Xの周辺確率分布関数
$f_y(y_j)$：Yの周辺確率分布関数

(5.3) 共分散と相関係数

・２つの確率変数 X、Y の間の関係の強さを表す尺度として、「共分散（ｷｮｳﾌﾞﾝｻﾝ、covariance)」があります。これは、積 $(X-\mu_x)*(Y-\mu_y)$ の期待値として定義されます。
本セミナーでは共分散を「Cov(X,Y)」で記述します。

確率変数 X、Y の共分散「Cov(X,Y)」

$$Cov(X,Y) = E((X-\mu_x)*(Y-\mu_y)) \quad \cdots(式5.3\text{-}1)$$

$$= E(XY) - \mu_x*\mu_y \quad \cdots(式5.3\text{-}2)$$

μ_x：確率変数 X の期待値
μ_y：確率変数 Y の期待値

（証明）
$$E((X-\mu_x)*(Y-\mu_y))$$
$$= E(XY - \mu_xY - \mu_yX + \mu_x\mu_y)$$
$$= E(XY) - \mu_xE(Y) - \mu_yE(X) + \mu_x\mu_y$$
$$= E(XY) - \mu_x\mu_y$$

・共分散 Cov(X,Y) は、各確率変数 X と Y の測定単位に依存しています。
測定単位に依存しない２変数間の尺度は、共分散を各変数の標準偏差 σ_x、σ_y で割って求めます。
これによって得られる尺度は X と Y の「相関係数（ｿｳｶﾝｹｲｽｳ, correlation coefficient)」と呼ばれます。本セミナーでは相関係数を「Corr(X,Y)」で記述します。

確率変数 X、Y の相関係数「Corr(X,Y)」

$$Corr(X,Y) = Cov(X,Y) / \sigma_x*\sigma_y \quad \cdots(式5.3\text{-}3)$$

σ_x：確率変数 X の標準偏差
σ_y：確率変数 Y の標準偏差

・以下に、X と Y の相関係数 Corr(X,Y) の性質を幾つか見てみましょう。

X、Y の相関係数 Corr(X,Y) の性質

(1) 相関係数 Corr(X,Y) は必ず -1 以上 1 以下の間の数である。

$$-1.0 \leqq Corr(X,Y) \leqq 1.0$$

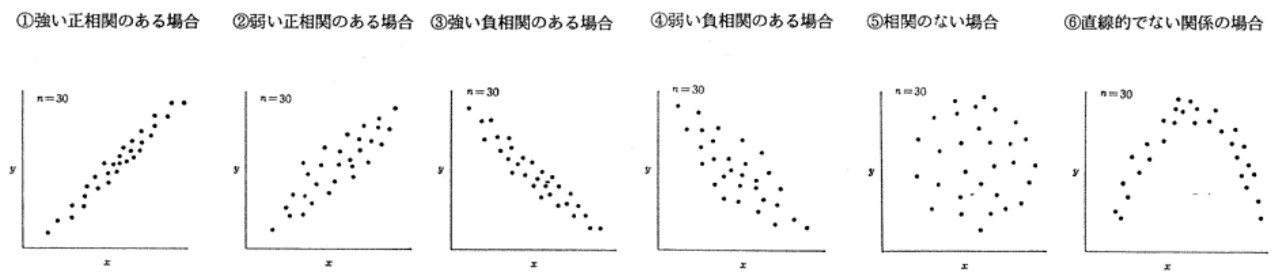
Corr(X,Y)＝+1.0 となるのは、X と Y が正の勾配を持つ直線関係にある場合である。
Corr(X,Y)＞ 0.0 となるのは、X と Y の間に正の相関関係がある場合である。
Corr(X,Y)＝ 0.0 となるのは、X と Y が無関係（独立）の場合である。
Corr(X,Y)≒ 0.0 となるのは、X と Y の間の相関関係が弱い場合である。
Corr(X,Y)＜ 0.0 となるのは、X と Y の間に負の相関関係がある場合である。
Corr(X,Y)＝-1.0 となるのは、X と Y が負の勾配を持つ直線関係にある場合である。

(2) 相関係数 Corr(X,Y) は確率変数 X, Y に定数が加えられても、同符号の定数が確率変数に掛けられても不変である。

$$Corr(X,Y) = Corr(a_xX+b_x,\ a_yY+b_y) \quad \cdots(式5.3\text{-}4)$$

a_x, a_y は同符号
a_x, a_y, b_x, b_y：定数値

・X と Y の相関を散布図で図示すると以下の様なものになります。
（引用元「https://matome.naver.jp/odai/2136538459573533501/2147046772966211003」）



(例5.2) 完全失業率(全国)と経済成長率の相関係数

・「(例5.1) 完全失業率(全国)と経済成長率の散布図と時系列」について、
その相関係数を計算してみましょう。

2000年～2019年 (標本数=20)
完全失業率(全国平均) 平均＝4.11, 標準偏差＝0.90
経済成長率(日本)　　平均＝0.90, 標準偏差＝1.84
完全失業率(全国平均)と経済成長率(日本)の相関係数(Numpy　使用) ＝-0.05207
完全失業率(全国平均)と経済成長率(日本)の相関係数(Numpy 不使用) ＝-0.05207

・完全失業率(全国平均)と経済成長率(日本)の相関係数は、約 -0.05 で負ではあるもののOに近く、
「負の相関」というより「相関が無い」と言ってもおかしくない値であることが分かります。

(ﾘｽﾄ06-(05)-2_相関係数)

```
#**************************************************
# ﾘｽﾄ06-(05)-2_相関係数
#**************************************************
import numpy as np

# 相関係数の計算 (numpy 使用版)
def fCorr1(xlist, ylist):
    xycorr = np.corrcoef(xlist, ylist)[0, 1]
    return xycorr

# 相関係数の計算 (numpy 不使用版で、式のとおりに計算)
def fCorr2(xlist, ylist):
    xm = np.mean(xlist)
    ym = np.mean(ylist)
    xs= np.std(xlist)
    ys= np.std(ylist)
    n = len(xlist)
    exy = 0
    for i in range(n):
        exy += xlist[i]*ylist[i]/n
    covxy = exy - (xm * ym)
    xycorr = covxy / (xs * ys)
    return xycorr

# (2000～2019) 経済成長率(日本)
X_list = [ 2.78, 0.41, 0.12, 1.53, 2.21, 1.66, 1.42, 1.65, -1.09, -5.42,
           4.19, -0.12, 1.50, 2.00, 0.38, 1.22, 0.52, 2.17, 0.28, 0.67 ]

# (2000～2019) 完全失業率(全国)
Y_list = [ 4.7, 5.0, 5.4, 5.3, 4.7, 4.4, 4.1, 3.9, 4.0, 5.1,
           5.1, 4.5, 4.3, 4.0, 3.6, 3.4, 3.1, 2.8, 2.4, 2.4 ]

# 標準偏差と相関係数を計算
X_mean = np.mean(X_list)
Y_mean = np.mean(Y_list)
X_std= np.std(X_list)
Y_std= np.std(Y_list)
xycorr1 = fCorr1(X_list, Y_list)
xycorr2 = fCorr2(X_list, Y_list)

print(f"2000年～2019年 (標本数={len(Y_list)!r})")
print(f"  完全失業率(全国平均) 平均＝{Y_mean:.2f}, 標準偏差＝{Y_std:.2f}")
print(f"  経済成長率(日本)　　平均＝{X_mean:.2f}, 標準偏差＝{X_std:.2f}")
print(f"  完全失業率(全国平均)と経済成長率(日本)の相関係数(Numpy　使用) ＝{xycorr1:.5f}")
print(f"  完全失業率(全国平均)と経済成長率(日本)の相関係数(Numpy 不使用) ＝{xycorr2:.5f}")
```

(5.4) 説明変数と相関関係

・何かの因果関係を探る場合、統計モデルでは、原因となる変数群を元に、結果となる変数群を説明しようと試みます。
原因となる変数のことを「説明変数（ｾﾂﾒｲﾍﾝｽｳ、explanatory variable)」と言います。
「独立変数（ﾄﾞｸﾘﾂﾍﾝｽｳ、independent variable)」、
「予測変数（ﾖｿｸﾍﾝｽｳ、predictor variable)」とも言います。
結果となる変数のことを「目的変数（ﾓｸﾃｷﾍﾝｽｳ、response variable)」と言います。
「従属変数（ｼﾞｭｳｿﾞｸﾍﾝｽｳ、dependent variable)」、
「結果変数（ｹｯｶﾍﾝｽｳ、outcome variable)」とも言います。

・説明変数を選ぶ際に気を付ける必要があるのは、多重共線性の問題です。
「多重共線性（ﾀｼﾞｭｳｷｮｳｾﾝｾｲ、multicollinearity)」（略称「ﾏﾙﾁｺ」）とは、
モデル内の一部の説明変数と他の説明変数の相関係数が高いときに起こる状態です。
多重共線性によって、重回帰分析では、回帰係数の分散を増加させて不安定にするため、
正しく推計できなくなるといった悪影響をもたらします。
この問題の最も一般的な解消法は、「相関関係が高いと考えられる説明変数を外すこと」です。

(例5.3) コンビニの月間のアイスクリームの売り上げ
・目的変数：(1) コンビニの月間のアイスクリームの売り上げ
・説明変数：(1) 来客数
(2) 最高気温が30℃以上の日数
(3) 降雨日数
(4) 月間降水量
(5) 平均気温
という分析モデルを考えたとき、説明変数で「(3) 降雨日数」と「(4) 月間降水量」は
相関が高く、多重共線性が発生する可能性が高いとみられます。
相関係数を計算して、(3)か(4)の何れか一方を説明変数として不採用とするか判断します。

【出典・参考】

散布図⇒ https://bellcurve.jp/statistics/glossary/7409.html
散布図⇒ https://ja.wikipedia.org/wiki/散布図
相関関係⇒ https://kotobank.jp/word/相関関係-5304
相関係数⇒ http://mt-net.vis.ne.jp/ADFE_mail/0208.htm
相関係数⇒ https://ja.wikipedia.org/wiki/相関係数
「e-Stat 統計で見る日本」⇒ https://www.e-stat.go.jp/dbview?sid=0003008332
「総務省統計局提供労働力調査」⇒ https://www.stat.go.jp/data/roudou/longtime/03roudou.html
「日本の経済成長率の推移」⇒ https://ecodb.net/country/JP/imf_growth.html
⇒ https://www.imf.org/en/Countries/JPN#countrydata
結合確率分布⇒ http://www.geisya.or.jp/~mwm48961/kou3/prob_joint1.htm
統計用語の英訳⇒ http://www.cottonpot01.com/JpnEng/EngJpnSta120160825.pdf
同時分布⇒ https://ja.wikipedia.org/wiki/同時分布
周辺分布⇒ http://ibisforest.org/index.php?%E5%91%A8%E8%BE%BA%E5%88%86%E5%B8%83
結合分布⇒「初等統計学」ﾊﾞﾀﾁｬﾘﾔ、ｼﾞｮﾝｿﾝ著 箕谷千鳳彦訳 東京図書 1980年
共分散⇒ https://bellcurve.jp/statistics/glossary/914.html
相関係数⇒「初等統計学」ﾊﾞﾀﾁｬﾘﾔ、ｼﾞｮﾝｿﾝ著 箕谷千鳳彦訳 東京図書 1980年
散布図⇒ https://matome.naver.jp/odai/2136538459573533501/2147046772966211003
説明変数、目的変数⇒ https://bellcurve.jp/statistics/course/1590.html
目的変数⇒ https://bellcurve.jp/statistics/glossary/551.html
説明変数⇒ https://bellcurve.jp/statistics/glossary/2109.html
多重共線性⇒ https://xica.net/vno4ul5p/
多重共線性⇒ https://support.minitab.com/ja-jp/minitab/18/help-and-how-to/modeling-statistics/regression/supporting-topics/model-assumptions/multicollinearity-in-regression/

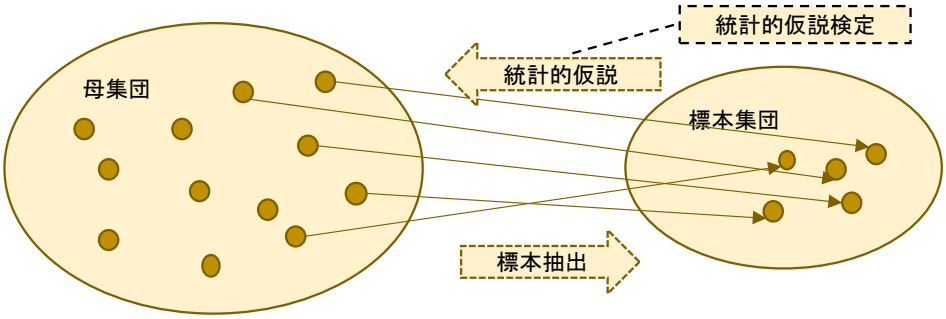
(6) 仮説検定

(6.1) 統計的仮説

・母集団についての仮説を「統計的仮説（ﾄｳｹｲﾃｷｶｾﾂ、statistical hypothesis)」と呼びます。
統計的仮説には、「帰無仮説」と帰無仮説の逆の内容を持つ「対立仮説」があります。
標本から何らかの主張を立証しようとしている時、主張自体を
「対立仮説（ﾀｲﾘﾂｶｾﾂ、alternative hypothesis)」と呼び、H_1 と記します。
主張を否定する(その主張は誤りであり、元々問題なく、主張は無に帰するという)仮説を
「帰無仮説（ｷﾑｶｾﾂ、Null hypothesis)」と呼び、H_0 と記します。

・対立仮説は、帰無仮説が「棄却（ｷｷｬｸ、reject)」された場合に「採択（ｻｲﾀｸ、accept)」されます。
この時、検定は「統計的に有意（ﾄｳｹｲﾃｷﾆﾕｳｲ、statistically significant)」であると言います。

・帰無仮説 H_0 と対立仮説 H_1 の設定についての指針があります。
帰無仮説 H_0 が実際に偽である時、それを棄却しないという過誤(「第二種の過誤」と言います)よりも、
帰無仮説 H_0 が実際に真である時、それを棄却するという過誤(「第一種の過誤」と言います)
の方を重要視します。
従って、帰無仮説 H_0 を対立仮説 H_1 に対して検定する時、
ﾃﾞｰﾀが H_0 に対して強く不利な証拠を示していない限り、H_0 は真であるとして支持し、
ﾃﾞｰﾀが H_0 に対して強く不利な証拠を示している場合には、H_1 を支持して H_0 を棄却します。
帰無仮説 H_0 と対立仮説 H_1 は、こういった指針に基づいて設定します。



(例6.1) サイコロの３の目が不公平に出るかどうかの仮説
・サイコロの３の目が出る確率を p としたとき
帰無仮説 H_0： $p=1/6$ である
対立仮説 H_1： $p\neq1/6$ である

(例6.2) ドラマの視聴率が７％を超えたかどうかの仮説
・ドラマの視聴率を p としたとき
帰無仮説 H_0： $p\leqq0.07$ である
対立仮説 H_1： $p>0.07$ である

(例6.3) 数学の試験でＡ高校とＢ高校の平均点が等しくないかどうかの仮説
・Ａ高校とＢ高校の数学の試験の平均点を各々 μ_A、μ_B としたとき
帰無仮説 H_0： $\mu_A=\mu_B$ である
対立仮説 H_1： $\mu_A\neq\mu_B$ である

【出典・参考】
仮説検定⇒ https://ja.wikipedia.org/wiki/仮説検定
仮説検定⇒「初等統計学」ﾊﾞﾀﾁｬﾘﾔ、ｼﾞｮﾝｿﾝ著 箕谷千鳳彦訳 東京図書 1980年
仮説検定⇒「統計学」森棟,照井,中川,西埜,黒住 共著 有斐閣 2017年12月 改訂版第3刷
有意⇒ https://ja.wikipedia.org/wiki/有意

(6.2) 統計的仮説検定

・統計的仮説の妥当性を、標本から検証することを　「統計的仮説検定 (ﾄｳｹｲﾃｷｶｾﾂｹﾝﾃｲ、statistical hypothesis testing)」、あるいは単に「検定」と呼びます。

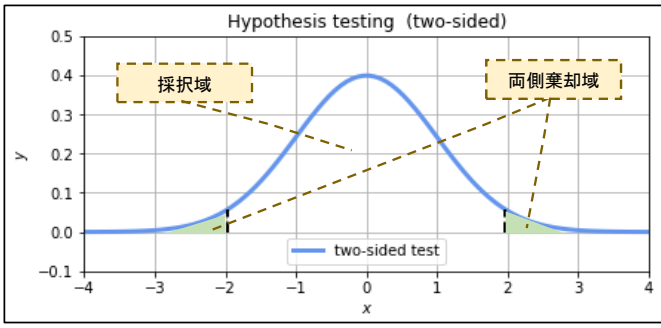
・統計的仮説検定に用いられる確率変数を「検定統計量 (ｹﾝﾃｲﾄｳｹｲﾘｮｳ、test statistic)」、あるいは、単に「統計量」と呼びます。
検定統計量は、標本を元に計算するので、標本統計量と同様に分布を持ちます。
帰無仮説の下で導かれた検定統計量の分布を元にして、検定を行います。

・検定で、帰無仮説 H_0 が棄却されると判断する検定統計量の値の範囲は、
帰無仮説の下で導かれた検定統計量の分布の裾（両裾または片裾）に設定され、
「棄却域 (ｷｷｬｸｲｷ,rejection region)」と呼びます。

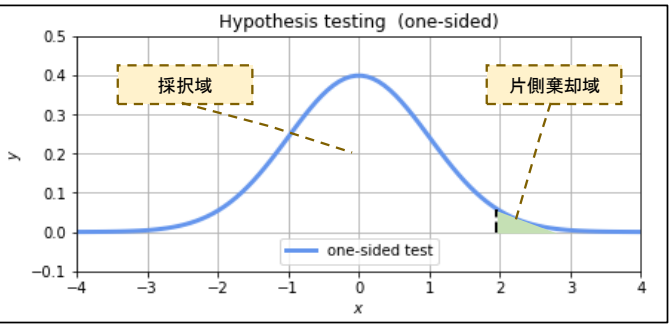
・棄却域以外の領域は帰無仮説 H_0 が棄却されない領域であり、
「採択域 (ｻｲﾀｸｲｷ, acceptance region)」または「受容域」と呼びます。
棄却域と採択域の境目を「臨界値 (ﾘﾝｶｲﾁ, critical value)」または「境界値」と呼びます。

・検定統計量の分布で、棄却域の面積(＝確率)を「有意水準 (ﾕｳｲｽｲｼﾞｭﾝ、significance level)」と呼び、αで表現します。αは通常、0.05(5%)、0.01(1%) といった小さな確率に定められます。
有意水準αと「第一種の過誤」が生じる確率は同じです。

(例6.4) 両側検定の棄却域の例



(例6.5) 片側検定の棄却域の例



・(例6.4) 両側検定で $X \leqq -1.96\sigma$、$1.96\sigma \leqq X$ を棄却域とする（$\alpha = 0.05(5\%)$）。
・(例6.5) 片側検定で　　$1.96\sigma \leqq X$ を棄却域とする（$\alpha = 0.025(2.5\%)$）。
参照⇒(ﾘｽﾄ06-(06)-1_仮説検定_両側検定_片側検定)

・統計的仮説の設定に従って、検定は両側検定と片側検定に分類されます。

・統計量がある閾値よりも大きい（あるいは小さい）かを判定するような対立仮説を
「片側対立仮説 (ｶﾀｶﾞﾜﾀｲﾘﾂｶｾﾂ、one-sided alternative hypothesis)」と言います。
片側対立仮説では、棄却域が統計量の分布の片裾にあるため、
この棄却域を「片側棄却域 (ｶﾀｶﾞﾜｷｷｬｸｲｷ、one-sided rejection region)」と言い、
この検定を「片側検定 (ｶﾀｶﾞﾜｹﾝﾃｲ、one-sided test)」と言います。

・統計量がある値と一致するかを判定するような対立仮説を
「両側対立仮説 (ﾘｮｳｶﾞﾜﾀｲﾘﾂｶｾﾂ、two-sided alternative hypothesis)」と言います。
両側対立仮説では、棄却域が統計量の分布の両裾にあるため、
この棄却域を「両側棄却域 (ﾘｮｳｶﾞﾜｷｷｬｸｲｷ、two-sided rejection region)」と言い、
この検定を「両側検定 (ﾘｮｳｶﾞﾜｹﾝﾃｲ、two-sided test)」と言います。

・観察されたデータを用いて計算された検定統計量が、棄却域に入る時は、帰無仮説 H_0 を棄却します。
検定統計量が、棄却域に入らない時は、帰無仮説 H_0 を棄却できません。

(6.3) 検定の手順

・これまでに述べた概念を元に検定の手順をまとめると、以下の様になります：

検定の手順

(手順1) 帰無仮説 H_0 と対立仮説 H_1 を設定する。
(手順2) 検定統計量を定める。
(手順3) 有意水準αを定める。
(手順4) 帰無仮説 H_0 の下での検定統計量の分布に基づき、
有意水準αに対応する棄却域を定める。
(手順5) 観察されたデータを用いて検定統計量を計算する。
(手順6) 計算された検定統計量が、棄却域に入る時は、帰無仮説 H_0 を棄却できると判定する。
検定統計量が、棄却域に入らない時は、帰無仮説 H_0 を棄却できないと判定する。

(例6.6) サイコロの３の目が不公平に出るかどうかの検定
（これは「(例6.1) サイコロの３の目が不公平に出るかどうかの仮説」の検定になります）

・元になる考え方
１回のサイコロを振る試行で、３の目が出る回数(0, 1 の何れか)を確率変数 X としたとき、
「サイコロの３の目が出る場合成功(X=1)、それ以外の目が出れば失敗(X=0)」という
ベルヌーイ試行として扱います(「(3.2.2) ベルヌーイ分布」を参照のこと)。
従って、サイコロを１回振って３の目が出る確率を p としたとき、
確率変数 X の平均 E(X)＝p、分散 V(X)＝p(1-p) で与えられます。

サイコロが公平な場合、以下の様な統計指標が得られ、
これが帰無仮説(p＝1/6)の下での、母集団の値になります。
確率変数 X の平均 E(X)＝p＝1/6
確率変数 X の分散 V(X)＝p(1-p)＝(1/6)*(5/6)

・観察されたデータ
500回サイコロを振ったところ、100回３の目が出ました。

(手順1) サイコロを１回振って３の目が出る確率を p としたとき
帰無仮説：p＝1/6 である
対立仮説：p≠1/6 である

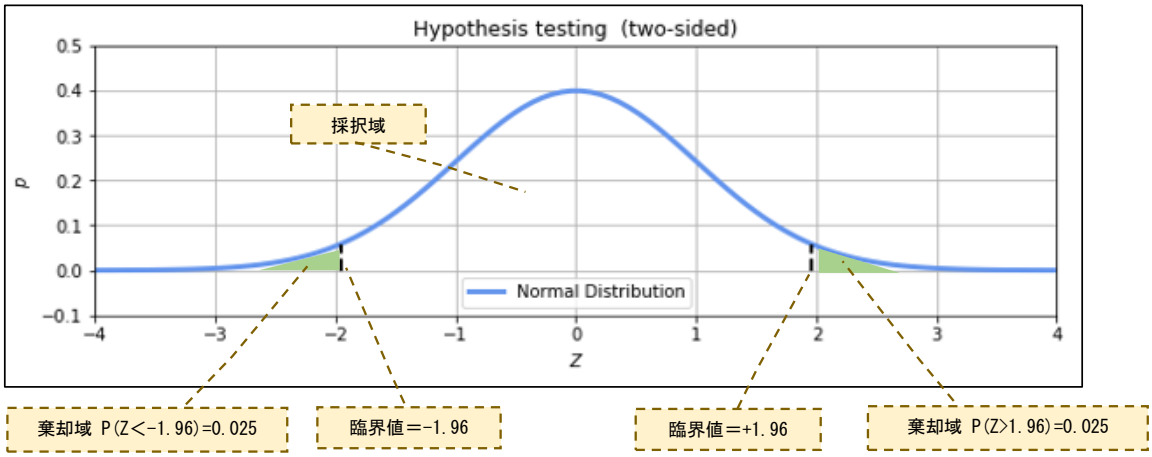
(手順2) 検定統計量 T として、サイコロを１回振って３の目が出る確率(標本確率)を選ぶ：
T ＝ ３の目が出た回数／サイコロを振った回数
標本確率が 1/6 に近ければ、帰無仮説 H_0 を棄却できないと判定します。

(手順3、4) 検定統計量 T は標本平均なので、標本平均の性質により、
標本平均の期待値＝ E(T) ＝ 母平均 E(X) ＝μ＝ 1/6
標本平均の分散＝ V(T) ＝ 母分散 V(X)／標本数n ＝ (1/6)*(5/6)／500
となり、これを以下の式により標準化確率変数 Z へ変換します
$Z = (T - E(T)) / \sqrt{V(T)}$

更に、観測個数(n＝500) は十分に大きいので、中心極限定理により、
標準化確率変数 Z の分布は標準正規分布 N(0,1) で近似できます。

近似した標準正規分布 N(0,1) の下で、有意水準 α ＝5%となるような棄却域は、
標準正規分布 N(0,1) で 1.96＜Z の確率 P(1.96＜Z)＝2.5%であることを用いると、
以下の様な範囲になります：

棄却域：　Z ＜-1.96、+1.96＜ Z
P(|Z|＞1.96) ＝ P(Z＜-1.96) ＋ P(1.96＜Z) ＝ 0.025 + 0.025 ＝ 0.05 (5%) ＝ α



(手順5) 観察されたデータを用いて各値を計算すると以下の様になります。
検定統計量 T ＝ ３の目が出た回数／サイコロを振った回数 ＝ 100／500
標本平均の期待値＝ E(T) ＝ 母平均 E(X) ＝μ＝ 1/6
標本平均の分散＝ V(T) ＝ 母分散 V(X)／標本数n ＝ (1/6)*(5/6)／500

これを元に、検定統計量 T を標準化確率変数 Z へ変換すると以下の様になります。

$$
\begin{aligned}
Z &= (T - E(T)) / \sqrt{V(T)} \\
&= (100/500 - 1/6) / \sqrt{(1/6)*(5/6)/500} \\
&= (1/5 - 1/6) / \sqrt{1/3600} \\
&= 2.0
\end{aligned}
$$

(手順6) 計算された検定統計量（標準化確率変数 Z）は、
Z ＝ 2.0 ＞ 1.96
となり、右側の棄却域に入っています。
従って、帰無仮説 H_0「p＝1/6 である」が棄却されました。
「サイコロの３の目が公平に出る」という仮説は棄却された、という判定結果となりました。

【出典・参考】

統計検定量⇒ https://bellcurve.jp/statistics/course/9317.html
仮説検定⇒「初等統計学」ﾊﾞﾀﾁｬﾘﾔ、ｼﾞｮﾝｿﾝ著 箕谷千鳳彦訳 東京図書 1980年
仮説検定⇒「統計学」森棟,照井,中川,西埜,黒住 共著 有斐閣 2017年12月 改訂版第3刷
棄却域・採択域⇒ https://bellcurve.jp/statistics/course/9317.html

(6.4) Ｐ値

・「Ｐ値（ﾋﾟｰﾁ、P-value)」は、観測された検定統計量 T の値が t_r の時、
帰無仮説 H_0 での検定統計量の分布の下で、
$T=t_r$ を臨界値とした場合の棄却域の面積（検定統計量 T が t_r より棄却域側にある割合）です。

・両側仮説検定の場合	Ｐ値 ＝ P(|T| ＞ t_r)
・右側の片側仮説検定の場合	Ｐ値 ＝ P(T ＞ t_r)
・左側の片側仮説検定の場合	Ｐ値 ＝ P(T ＜ t_r)

・Ｐ値は、「帰無仮説 H_0 での検定統計量の分布の下で、
観測された検定統計量以上に、(帰無仮説に反する)偏った検定統計量が得られる確率」を示しています。
Ｐ値が有意水準αを下回ったときに、はじめて「統計的有意差があった」と言うことができます。

(例6.7)「(例6.6) サイコロの３の目が不公平に出るかどうかの検定」でのＰ値

(例6.6) では、

検定統計量 T ＝ ３の目が出た回数／サイコロを振った回数 ＝ 100／500

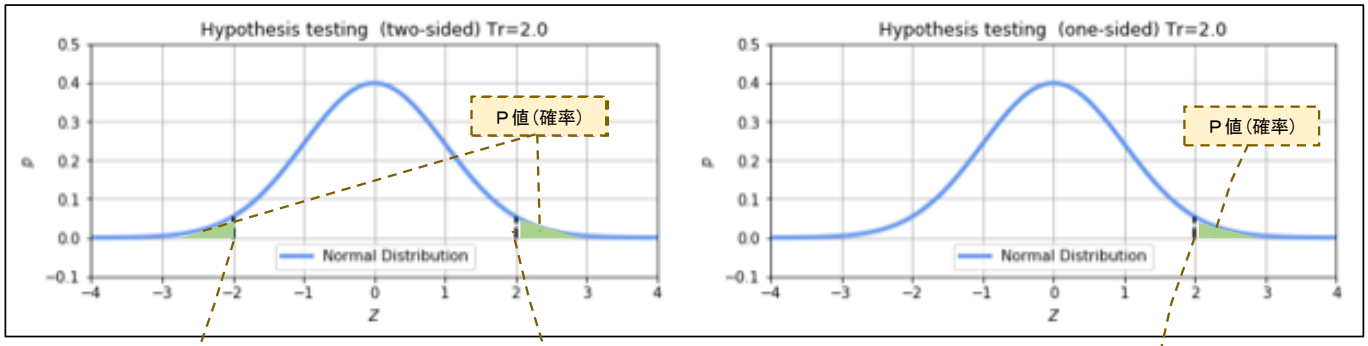
で、T を標準化検定統計量 Z へ変換したものは

標準化検定統計量 Z ＝ 2.0

でした。この例で、Ｐ値は

Ｐ値 ＝ P(|Z| ＞ 2.0) ＝ 0.0456

となっていて、有意水準αの 5% を下回っており、
「統計的有意差があった」と言うことができます。



参照⇒(ﾘｽﾄ06-(06)-2_仮説検定とP値)

【出典・参考】

Ｐ値⇒ https://haru-reha.com/p-value-mean/
Ｐ値⇒ https://xica.net/magellan/marketing-idea/stats/tvalue-and-pvalue/
Ｐ値、有意水準⇒ https://bellcurve.jp/statistics/blog/14004.html
Ｐ値、有意水準⇒ https://ja.wikipedia.org/wiki/有意#p値
Ｐ値⇒「統計学」森棟,照井,中川,西埜,黒住 共著 有斐閣 2017年12月 改訂版第3刷
標準正規分布表⇒ https://www.koka.ac.jp/morigiwa/sjs/standard_normal_distribution.htm

(6.5) z検定、t検定

・上記では主に、正規分布を用いて検定の説明をしてきました。
検定には、標本の大きさ、仮定する分布や既知項目などによって様々な検定があります。
以下では、z検定、t検定について述べます。

(6.5.1) z検定

「z検定 (ｾﾞｯﾄｹﾝﾃｲ、z-test)」は、
・母集団が、正規分布に従うと仮定されたデータに対して用います。
・母分散 σ^2 が既知であるときに用います。
・標本の大きさが大きい場合（≧30程度）に用います。
・この時、以下の標本平均についての(標準化された)検定統計量 z は、
標準正規分布 N(0,1) に従い、検定ではこの分布を用います。

z検定での検定統計量 z

$$z = (x_{mean} - \mu_0) / (\sigma/\sqrt{n})$$ …(式6.5-1)

μ_0：帰無仮説による母平均
x_{mean}：標本の平均値
σ：母標準偏差
n：標本の大きさ

(6.5.2) t検定

「t検定 (ﾃｨｰｹﾝﾃｲ, t-test)」は、
・母集団が、正規分布に従うと仮定されたデータに対して用います。
・母分散 σ^2 が未知であるときに用い、その推定値として標本不偏分散 u^2 を使います。
・標本の大きさが小さい場合（＜30程度）に用います。
・この時、以下の標本平均についての検定統計量 t (t値)は、
自由度n-1の「t分布 (ﾃｨｰﾌﾞﾝﾌﾟ、t-distribution)」に従い、検定ではこの分布を用います。

t検定での検定統計量 t

$$t = (x_{mean} - \mu_0) / (u/\sqrt{n})$$ …(式6.5-2)

μ_0：帰無仮説による母平均
x_{mean}：標本の平均値
u：標本不偏分散の平方根
n：標本の大きさ

標本不偏分散

$$u^2 = (1/(n-1)) \sum_{i=1}^{n} (x_i - x_{mean})^2$$ …(式6.5-3)
…(式2.4-5)再掲

（尚、t分布は自由度 n が大きくなるにつれ、標準正規分布の形に近づいていきます。）

【出典・参考】
z検定、t検定⇒ http://www.geisya.or.jp/~mwm48961/linear_algebra/t_test2.htm
z検定、t検定⇒ https://qiita.com/Haji-_/items/5e3c0a7f2108ae882bff
z検定⇒ https://to-kei.net/hypothesis-testing/z-test/
t検定⇒ https://to-kei.net/hypothesis-testing/t-test/
t分布⇒ https://to-kei.net/distribution/t-distribution/t-distribution/

(ﾘｽﾄ06-(06)-1_仮説検定_両側検定_片側検定)

```
#************************************************
# ﾘｽﾄ06-(06)-1_仮説検定_両側検定_片側検定
#************************************************
import numpy as np
import matplotlib.pyplot as plt
from mpl_toolkits.mplot3d import Axes3D
%matplotlib inline

# ガウス分布関数
def gauss(xlist, mu, sigma):
    return (1 / (np.sqrt(2 * np.pi) * sigma))¥
            * np.exp(-(xlist - mu)**2 / (2 * sigma**2))

# データとグリッド
xn = 100
xlist = np.linspace(-4, 4, xn)
mu = 0
sigma = 1

# 臨界値 (±1.96σ)
fiveper = 1.96
cv1xlist = np.array([ fiveper, fiveper ])
cv1ylist = gauss(cv1xlist, mu, sigma)
cv1ylist[0] = 0
cv2xlist = np.array([ -fiveper, -fiveper ])
cv2ylist = gauss(cv2xlist, mu, sigma)
cv2ylist[0] = 0

# グラフ描画
plt.figure(figsize=(15, 3))

# 分布曲線と臨界値 (両側検定)
plt.subplot(1, 2, 1)
plt.plot( xlist, gauss(xlist, mu, sigma), 'cornflowerblue', linewidth=3,¥
         label='two-sided test' )
plt.plot( cv1xlist, cv1ylist, color='black', linestyle='--', linewidth=2 )
plt.plot( cv2xlist, cv2ylist, color='black', linestyle='--', linewidth=2 )

plt.title('Hypothesis testing  (two-sided)')
plt.legend(loc='lower center')
plt.ylim(-0.1, 0.5)
plt.xlim(-4, 4)
plt.xlabel('$x$')
plt.ylabel('$y$')
plt.grid(True)

# 分布曲線 (片側検定)
plt.subplot(1, 2, 2)
plt.plot( xlist, gauss(xlist, mu, sigma), 'cornflowerblue', linewidth=3,¥
         label='one-sided test' )
plt.plot( cv1xlist, cv1ylist, color='black', linestyle='--', linewidth=2 )

plt.title('Hypothesis testing  (one-sided)')
plt.legend(loc='lower center')
plt.ylim(-0.1, 0.5)
plt.xlim(-4, 4)
plt.xlabel('$x$')
plt.ylabel('$y$')
plt.grid(True)

plt.show()
```

(ﾘｽﾄ06-(06)-2_仮説検定とP値)

```
#************************************************
# ﾘｽﾄ06-(06)-2_仮説検定とP値
#************************************************
import numpy as np
import matplotlib.pyplot as plt
from mpl_toolkits.mplot3d import Axes3D
%matplotlib inline

# ガウス分布関数
def gauss(xlist, mu, sigma):
    return (1 / (np.sqrt(2 * np.pi) * sigma))¥
            * np.exp(-(xlist - mu)**2 / (2 * sigma**2))

# データとグリッド
xn = 100
xlist = np.linspace(-4, 4, xn)
mu = 0
sigma = 1

# P値 (2.0σ)
pValue = 2.0
cv1xlist = np.array([ pValue, pValue ])
cv1ylist = gauss(cv1xlist, mu, sigma)
cv1ylist[0] = 0
cv2xlist = np.array([ -pValue, -pValue ])
cv2ylist = gauss(cv2xlist, mu, sigma)
cv2ylist[0] = 0

# グラフ描画
plt.figure(figsize=(15, 3))

# 分布曲線と臨界値 (両側検定)
plt.subplot(1, 2, 1)
plt.plot( xlist, gauss(xlist, mu, sigma), 'cornflowerblue', linewidth=3,¥
         label='Normal Distribution' )
plt.plot( cv1xlist, cv1ylist, color='black', linestyle='--', linewidth=2 )
plt.plot( cv2xlist, cv2ylist, color='black', linestyle='--', linewidth=2 )

plt.title('Hypothesis testing  (two-sided) Tr=2.0')
plt.legend(loc='lower center')
plt.ylim(-0.1, 0.5)
plt.xlim(-4, 4)
plt.xlabel('$Z$')
plt.ylabel('$p$')
plt.grid(True)

# 分布曲線 (片側検定)
plt.subplot(1, 2, 2)
plt.plot( xlist, gauss(xlist, mu, sigma), 'cornflowerblue', linewidth=3,¥
         label='Normal Distribution' )
plt.plot( cv1xlist, cv1ylist, color='black', linestyle='--', linewidth=2 )

plt.title('Hypothesis testing  (one-sided) Tr=2.0')
plt.legend(loc='lower center')
plt.ylim(-0.1, 0.5)
plt.xlim(-4, 4)
plt.xlabel('$Z$')
plt.ylabel('$p$')
plt.grid(True)

plt.show()
```

(7) その他

この章では、計算のﾉｳﾊｳなどを記します。

(7.1) 観測値追加時の平均値

・「平均値 (ﾍｲｷﾝﾁ、mean value)」は、観測値の総和を観測値の個数で割ったもので、(式2.1-1)のとおりです。

平均値 (n個の観測値)

$$x_{mean\_n} = (\sum_{i=1}^{n} x_i)/n$$

x_{mean_n} : n個の観測値の平均値 …(式2.1-1')

n : 観測値の個数

x_i : i番目の観測値 (i=1～n)

これにもう一個観測値を加えた時の平均値は、計算済の平均値を用いて、以下のように計算できます。

平均値 (n+1個の観測値)

$$x_{mean\_n+1} = x_{mean\_n} + (x_{n+1} - x_{mean\_n})/(n+1)$$

x_{mean_n+1} : n+1個の観測値の平均値 …(式7.1)

x_{mean_n} : n個の観測値の平均値

n+1 : 観測値の個数

x_i : i番目の観測値 (i=1～n+1)

(8) 確認問題

(1) 基本的な統計指標

基本的な統計指標について見てみましょう。
以下の空欄に最もあてはまる語句を選択肢から選び、その記号を回答欄に記入してください。

・(1.1)＿＿＿＿＿＿＿＿＿＿は、観測値を小さい順に並べたとき中央に位置する値です。
(1.2)＿＿＿＿＿＿＿＿＿＿は、観測値の総和を観測値の個数で割ったものです。
データの分布が対称である場合は、中央値は平均値に等しくなります。
分布が対称でなくても、中央値と平均値が等しくなる事もあります。
中央値は平均値と類似した目的で使いますが、全体の傾向を表す代表値として適切である場合が多いです。

・データを小さい順に並び替えたときに、データの数で四等分した時の区切り値を
(1.3)＿＿＿＿＿＿＿＿＿＿と言います。四等分すると三つの区切りの値が得られ、
小さいほうから「25パーセンタイル（第一四分位数）」、「50パーセンタイル（中央値）」、
「75パーセンタイル（第三四分位数）」とよびます。
第一四分位数・第三四分位数の差は、「四分位範囲（ｼﾌﾞﾝｲﾊﾝｲ、interquartile range, IQR)」といい、
分布のばらつきの代表値となっています。

・各データと平均値μとの差を(1.4)＿＿＿＿＿＿と言いますが、これの総和は常に0になるため、
データのばらつき具合の指標として、そのままで扱うのには問題があります。

・偏差の代わりに、偏差の二乗和の平均を取ったものを(1.5)＿＿＿＿＿＿＿＿と呼び、
これをデータのばらつき具合の指標として使い、σ^2 (ｼｸﾞﾏの2乗)または V (ﾌﾞｲ)で記述します。
分散はデータがどの程度平均値の周りにばらついているかの指標になります。

・分散の平方根を(1.6)＿＿＿＿＿＿＿＿＿＿と呼び、σ(ｼｸﾞﾏ)で記述します。
二乗和を取った分散とは異なり、元のデータと同じ次元（単位）になっています。

（選択肢）

(a) 「四分位数（ｼﾌﾞﾝｲｽｳ、quartile)」
(b) 「中央値（ﾁｭｳｵｳﾁ、median)」
(c) 「標準偏差（ﾋｮｳｼﾞｭﾝﾍﾝｻ, standard deviation)」
(d) 「分散（ﾌﾞﾝｻﾝ, variance)」
(e) 「平均値（ﾍｲｷﾝﾁ、mean value)」
(f) 「偏差（ﾍﾝｻ, deviation)」

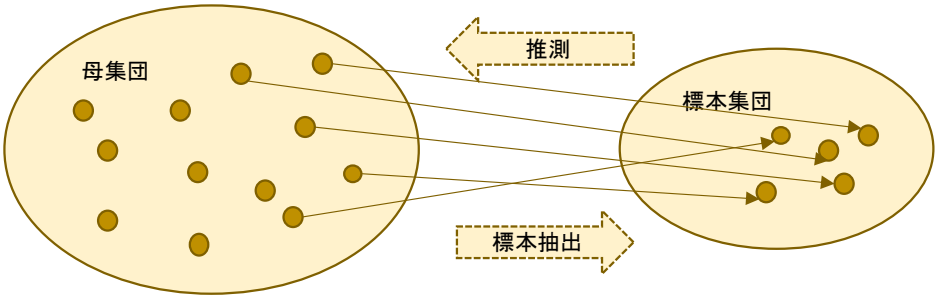
（回答）

(1.1)	
(1.2)	
(1.3)	
(1.4)	
(1.5)	
(1.6)	

(2) 母集団と標本集団

母集団とそれからの標本抽出について見てみましょう。

以下の空欄に最もあてはまる語句を選択肢から選び、その記号を回答欄に記入してください。



・(2.1) ＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿とは、調査対象となる集合全体を言います。

母集団全体を調査の対象とすることが出来ない時、母集団の中から一定の数のデータを抜き出します。

抜き出した集団を(2.2) ＿＿＿＿＿＿＿＿＿＿＿＿ 、

標本集団を選ぶことを(2.3) ＿＿＿＿＿＿＿＿＿＿＿＿、

抜き出したデータの個数を(2.4) ＿＿＿＿＿＿＿＿＿＿＿＿と呼びます。

・データの集合が母集団であるとき、

データの集合の平均を(2.5) ＿＿＿＿＿＿＿＿＿＿＿＿（μ（ﾐｭｰ））、

データの集合の標準偏差を(2.6) ＿＿＿＿＿＿＿＿＿＿＿＿（σ（ｼｸﾞﾏ））、

母標準偏差の二乗を(2.7) ＿＿＿＿＿＿＿＿＿＿＿＿（σ^2）と呼びます。

・データの集合が標本集団であるとき、

データの集合の平均を(2.8) ＿＿＿＿＿＿＿＿＿＿＿＿（x_{mean}）、

データの集合の標準偏差を(2.9) ＿＿＿＿＿＿＿＿＿＿＿＿（ｓ（ｴｽ））、

標本標準偏差の二乗を(2.10) ＿＿＿＿＿＿＿＿＿＿＿＿（s^2（ｴｽのﾆｼﾞｮｳ））と呼びます。

（選択肢）

(a) 「標本集団（ﾋｮｳﾎﾝｼｭｳﾀﾞﾝ、sample）」
(b) 「標本の大きさ（sample size）」
(c) 「標本抽出（ﾋｮｳﾎﾝﾁｭｳｼｭﾂ、sampling）」
(d) 「標本標準偏差（ﾋｮｳﾎﾝﾋｮｳｼﾞｭﾝﾍﾝｻ, sample standard deviation）」
(e) 「標本分散（ﾋｮｳﾎﾝﾌﾞﾝｻﾝ, sample variance）」
(f) 「標本平均（ﾋｮｳﾎﾝﾍｲｷﾝ、sample mean）」
(g) 「母集団（ﾎﾞｼｭｳﾀﾞﾝ、population）」
(h) 「母標準偏差（ﾎﾞﾋｮｳｼﾞｭﾝﾍﾝｻ, population standard deviation）」
(i) 「母分散（ﾎﾞﾌﾞﾝｻﾝ, population variance）」
(j) 「母平均（ﾎﾞﾍｲｷﾝ、population mean）」

（回答）

(2.1)	
(2.2)	
(2.3)	
(2.4)	
(2.5)	
(2.6)	
(2.7)	
(2.8)	
(2.9)	
(2.10)	

(3) 確率変数と確率分布

確率変数と確率分布について見てみましょう。

・「確率変数（ｶｸﾘﾂﾂﾍﾝｽｳ、random variable)」とは、ある確率で値を取る変数のことです。
確率変数を X、その確率を P(X) で表現します。
確率変数がとる値とその出現確率の対応を「確率分布（ｶｸﾘﾂﾂﾌﾞﾝﾌﾟ、probability distribution)」
と言います。確率分布には、確率変数が離散的である離散型確率分布と、
連続的である連続型確率分布があります。

・確率変数が離散的である（飛び飛びの値を取る）場合の確率分布が
「離散型確率分布（ﾘｻﾝｶﾞﾀｶｸﾘﾂﾂﾌﾞﾝﾌﾟ、discrete probability distribution)」です。
離散型確率変数の各値 x_i に対してその発生確率$P(x_i)$ を表すのが、
「確率質量関数（ｶｸﾘﾂﾂｼﾂﾘｮｳｶﾝｽｳ、probability mass function)」であり、離散型確率分布を表します。
離散型確率分布では、各離散値 x_i に対する発生確率の総和は１となります。

$$\sum_i P(x_i) = 1 \quad (0 \leqq P(x_i) \leqq 1)$$

・確率変数が連続的である（連続的な値を取る）場合の確率分布が
「連続型確率分布（ﾚﾝｿﾞｸｶﾞﾀｶｸﾘﾂﾂﾌﾞﾝﾌﾟ, continuous probability distribution)」です。
連続型確率変数の各値 x に対してその発生確率f(x) を表すのが、
「確率密度関数（ｶｸﾘﾂﾂﾐﾂﾄﾞｶﾝｽｳ, probability density function)」であり、連続型確率分布を表します。
連続型確率分布では、確率変数の定義域全体に渡る確率密度関数 f(x)の積分は１となります。

$$\int_{-\infty}^{+\infty} f(x)\ dx = 1 \quad (0 \leqq f(x) \leqq 1)$$

・確率変数が従う確率分布を記した以下の表を、（選択肢）から完成させてください。

確率変数 X	離散/連続	従う確率分布
１回のベルヌーイ試行で成功する回数 X （0, 1 の何れか）	離散型	(3.1)
n 回のベルヌーイ試行において成功する回数 X （0, 1, .., n の何れか）	離散型	(3.2)
ベルヌーイ試行を何回か繰り返すときに、初めて成功するまでの回数 X（1,2,...）	離散型	(3.3)
発生が極めてまれな事象が単位時間あたりに起こる回数 X	離散型	(3.4)
着目している事象が次に起こるまでの期間 X	連続型	(3.5)
平均値の付近に集積するような、釣り鐘型のデータ分布を表した連続的な確率変数 X	連続型	(3.6)
定義域内の任意の値に対し発生確率が等しい離散的な確率変数 X	離散型	(3.7)
定義域内の任意の値に対し発生確率が等しい連続的な確率変数 X	連続型	(3.8)

（選択肢）

(a) 「ベルヌーイ分布（ﾍﾞﾙﾇｰｲﾌﾞﾝﾌﾟ、Bernoulli distribution)」
(b) 「ポアソン分布（ﾎﾟｱｿﾝﾌﾞﾝﾌﾟ、Poisson distribution)」
(c) 「幾何分布（ｷｶﾌﾞﾝﾌﾟ、geometric distribution)」
(d) 「指数分布（ｼｽｳﾌﾞﾝﾌﾟ、exponential distribution)」
(e) 「正規分布（ｾｲｷﾌﾞﾝﾌﾟ、normal distribution)」
(f) 「二項分布（ﾆｺｳﾌﾞﾝﾌﾟ、binominal distribution)」
(g) 「離散一様分布（ﾘｻﾝｲﾁﾖｳﾌﾞﾝﾌﾟ、discrete uniform distribution)」
(h) 「連続一様分布（ﾚﾝｿﾞｸｲﾁﾖｳﾌﾞﾝﾌﾟ、continuous uniform distribution)」

（回答）

(3.1)	
(3.2)	
(3.3)	
(3.4)	
(3.5)	
(3.6)	
(3.7)	
(3.8)	

(4) 確率変数の期待値と分散、標準化

確率変数の期待値と分散、標準化について見てみましょう。

以下の空欄に最もあてはまる語句を選択肢から選び、その記号を回答欄に記入してください。

・確率変数 X の取り得る値にその発生確率 P(X) を掛けた値の総和を (4.1) と言い、
E(X) で表します。
これは、確率変数のすべての値 X に確率 P(X) の重みをつけた加重平均となっています。

・離散型確率分布の場合

$$E(X) = \sum_{i=1}^{N} x_i * P(x_i)$$

$P(x_i)$：確率質量関数

・連続型確率分布の場合

$$E(X) = \int_{-\infty}^{+\infty} x * f(x)\, dx$$

$f(x)$：確率密度関数

・確率変数 X の (4.2) は、
確率変数 X の分布が平均値(期待値) μ からどれだけ散らばっているかを示すもので、
確率変数 X の観測値 xi の偏差の二乗和の平均を取ったもので、偏差の二乗の期待値となっています。

・離散型確率分布の分散

$$V(X) = \sum_{i=1}^{N} (x_i - \mu)^2 * P(x_i)$$

$P(x_i)$：確率質量関数

・連続型確率分布の分散

$$V(X) = \int_{-\infty}^{+\infty} (x - \mu)^2 * f(x)\, dx$$

$f(x)$：確率密度関数
μ ：平均値

・平均μ、標準偏差σの確率変数 X を、
平均０、標準偏差１の確率変数 Z へ、次式により変換することが出来ます。
この Z を(4.3) と言います。
平均値がμ、標準偏差がσの正規分布に従う確率変数 X に対し、Z が従う確率分布は
(4.4) と言います。

$$Z = (X - \mu) / \sigma$$

μ ：平均値
σ ：標準偏差

(選択肢)

(a) 「期待値 (ｷﾀｲﾁ、expected value)」
(b) 「標準化確率変数 (ﾋｮｳｼﾞｭﾝｶｶｸﾘﾂﾍﾝｽｳ, standardized random variable)」
(c) 「標準正規分布 (ﾋｮｳｼﾞｭﾝｾｲｷﾌﾞﾝﾌﾟ, Standard Normal Distribution)」
(d) 「分散 (ﾌﾞﾝｻﾝ, variance)」

(回答)

(4.1)	
(4.2)	
(4.3)	
(4.4)	

(5) 確率論・統計学における基本定理

確率論・統計学における基本定理について見てみましょう。
以下の空欄に最もあてはまる語句を選択肢から選び、その記号を回答欄に記入してください。

・事象Ａの発生確率をｐとします。
事象Ａについてのｎ回の独立試行で、事象Ａが起こった回数をｒとします。
試行回数ｎが非常に大きくなると、相対度数　ｒ／ｎ　は、事象の発生確率ｐに限りなく近づきます。
これを (5.1)＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿と呼びます。

・平均μ、標準偏差σを持つ正規母集団からの、大きさｎの無作為標本において、
標本平均　x_{mean} の分布は、平均μ、標準偏差$\sigma/\sqrt{n}$　の正規分布になります：

$$Z = (x_{mean} - \mu) / (\sigma/\sqrt{n})$$

としたとき、 Z の確率分布 ＝Ｎ（0, 1）

一方、平均μ、標準偏差σを持つ任意の母集団からの、大きさｎの無作為標本において、
標本平均　x_{mean} の分布は、各標本の大きさｎが大きい時、
近似的に平均μ、標準偏差$\sigma/\sqrt{n}$　の正規分布になります：

$$Z = (x_{mean} - \mu) / (\sigma/\sqrt{n})$$

としたとき、 Z の確率分布 ≒ Ｎ（0, 1）
これを (5.2)＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿と呼びます。

（選択肢）

(a)　「ベイズの定理（ﾍﾞｲｽﾞﾉﾃｲﾘ、Bayes' theorem)」
(b)　「大数の法則（ﾀｲｽｳﾉﾎｳｿｸ、Law of Large Numbers、LLN)」
(c)　「中心極限定理（ﾁｭｳｼﾝｷｮｸｹﾞﾝﾃｲﾘ、central limit theorem、CLT)」

（回答）

(5.1)	
(5.2)	

(6) 相関分析

２つの確率変数 X、Y の間の相関について見てみましょう。

・２つの確率変数 X、Y の間に何らかの関係がある場合、
これらのデータ間には「相関関係 (ｿｳｶﾝｶﾝｹｲ, correlation)」があるといいます。
一方が増加すれば他方も増加する時、「正の相関関係 (positive correlation)」があるといい、
一方が増加すれば他方が減少する時、「負の相関関係 (negative correlation)」があると言います。

・２つの確率変数 X、Y の間の関係の強さを表す尺度として、「共分散 (ｷｮｳﾌﾞﾝｻﾝ、covariance)」
があります。これは、積 $(X-\mu_x)*(Y-\mu_y)$ の期待値として定義されます。

確率変数 X、Y の共分散「Cov(X,Y)」

$$Cov(X,Y) = E((X-\mu_x)*(Y-\mu_y))$$

μ_x：確率変数 X の期待値
μ_y：確率変数 Y の期待値

・共分散 Cov(X,Y) は、各確率変数 X と Y の測定単位に依存しています。
測定単位に依存しない２変数間の尺度は、共分散を各変数の標準偏差σ_x、σ_yで割って求めます。
これによって得られる尺度は X と Y の「相関係数 (ｿｳｶﾝｹｲｽｳ, correlation coefficient)」
と呼ばれます。

確率変数 X、Y の相関係数「Corr(X,Y)」

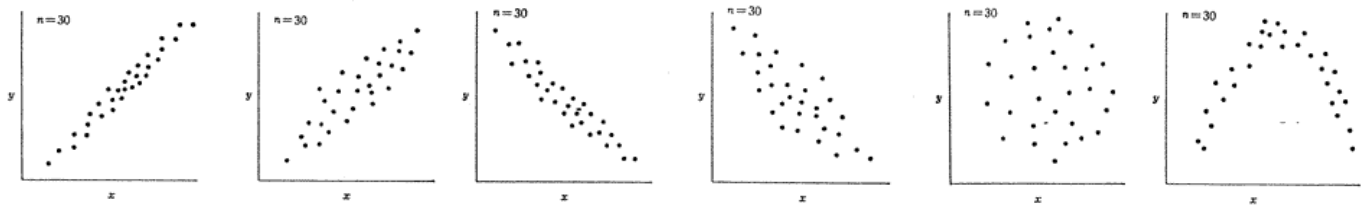
$$Corr(X,Y) = Cov(X,Y) / \sigma_x * \sigma_y$$

σ_x：確率変数 X の標準偏差
σ_y：確率変数 Y の標準偏差

・何かの因果関係を探る場合、統計モデルでは、原因となる変数群を元に、結果となる変数群を説明
しようと試みます。
原因となる変数のことを「説明変数 (ｾﾂﾒｲﾍﾝｽｳ、explanatory variable)」と言います。
結果となる変数のことを「目的変数 (ﾓｸﾃｷﾍﾝｽｳ、response variable)」と言います。

・説明変数を選ぶ際に気を付ける必要があるのは、多重共線性の問題です。
「多重共線性 (ﾀｼﾞｭｳｷｮｳｾﾝｾｲ、multicollinearity)」(略称「ﾏﾙﾁｺ」)とは、
モデル内の一部の説明変数と他の説明変数の相関係数が高いときに起こる状態です。
多重共線性によって、正しく推計できなくなるといった悪影響をもたらします。
この問題の最も一般的な解消法は、「相関関係が高いと考えられる説明変数を外すこと」です。

・以下の散布図は、どのような相関関係を表しているのか、（選択肢）から選んで場合分けしてください：



(6.1)　(6.2)　(6.3)　(6.4)　(6.5)　(6.6)

（選択肢）

(a)　相関の無い場合
(b)　直線的でない関係の場合
(c)　強い正相関のある場合
(d)　強い負相関のある場合
(e)　弱い正相関のある場合
(f)　弱い負相関のある場合

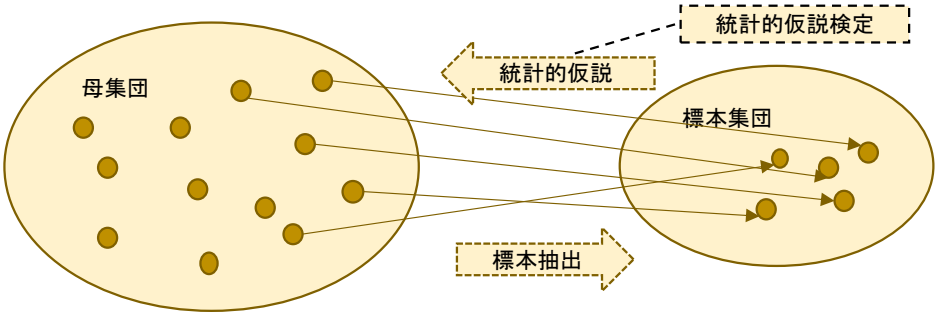
（回答）

(6.1)	
(6.2)	
(6.3)	
(6.4)	
(6.5)	
(6.6)	

(7) 統計的仮説検定

統計的仮説検定について見てみましょう。
以下の空欄に最もあてはまる語句を選択肢から選び、その記号を回答欄に記入してください。

・母集団についての仮説を「統計的仮説（ﾄｳｹｲﾃｷｶｾﾂ、statistical hypothesis)」と呼びます。
統計的仮説には、「帰無仮説」と帰無仮説の逆の内容を持つ「対立仮説」があります。
標本から何らかの主張を立証しようとしている時、主張自体を
(7.1)＿＿＿＿＿＿＿＿＿＿＿＿と呼び、H_1 と記します。
主張を否定する(その主張は誤りであり、元々問題なく、主張は無に帰するという)仮説を
(7.2)＿＿＿＿＿＿＿＿＿＿＿＿と呼び、H_0 と記します。
対立仮説は、帰無仮説が「棄却（ｷｷｬｸ、reject)」された場合に「採択（ｻｲﾀｸ、accept)」されます。
この時、検定は(7.3)＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿であると言います。



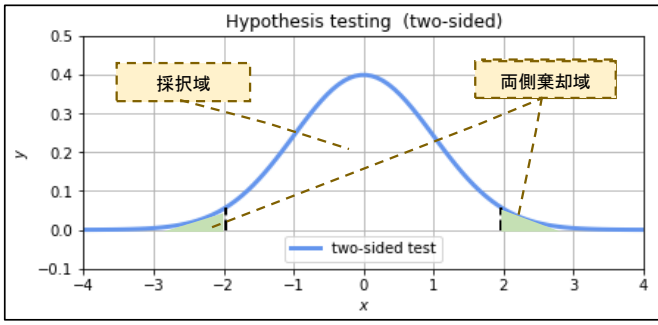
・統計的仮説の妥当性を、標本から検証することを(7.4)＿＿＿＿＿＿＿＿＿＿＿＿、
あるいは単に「検定」と呼びます。

・統計的仮説検定に用いられる確率変数を「検定統計量（ｹﾝﾃｲﾄｳｹｲﾘｮｳ、test statistic)」、あるいは単に「統計量」と呼びます。
検定統計量は、標本を元に計算されるので、標本統計量と同様に分布を持ちます。
帰無仮説の下で導かれた検定統計量の分布を元にして、検定を行います。

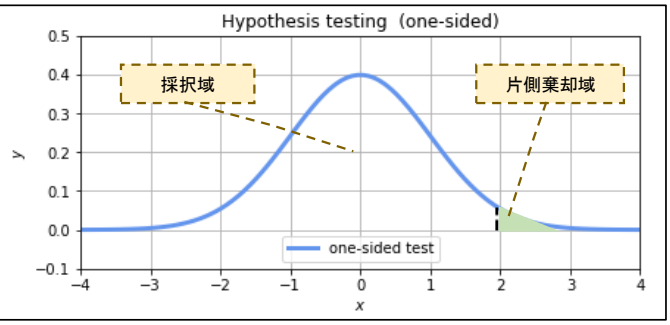
・検定で、帰無仮説 H_0 が棄却されると判断する検定統計量の値の範囲は、
帰無仮説の下で導かれた検定統計量の分布の裾（両裾または片裾）に設定され、
(7.5)＿＿＿＿＿＿＿＿＿＿＿＿と呼びます。
棄却域以外の領域は帰無仮説 H_0 が棄却されない領域であり、
(7.6)＿＿＿＿＿＿＿＿＿＿＿＿または「受容域」と呼びます。
棄却域と採択域の境目を「臨界値（ﾘﾝｶｲﾁ, critical value)」または「境界値」と呼びます。

・検定統計量の分布で、棄却域の面積(＝確率)αを「有意水準（ﾕｳｲｽｲｼﾞｭﾝ、significance level)」と呼び、通常、0.05(5%)、0.01(1%) といった小さな確率に定められます。

(例7.1) 両側検定の棄却域の例



(例7.2) 片側検定の棄却域の例



・(例7.1) 両側検定で $X \leqq -1.96\sigma$、$1.96\sigma \leqq X$ を棄却域とする（$\alpha = 0.05$(5%)）。
・(例7.2) 片側検定で　　$1.96\sigma \leqq X$ を棄却域とする（$\alpha = 0.025$(2.5%)）。

・統計的仮説の設定に従って、検定は両側検定と片側検定に分類されます。

・統計量がある閾値よりも大きい（あるいは小さい）かを判定するような対立仮説を
「片側対立仮説（ｶﾀｶﾞﾜﾀｲﾘﾂｶｾﾂ、one-sided alternative hypothesis)」と言います。
片側対立仮説では、棄却域が統計量の分布の片裾にあるため、
この棄却域を「片側棄却域（ｶﾀｶﾞﾜｷｷｬｸｲｷ、one-sided rejection region)」と言い、
この検定を「片側検定（ｶﾀｶﾞﾜｹﾝﾃｲ、one-sided test)」と言います。

・統計量がある値と一致するかを判定するような対立仮説を
「両側対立仮説（ﾘｮｳｶﾞﾜﾀｲﾘﾂｶｾﾂ、two-sided alternative hypothesis)」と言います。
両側対立仮説では、棄却域が統計量の分布の両裾にあるため、
この棄却域を「両側棄却域（ﾘｮｳｶﾞﾜｷｷｬｸｲｷ、twp-sided rejection region)」と言い、
この検定を「両側検定（ﾘｮｳｶﾞﾜｹﾝﾃｲ、two-sided test)」と言います。

・観察されたデータを用いて計算された検定統計量が、棄却域に入る時は、帰無仮説 H_0 を棄却します。
検定統計量が、棄却域に入らない時は、帰無仮説 H_0 を棄却できません。

・検定の手順をまとめると、以下の様になります：

検定の手順

（手順1）帰無仮説 H_0 と対立仮説 H_1 を設定する。
（手順2）検定統計量を定める。
（手順3）有意水準αを定める。
（手順4）帰無仮説 H_0 の下での検定統計量の分布に基づき、
有意水準αに対応する棄却域を定める。
（手順5）観察されたデータを用いて検定統計量を計算する。
（手順6）計算された検定統計量が、棄却域に入る時は、帰無仮説 H_0 を棄却できると判定する。
検定統計量が、棄却域に入らない時は、帰無仮説 H_0 を棄却できないと判定する。

・(7.7)＿＿＿＿＿＿＿＿＿＿＿＿は、観測された検定統計量 T の値が t_r の時、
帰無仮説 H_0 での検定統計量の分布の下で、
$T=t_r$ を臨界値とした場合の棄却域の面積（検定統計量 T が t_r より棄却域側にある割合）です。

・両側仮説検定の場合　　Ｐ値 ＝ $P(|T| > t_r)$
・右側の片側仮説検定の場合　　Ｐ値 ＝ $P(T > t_r)$
・左側の片側仮説検定の場合　　Ｐ値 ＝ $P(T < t_r)$

・Ｐ値は、「帰無仮説 H_0 での検定統計量の分布の下で、
観測された検定統計量以上に、(帰無仮説に反する)偏った検定統計量が得られる確率」を示しています。
Ｐ値が有意水準αを下回ったときに、はじめて「統計的有意差があった」と言うことができます。

（選択肢）

(a)　「Ｐ値（ﾋﾟｰﾁ、P-value)」
(b)　「棄却域（ｷｷｬｸｲｷ,rejection region)」
(c)　「帰無仮説（ｷﾑｶｾﾂ、Null hypothesis)」
(d)　「採択域（ｻｲﾀｸｲｷ, acceptance region)」
(e)　「対立仮説（ﾀｲﾘﾂｶｾﾂ、alternative hypothesis)」
(f)　「統計的に有意（ﾄｳｹｲﾃｷﾆﾕｳｲ、statistically significant)」
(g)　「統計的仮説検定（ﾄｳｹｲﾃｷｶｾﾂｹﾝﾃｲ、statistical hypothesis testing)」

（回答）

(7.1)	
(7.2)	
(7.3)	
(7.4)	
(7.5)	
(7.6)	
(7.7)	

以上。

(9) 確認問題回答用紙

提出者　：

提出日　：　　年　　月　　日

回答

（全 43 問）

No.	回答
(1.1)	
(1.2)	
(1.3)	
(1.4)	
(1.5)	
(1.6)	
(2.1)	
(2.2)	
(2.3)	
(2.4)	
(2.5)	
(2.6)	
(2.7)	
(2.8)	
(2.9)	
(2.10)	
(3.1)	
(3.2)	
(3.3)	
(3.4)	
(3.5)	
(3.6)	
(3.7)	
(3.8)	

No.	回答
(4.1)	
(4.2)	
(4.3)	
(4.4)	
(5.1)	
(5.2)	
(6.1)	
(6.2)	
(6.3)	
(6.4)	
(6.5)	
(6.6)	
(7.1)	
(7.2)	
(7.3)	
(7.4)	
(7.5)	
(7.6)	
(7.7)	

（※黄色の枠　のみ記入をお願いします）

※ ご意見・ご要望など有りましたら、下欄に記してください。