第1課題：統計的分析

岩崎淳
iwasaki.atsushi.4x@kyoto-u.ac.jp

1　はじめに

1.1　課題について

2章以降の各章ごとに課題がついている．それらの全てに取り組むこと．なお，プログラミングはすべてC言語を用いて行うこと．
描画（グラフの作成）の方法は各自にまかせる．

1.2　レポート

レポートに基づいて成績を評価する．課題への解答をまとめたものをレポートととして提出すること．レポートの作成に当たっては，この資料の他，文献等を調査してもよい．ただし，引用は引用元と引用箇所を明示すること．また，課題で作成したソースコードもzip形式でひとまとめにし，あわせて提出すること．

提出期限：5月27日（水）

提出方法：PandA上で提出

過去には，プログラムが間違っているのに正しい答えのレポートが提出されるなどの不正行為をうかがわせる例があった．そういったことは絶対に行わないように．不正行為には厳格に対応する．

1.3　データの取得方法

課題で必要なデータファイルの取り扱いについて説明しておく．例えば，ファイル名が「dat1_double_1000」であれば真ん中の「double」がdouble型を，最後の「1000」がデータ数を表している．すなわち，「dat1_double_1000」にはdouble型のデータが1000個分入っている．同様に，「dat32_int_555」であればint型のデータが555個入っている．

データファイルはバイナリ形式になっている．そのためテキストエディタ（メモ帳など）では（正常に）開くことができない．データの読み出しは以下のコードを参考にされたい:

dat1_double_1000 から配列 x[0],x[1],···,x[999] にデータを取得する例

```
double x[1000];
 FILE *fp;

 fp=fopen("dat1_double_1000", "rb");

 fread(&x[0],8,1000,fp);

 fclose(fp);
```

dat32_int_555 から配列 y[0],y[1],···,y[554] にデータを取得する例

```
int y[555];
 FILE *fp;

 fp=fopen("dat32_int_555", "rb");

 fread(&y[0],4,555,fp);

 fclose(fp);
```

また,「dat3_double_100_dim2」のように後ろに「dim2」と書かれているものは 2 次元のデータであることを表す．2 次元のデータが 100 個なので，合計 200 個分のデータが含まれていることに注意.

dat3_double_100_dim2 から配列 z[0][0],z[0][1],···z[99][0],z[99][1] にデータを取得する例

```
double z[100][2];
 FILE *fp;

 fp=fopen("dat3_double_100_dim2", "rb");

 fread(&z[0][0],8,200,fp);

 fclose(fp);
```

この例では，(z[0][0],z[0][1])，(z[1][0],z[1][1])，(z[2][0],z[2][1]) のように，z[i][0] と z[i][1] がペアとなり 2 次元のデータを構成している．

2　基礎編

2.1　度数・ヒストグラム

データ（標本・サンプル）が取りうる値の範囲を複数の**階級**に分類し，階級ごとにその範囲に入ったデータがいくつあるかを数えた値をその階級の**度数**という．また，度数をデータの総数で割った値を**相対度数**という．

考えている階級とそれより下の（値が小さい）階級の度数を全て足した値を**累積度数**という．**累積相対度数**も同様に定義される．

階級と度数・相対度数・累積度数・累積相対度数を表にまとめたものを度数分布表，グラフに描いたものをヒストグラムと呼ぶ．

2.2　平均・中央値・分散・標準偏差

データ $x_1, x_2, \cdots, x_n$ が与えられたとしよう．各 x_i は実数値とする．データから計算される最も基礎的・一般的な統計量には,「平均」「中央値」「分散」「標準偏差」が挙げられる．

平均

「平均」の定義は状況によって多様な定義が使われるが，平均という言葉から直感的に想像がつく以下で定義される算術平均であろう：

$$(\text{平均}) := \frac{(\text{データの総和})}{(\text{データ数})} \tag{1}$$

$$= \frac{1}{n}\sum_{i=1}^{n} x_i \tag{2}$$

中央値

平均は広く使われるが，データの中に少数の大きな値（または，マイナス方向に大きな値）が含まれていた場合にその影響を受けやすい．例えば，日本人の平均年収を考えると，ごく一部の高所得者の影響を受けて「標準的な」日本人の年収よりも平均年収は高くなる．場合によってはそのような影響を受けにくい統計量を用いたいときもある．

そのためによく用いられるものとして**中央値**がある．データを大きい順に並び替え，中央に来る値として定義される．データ数が偶数の場合には中央を挟む二つの値の平均値として定義される:

$$(\text{中央値}) := \begin{cases} z_{\frac{n+1}{2}} & (n \text{ が奇数}) \\ \dfrac{z_{\frac{n}{2}} + z_{\frac{n+2}{2}}}{2} & (n \text{ が偶数}) \end{cases} \tag{3}$$

ただし，並び替えた後のデータを $z_1, z_2, \cdots, z_n$ としている．

分散

分散はデータのばらつき度合いを示す統計量で，値が大きい方がより広くばらついていることを表す．分散は非負の値しかとらない．値がちょうど 0 になるのは，データに全くばらつきがない，すなわち，$x_1 = x_2 = \cdots = x_n$ の場合である．分散は以下のように定義される．

$$
\begin{aligned}
(\text{分散}) :=& \frac{(\text{各データと平均との差の二乗の総和})}{(\text{データ数})} && (4)\\
=& \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \frac{1}{n}\sum_{j=1}^{n} x_j\right)^2 && (5)\\
=& \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \left(\frac{1}{n}\sum_{j=1}^{n} x_j\right)^2 && (6)\\
=& (\text{各データを二乗した値の平均}) - (\text{平均})^2 && (7)
\end{aligned}
$$

標準偏差

標準偏差も分散と同じくデータのばらつき度合いを示す統計量である．ただし，分散とは異なり，元のデータと同じ「次元」の値となる．どういうことかというと，例えば元のデータが長さを表す（“メートル” などの単位で表される）としよう．このとき，分散は面積の次元（“平方メートル”) であるが，標準偏差の単位は元のデータと同じ長さの次元（“メートル”）になる．つまり，標準偏差は元のデータと「比較」できる．

標準偏差は分散の平方根で表される：

$$
\begin{aligned}
(\text{標準偏差}) :=& \sqrt{(\text{分散})} && (8)\\
=& \sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \left(\frac{1}{n}\sum_{j=1}^{n} x_j\right)^2} && (9)
\end{aligned}
$$

2.3　2 次元のデータ

2 次元のデータを $(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$ を考えよう．第一成分（x_i の方）についての統計量・第二成分（y_i の方）についての統計量の他に，今度は第一成分と第二成分の関係性についての統計量も考えることが出来る（考

えなければならない）．最もベーシックなものとして,「共分散」と「相関係数」が挙げられる：

$$\begin{aligned}&(\text{共分散})\\&:= (\text{“第一成分の平均との差} \times \text{第二成分の平均との差” の平均})\end{aligned} \tag{10}$$

$$= \frac{1}{n}\sum_{k=1}^{n}\left(x_k - \frac{1}{n}\sum_{i=1}^{n}x_i\right)\left(y_k - \frac{1}{n}\sum_{j=1}^{n}y_j\right) \tag{11}$$

$$= \frac{1}{n}\sum_{k=1}^{n}x_k y_k - \left(\frac{1}{n}\sum_{i=1}^{n}x_i\right)\left(\frac{1}{n}\sum_{j=1}^{n}y_j\right) \tag{12}$$

$$\begin{aligned}&= (\text{“第一成分} \times \text{第二成分” の平均})\\&\qquad - (\text{第一成分の平均}) \times (\text{第二成分の平均})\end{aligned} \tag{13}$$

$$(\text{相関係数}) := \frac{(\text{共分散})}{(\text{第一成分の標準偏差}) \times (\text{第二成分の標準偏差})} \tag{14}$$

$$= \frac{\frac{1}{n}\sum_{i=1}^{n}x_i y_i - \left(\frac{1}{n}\sum_{j=1}^{n}x_j\right)\left(\frac{1}{n}\sum_{k=1}^{n}y_k\right)}{\sqrt{\frac{1}{n}\sum_{s=1}^{n}x_s^2 - \left(\frac{1}{n}\sum_{t=1}^{n}x_t\right)^2}\sqrt{\frac{1}{n}\sum_{u=1}^{n}y_u^2 - \left(\frac{1}{n}\sum_{v=1}^{n}y_v\right)^2}} \tag{15}$$

相関係数は [-1,1] の値しかとらない．

2.3.1 最小二乗法

2 次元のデータを $(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$ を考え，データを近似する xy 平面上の直線を考えよう．直線は一般に

$$y = ax + b \tag{16}$$

の形で書くことができる．問題は a と b の定め方だ．最小二乗法は最もよく使われる方法で，以下の量が最小になるように a と b を定める：

$$\sum_{i=1}^{n}\{y_i - (ax_i + b)\}^2 \tag{17}$$

この最小化問題は難しい最適化法を使わなくても解くことができて，

$$a = \frac{(\text{第一成分と第二成分の共分散})}{(\text{第一成分の分散})} \tag{18}$$

$$b = (\text{第二成分の平均}) - a \times (\text{第一成分の平均}) \tag{19}$$

が解となる．

2.4 自己相関

とある現象から時間とともに取得したデータ $x_1, x_2, \cdots, x_n$ を考えよう．(x_1 は時刻 1 のときの値，x_2 は時刻 2 のときの値，といったぐあいである．) そのようなデータを時系列データという．例えば,「1 月 1 日から 10 月 15 日までのとある銘柄の株価のデータ」のようなものをイメージすればよい．

時系列データの場合，データの “順番” にも意味がある.「変化の過程」が記録されているからだ．その「変化の過程」を分析するにあたって，最も単純な視点は「時間的に前の値は後ろにどう影響しているか」であろう．それを捉える手法の一つとして，相関係数を用いたものがある．

“1 次元のデータなのに相関係数?” という疑問は自然である（むしろ疑問を感じてほしい）．どうするかというと，“順番” を利用してデータを 2 次元に水増しするのである．具体的には以下のようなデータを構成する：

$$(x_1, x_{h+1}),\ (x_2, x_{h+2}),\ (x_3, x_{h+3}),\ \cdots,\ (x_{n-h-1}, x_{n-1}),\ (x_{n-h}, x_n) \quad (20)$$

時刻 i での値 x_i と時刻 $i+h$ での値 x_{i+h} をペアにして 2 次元のデータを構成している．ここで，h は定数である．この方法で時系列データに対して計算された相関係数のことを自己相関と呼ぶ.

ただし，データ数 n が十分に大きく h が n に対して十分に小さい場合には，相関係数を計算するにあたって用いられる第一成分・第二成分それぞれの平均と標準偏差は，1 次元のデータ $x_1, x_2, \cdots, x_n$ の平均と分散で代用できる．

2.5 課題

1. データファイル dat1_double_1000 からデータを取得し，平均・中央値・分散・標準偏差を求めよ．

2. データファイル dat2_int_1000 からデータを取得し，平均・中央値・分散・標準偏差を求めよ．

   ヒント：オーバフローに注意せよ．

3. データファイル dat3_double_100_dim2 からデータを取得し，共分散・相関係数を求めよ．

4. データファイル dat4_double_100_dim2 からデータを取得し，共分散・相関係数を求めよ．

5. dat3_double_100_dim2 と dat4_double_100_dim2 をそれぞれプロットし，最小二乗法で求めた近似直線を描け．また，相関係数と図の関係を考察せよ．

6. データファイル dat5_double_1000 からデータを取得し，時系列データとみなしたうえで，時刻を一つずらしたときの（すなわち，2.4 節における $h = 1$ として）自己相関を求めよ．

3 推定

3.1 分布とデータ生成

推定の話をする前に，必要な知識を確認しておこう．

3.1.1 連続型の一様分布

連続確率変数 X が実数値上の区間 $[a, b]$ の値を一様な確率でとるとき，X が従う分布を $[a, b]$ 上の一様分布と呼ぶ．確率密度関数は

$$p(x) = \begin{cases} \frac{1}{b-a} & (a \leq x \leq b) \\ 0 & (x < a, x > b) \end{cases} \tag{21}$$

で表される．

3.1.2 正規分布

平均 μ，分散 σ に従う正規分布の確率密度関数は

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{22}$$

と書き表される．平均が 0，分散が 1 の正規分布のことをとくに，標準正規分布と呼ぶ．

3.1.3 Box-Muller 法

確率変数 X と Y は互いに独立に $[0, 1]$ 上の一様分布に従うとする．このとき，確率変数 Z_1 と Z_2 を

$$Z_1 := \sqrt{-2\log X}\cos 2\pi Y, \tag{23}$$

$$Z_2 := \sqrt{-2\log X}\sin 2\pi Y \tag{24}$$

と定義すると，Z_1 と Z_2 は互いに独立に標準正規分布に従う．

3.1.4 χ^2 分布

非負の値をとる確率変数 X が従う分布の確率密度関数が

$$p(x) = \frac{x^{\frac{k}{2}-1}e^{-\frac{x}{2}}}{2^{\frac{k}{2}}\Gamma\left(\frac{k}{2}\right)} \tag{25}$$

で表されるとき，X が従う分布を**自由度 k の χ^2 分布**という．ここで，Γ はガンマ関数である．

確率変数 $X_1, X_2, \cdots, X_k$ が独立に標準正規分布に従うとき，

$$Y := {X_1}^2 + {X_2}^2 + {X_3}^2 + \cdots + {X_k}^2 \tag{26}$$

で定義される Y は自由度 k の χ^2 分布に従う．

3.1.5 t 分布

確率密度関数が

$$p(x) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}} \tag{27}$$

で表される分布を**自由度 ν の t 分布**という．

確率変数 X が標準正規分布に従い，Y が自由度 k の χ^2 分布に従うとする．また，X と Y は独立であるとする．このとき，

$$Z := \frac{X}{\sqrt{\frac{Y}{k}}} \tag{28}$$

で定義される Z は自由度 k の t 分布に従う．

3.2 推定

確率的な現象を観測したり，大きな集団から一部分を抽出したりしてサンプルを取得した状況を考えよう．大元の現象・集団の分布については未知とする．（分布について完全に未知の場合と，分布の形状はわかっておりパラメータのみが未知の場合の 2 通りある．）

しようとしていることは，得られたサンプルから元の分布を推定することである．

3.2.1 点推定・不偏推定量

多くの場合，元の分布についてまず知りたい情報は平均と分散であろう．その元の分布の平均と分散のことをそれぞれ母平均，母分散と呼ぶ．

元の分布についてのとあるパラメータ A の値は a だとしよう．サンプルから推定したパラメータ A の値 $\hat{a}$ はランダムに取得されたサンプルに依存する．そのため，“$\hat{a}$ の平均” を考えることができる．その “$\hat{a}$ の平均” が a に一致するとき，すなわち，

$$\mathbb{E}\left[\hat{a}\right] = a \tag{29}$$

が成り立つとき，$\hat{a}$ のことを不偏推定量と呼ぶ．

母平均については，通常のサンプルの平均が不偏推定量となる：

$$\mathbb{E}\left[\frac{X_1+X_2+\cdots+X_n}{n}\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[X_i\right] \tag{30}$$

$$= \mu \text{ (母平均)}. \tag{31}$$

一方，母分散については，通常のサンプルの分散は不偏推定量とならず，標本分散の $\frac{n}{n-1}$ 倍が不偏推定量（不偏標本分散）となる：

$$\text{(母分散) } \sigma^2 = \mathbb{E}\left[\frac{n}{n-1}\frac{\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2}{n}\right], \tag{32}$$

ここで，$\bar{X}$ は標本平均 $\bar{X} = \frac{X_1+X_2+\cdots+X_n}{n}$ である．よって，

$$\text{(不偏標本分散) } s^2 = \frac{\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2}{n-1} \tag{33}$$

で与えられる．

3.2.2 区間推定

前述のとおり，標本から推定した元の分布のパラメータ値は，それ自体もまた確率変数とみなせる．そのため，具体的なサンプルから実際に計算したパラメータ値は必ずしも正確な値であるとは限らない．そこで，「真のパラメータ値は 95%の確率で $[a,b]$ の区間内にある」のように幅を持たせた議論をしたい．ちなみに，このような確率のことを**信頼係数**、区間のことを**信頼区間**と呼ぶ．

母平均と母分散がそれぞれ μ と σ^2 の分布からランダムにサンプリングしたデータ $X_1, X_2, \cdots, X_n$ を独立な確率変数列とみなそう．

このとき，標本平均

$$\bar{X} = \frac{X_1+X_2+\cdots+X_n}{n} \tag{34}$$

を考えると，中心極限定理から $n\to\infty$ で $\bar{X}$ は平均 μ・分散 $\frac{\sigma^2}{n}$ の正規分布に従う．実際にはサンプル数 n は有限であるが，十分大きいとして $\bar{X}$ は正規分布に従うと考えれば（比較的）簡単に信頼区間が求まる．具体的には，信頼係数 $1-\alpha$ としたときの母平均の信頼区間は $[\bar{X}-\frac{\sigma}{\sqrt{n}}Z_{\frac{\alpha}{2}}, \bar{X}+\frac{\sigma}{\sqrt{n}}Z_{\frac{\alpha}{2}}]$ で与えられる．こで，$Z_{\frac{\alpha}{2}}$ は標準正規分布の上 $100\frac{\alpha}{2}$%点である．すなわち，$p(z)$ を標準正規分布の確率密度関数とすると，

$$\int_{Z_{\frac{\alpha}{2}}}^{\infty} p(z)dz = \frac{\alpha}{2} \tag{35}$$

を満たす．

今の議論では母分散 σ^2 が既知であることが前提になっている．母分散が未知の場合には，不偏分散

$$s^2 = \frac{\sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2}{n-1} \tag{36}$$

で代用すればよい．ただし，s^2 自体も確率変数となるので変更が必要である．結論をいうと，信頼係数 $1-\alpha$ としたときの母平均の信頼区間は $[\bar{X} - \frac{s}{\sqrt{n}} t_{\frac{\alpha}{2}}(n-1), \bar{X} + \frac{s}{\sqrt{n}} t_{\frac{\alpha}{2}}(n-1)]$ で与えられる．ここで，$t_{\frac{\alpha}{2}}(n-1)$ は自由度 $n-1$ の t 分布の上 $100\frac{\alpha}{2}\%$点である．すなわち，$p(t)$ を自由度 $n-1$ の t 分布の確率密度関数とすると，

$$\int_{t_{\frac{\alpha}{2}}(n-1)}^{\infty} p(t)dt = \frac{\alpha}{2} \tag{37}$$

を満たす．

3.3 課題

1. Box-Muller 法を実装して乱数を生成せよ．その後，生成した乱数のヒストグラムを作成し，標準正規分布に従っていることを確認せよ．

   なお，一様分布に従う乱数が必要ならば “メルセンヌツイスタ” を使用せよ．(以降のすべての問題に共通.）ライブラリ MT.h を用意してある．使い方は以下のサイトを参照されたい：

   `https://omitakahiro.github.io/random/random_variables_generation.html`

   （C 言語の rand 関数の作る乱数には単純な規則性があるので、rand 関数を使用する癖をつけないためにもここではあえてメルセンヌツイスタを使用する.）

2. 平均 3.7，分散 2.5 の正規分布に従う乱数を 100 万サンプル生成し，10 サンプルずつの 10 万セットに分け，

   (a) 各セットごとに標本平均を求め，標本平均が母平均の不偏推定量であることを確認せよ．

   (b) 各セットごとに標本分散を求め，標本分散が母分散の不偏推定量でないことを確認せよ．

   (c) 各セットごとに不偏標本分散を求め，不偏標本分散が母分散の不偏推定量であることを確認せよ．

   ヒント: 確率変数 X の平均が 0 なら $X+a$ の平均は a，X の分散が 1 なら aX の分散は a^2 である．X が正規分布に従っているなら $X+a$ と aX も正規分布に従う（$a \neq 0$）．

3. 区間 $[2, 4]$ 上の一様分布について，前問と同様の確認をせよ．

4. データファイル dat6_double_101 からデータを取得し，信頼係数 95%での母平均の信頼区間を求めよ．

5. データファイル dat7_double_200 からデータを取得し，信頼係数 95%での母平均の信頼区間を求めよ．ただし，母分散は 3.2 である．

4　仮説検定

例えば，コインを投げて表か裏かどちらが出るかを観測することを考えよう．表が出る確率は $\frac{1}{2}$，裏が出る確率も $\frac{1}{2}$ だ．そうすると，100 回投げれば表と裏が 50 回ずつ出てほしい．しかしながら，実際にはぴったり 50 回ずつとなるとは限らず，ズレが生じうる．表 51 回・裏 49 回とか表 60 回・裏 40 回とか．では，極端に偏りがでて，表 100 回・裏 0 回であったならどうか.「たまたまそうなった」と考えるより，そもそも「表裏が出る確率は $\frac{1}{2}$ ずつ，という 前提が間違っていた」と考えるほうが合理的 であろう．

仮説検定とは，まさしく今のような例だ．まず与えられる前提のことを**帰無仮説**という．帰無仮説の下で，実際に起こった（観測した）現象がどれくらい稀なものなのかを求め，あらかじめ定めた基準（**有意水準**）に従って帰無仮説が間違っていたとするかどうかの判定を下す．

なお，仮説検定では誤った結果を導く可能性が否定できない．本当は帰無仮説が正しいのに帰無仮説を間違っていると判定する場合（第一種の誤り）、帰無仮説が間違っているのに正しいと判定する（第二種の誤り）が生じる確率はどちらも 0 にすることはできない．検定を用いるときには常にその点を念頭に置いておかなければならない．

仮説検定は検定対象に応じて無数に作られうる．ここでは，代表的な検定の中からさらに選抜したものだけを取り扱うことにする．

4.1　母平均の検定

以下の帰無仮説に対する検定を考えよう：

帰無仮説

母平均（母集団の平均）は μ である

なお，両側検定を考えることにする．

検定の方法は，母分散 σ^2 が既知の場合と未知の場合で異なる．

4.1.1　母分散 σ^2 が既知の場合

検定統計量 T を以下で計算する：

$$T = \frac{\bar{X} - \mu}{\sqrt{\frac{\sigma^2}{n}}} \tag{38}$$

ここで，n は標本数，$\bar{X}$ は標本平均を表す．標本数 $n \to \infty$ としたとき，帰無仮説の下では，T は標準正規分布に従う．それを踏まえて，有意水準を α と

すると，

$$|T| \leq Z_{\frac{\alpha}{2}} \quad \text{のとき，帰無仮説を採択}$$
$$|T| > Z_{\frac{\alpha}{2}} \quad \text{のとき，帰無仮説を棄却}$$

と判定する．ここで，$Z_{\frac{\alpha}{2}}$ は標準正規分布の上 $100\frac{\alpha}{2}\%$点である．すなわち，$p(z)$ を標準正規分布の確率密度関数とすると，

$$\int_{Z_{\frac{\alpha}{2}}}^{\infty} p(z)dz = \frac{\alpha}{2} \tag{39}$$

を満たす．

4.1.2 母分散 σ^2 が未知の場合

母分散が未知なので，代わりに不偏標本分散 s^2 を用いる．検定統計量 T' を以下で計算する：

$$T' = \frac{\bar{X} - \mu}{\sqrt{\frac{s^2}{n}}} \tag{40}$$

標本数 n が十分に大きいとき，帰無仮説の下では，T' は自由度 $n-1$ の t 分布に従う．それを踏まえて，有意水準を α とすると，

$$|T| \leq t_{\frac{\alpha}{2}}(n-1) \quad \text{のとき，帰無仮説を採択}$$
$$|T| > t_{\frac{\alpha}{2}}(n-1) \quad \text{のとき，帰無仮説を棄却}$$

と判定する．ここで，$t_{\frac{\alpha}{2}}(n-1)$ は自由度 $n-1$ の t 分布の上 $100\frac{\alpha}{2}\%$点である．すなわち，$p(t)$ を自由度 $n-1$ の t 分布の確率密度関数とすると，

$$\int_{t_{\frac{\alpha}{2}}(n-1)}^{\infty} p(t)dt = \frac{\alpha}{2} \tag{41}$$

を満たす．

4.2 χ^2 検定

ありうる背反な現象が K 通りあり，現象 k が起きる確率は p_k であるとしよう．(つまり，$\sum_{k=1}^{K} p_k = 1.$) なお，$k = 1, 2, \cdots, K$ に対して，$p_k \neq 0$ を仮定する．例えばさいころを一回振ると，「1 の目が出る」,「2 の目が出る」, $\cdots$,「6 の目が出る」の 6 通りの現象がありえて，$p_1 = p_2 = \cdots = p_6 = \frac{1}{6}$ となる．

この現象に対して試行回数 N 回の観測を行ったとする．現象 k が起きた回数を f_k とおくと，必ずしも $f_k = Np_k$ が厳密に成り立つとは限らない．そこ

で生じる “ズレ” が帰無仮説の下で説明できるかどうかを見る仮説検定の一つが χ^2 検定だ．すなわち，実際に観測している分布が与えられた理論分布と（誤差の範囲内で）一致しているかを見る**適合度検定**の一種である．

帰無仮説を確認しておこう：

帰無仮説

$k = 1, 2, \cdots, K$ に対して，現象 k が起きる確率は p_k である．

ただし，現象は互いに背反，かつ，$\sum_{k=1}^{K} p_k = 1$，$p_k \neq 0$ $(k = 1, 2, \cdots, K)$ であることに注意．

検定統計量 χ_{obs} は以下で与えられる：

$$\chi_{obs} = \sum_{k=1}^{K} \frac{(f_k - Np_k)^2}{Np_k} \tag{42}$$

この χ_{obs} は，帰無仮説の下では，$N \to \infty$ で自由度 $K-1$ の χ^2 分布に従う．その事実に基づき，有意水準を α とすると，

$$\chi_{obs} \leq \chi^2_\alpha \quad \text{のとき，帰無仮説を採択}$$
$$\chi_{obs} > \chi^2_\alpha \quad \text{のとき，帰無仮説を棄却}$$

と判定する．ここで，χ^2_α は自由度 $K-1$ の χ^2 分布の上 $100\alpha\%$点である．すなわち，$p(z)$ を自由度 $K-1$ の χ^2 分布の確率密度関数とすると，

$$\int_{\chi^2_\alpha}^{\infty} p(z)dz = \alpha \tag{43}$$

を満たす．

4.3 Kolmogorov−Smirnov 検定

Kolmogorov—Smirnov 検定（KS 検定）も χ^2 検定と同じく適合度検定の一種だ．KS 検定ではより直接的に，得られた経験分布と理論分布を比較する．

帰無仮説は χ^2 検定のものと本質的には同じだが，言い回しを変えておこう：

帰無仮説

母集団の累積分布関数は $F(x)$ である

標本を N 個とり，その経験累積分布関数（累積相対度数を表す関数のこと）を $F_N(x)$ とする．もう少し具体的に言うと，データを x_1，x_2，$\cdots$，x_N とすると，

$$F_N(x) = \frac{\#\{i \mid x_i < x,\ i = 1, 2, \cdots, N\}}{N} \tag{44}$$

として与えられる．

この $F_N(x)$ と $F(x)$ を用いて，検定統計量 D を以下の式で求める：

$$D = \sqrt{N} \sup_x |F(x) - F_N(x)| . \tag{45}$$

この D は明らかに非負の値しかとらない．非負の範囲で D が従う分布は，$N \to \infty$ としたとき，

$$\mathrm{Prob}\{D < x\} = 1 - 2\sum_{i=1}^{\infty}(-1)^{i-1}e^{-2i^2x^2} \tag{46}$$

となる．

これに基づき，有意水準を α とすると，

$$D_N \leq D_\alpha \quad \text{のとき，帰無仮説を採択}$$
$$D_N > D_\alpha \quad \text{のとき，帰無仮説を棄却}$$

と判定する．ここで，D_α は，(46) で与えられる分布の上 $100\alpha\%$点である．すなわち，

$$1 - 2\sum_{i=1}^{\infty}(-1)^{i-1}e^{-2i^2{D_\alpha}^2} = 1 - \alpha \tag{47}$$

を満たす．

4.4 課題

1. データファイル dat8_double_10000 からデータを取得し，「母平均が 3.4 である」との帰無仮説の下で母平均の検定を有意水準 1%で行え．なお，母分散は 5.0 とする．（分布表を用いても良い．）

2. データファイル dat9_double_101 からデータを取得し，「母平均が-2.8 である」との帰無仮説の下で母平均の検定を有意水準 1%で行え．（分布表を用いても良い．）

3. データファイル dat10_double_10000 からデータを取得し，「母集団は [0,1] 一様分布に従っている」との帰無仮説の下で χ^2 検定を有意水準 1%で行え．（分布表を用いても良い．）

   ヒント：[0,1] 区間を 10 個程度に分割し，分割後の各区間に入っているデータの個数を数える．

4. データファイル dat11_double_1000 からデータを取得し，「母集団は標準正規分布に従っている」との帰無仮説の下で Kolmogorov—Smirnov 検定を有意水準 1%で行え．

ヒント 1 ：式 (45) における sup を計算するために $(-\infty, \infty)$ 上のすべての x を調べる必要はない．高々データの個数と同じだけ調べれば十分である．なぜならば，F は滑らかな単調増加関数であり，F_N は階段状の単調増加関数だから．

ヒント 2 : 標準正規分布の累積分布関数を表すのに必要な誤差関数 erf は math.h をインクルードすれば使える．

ヒント 3 ：式 (46) は累積分布関数なので x の関数として単調増加である．そのことを利用すれば D_α の値を求めなくても検定できる．

5　並列計算と乱数検定

並列計算は必ずしも統計的分析だけの手法ではない．また，この課題で並列計算を用いるほどの計算量は必要ない．が，カリキュラムの都合上ここで並列計算の 1 メソッドである OpenMP について学び，検定の一種である乱数検定に応用してみよう．

なお、並列計算を行うにあたっては実行環境に依存する問題が生じることが予想される．本演習では大学提供の仮想環境を基準にしているので，必要に応じて利用されたい．

5.1　OpenMP

計算機上で実行すべき計算が複数あるとき，OS によって振り分けられる実行単位をプロセス，あるいは，スレッドど呼ぶ．プロセス（あるいはスレッド）の生成・消滅にはコストがかかる．近年の CPU には単体で計算を行える “コア” が複数搭載されている．並列計算では一つの計算を複数のプロセス（スレッド）に分割し，複数のコアを用いて同時に計算することで計算時間の短縮を図る．

OpenMP の特徴としては，

- 並列計算には大きく分けて，メモリ分散型とメモリ共有型がある．OpenMP は後者．複数のスレッドで計算を別々に行うが，メモリは共有される．
- C 言語，C++，Fortran でディレクトリ文を挿入する言語拡張．C 言語においては，ディレクトリ文はコンパイラに指示を与えるプラグマ文であり，それを理解しないコンパイラでは無視される．

とはいってもイメージが湧かないと思うので，実際のコードを見ていこう．(ただし，OpenMP のすべてを紹介することはできない．必要に応じて文献等を参照されたい．)

parallel ディレクティブ

```
命令 A
#pragma omp parallel
{
  命令 B
}
命令 C
```

スレッド数を p とすると，

1. 命令 A

2. 命令 B_1，命令 B_2，$\cdots$，命令 B_p
3. 命令 C

の順番で実行される．ここで，命令 B_i はスレッド i で実行される命令 B のこと．つまり，命令 B が合計 p 回，同時に実行される．例えば，

```
printf(”A”);
#pragma omp parallel
{
  printf(”B”);
}
printf(”C”);
```

を実行すると，「ABBBBBBBBC」が表示される．（B の個数は環境依存）

for ディレクティブ

```
#pragma omp parallel
{
  ...
  #pragma omp for
  for(int i=0;i<N;++i)
  {
    ...
  }
  ...
}
```

ループ変数 i について，N をスレッド数 p で割った単位ごとに，ループが実行される．例えば，N= 12，p= 3 とすると，「スレッド 1 で i= $0, 1, 2, 3$ が，スレッド 2 で i= $4, 5, 6, 7$ が，スレッド 3 で i= $8, 9, 10, 11$ が実行される」といった具合である．

sections ディレクティブ

```
#pragma omp parallel
{
  ...
  #pragma omp sections
  {
    #pragma omp section
    {
```

```
    命令 A
  }
  #pragma omp section
  {
    命令 B
  }
 }
 ...
}
```

この例では 2 つのスレッドが生成され，それぞれが命令 A と命令 B を実行する．

single ディレクティブ

```
#pragma omp parallel
{
 ...
 #pragma omp single
 {
  命令 A
 }
 ...
}
```

この例では命令 A を一つのスレッドだけが実行し，それ以外のスレッドはそれが終わるのを待つ．

barrier ディレクティブ

```
#pragma omp parallel
{
 命令 A
 #pragma omp barrier
 命令 B
}
```

すべてのスレッドで命令 A が終わるのを待つ（同期をとる）．parallel，for，

sections，single ディレクティブの終了時には，barrier ディレクティブが暗黙的に含まれる．

critical ディレクティブ

```
#pragma omp parallel
{
  ...
  #pragma omp critical
  {
    命令 A
  }
  ...
}
```

命令 A を同時に実行するのは 1 スレッドのみにする（排他制御）．

private 指示節

```
#pragma omp parallel
{
  double tmp=0;
  #pragma omp for private(tmp)
  for(int i=0;i<N;++i)
  {
    tmp=···
    ...
  }
}
```

private(変数) で変数がプライベート化される．すなわち，同じ名前の変数であっても，スレッドごとに別の変数と認識される．上の例だと，for 文が複数のスレッドに分割されるが，それぞれのスレッドで使われる tmp は別物になる（スレッド 1 の tmp，スレッド 2 の tmp，といった具合に区別される）．

reduction 指示節

```
#pragma omp parallel
{
```

```
  double total=0;
  #pragma omp for reduction(+:total)
  for(int i=0;i<N;++i)
  {
    total+=ary[i];
  }
}
```

reduction 指示節は reduction(演算子:変数) の形で書かれる．reduction 指示節により指定される変数は，一旦プライベート化されたのち，最後に指定された演算で一つの変数にまとめられる．上の例だと，まず total がプライベート化され，スレッド 1 の total，スレッド 2 の total，···，スレッド p の total に分けられる．指定されている演算が足し算なので，for 文が終わったところで，

total = (スレッド 1 の total)+(スレッド 2 の total)+···+(スレッド p の total)

として，ただ一つの total にまとめられる．

nowait 指示節

```
#pragma omp parallel
{
  double total=0;
  #pragma omp for nowait
  for(int i=0;i<N1;++i)
  {
    命令 A
  }
  for(int i=0;i<N2;++i)
  {
    命令 B
  }
}
```

前述したが#pragma omp for の後ろでは暗黙的に同期がとられる．nowait 指示節によって，そのような同期が外される．上の例では，一つ目の for 文の実行が終わったスレッドから順次，二つ目の for 文を実行する．

実行環境関数
ヘッダファイル omp.h をインクルードすれば，以下の関数が使用できる．

- omp_ get_ num_threads()　　スレッド数を取得
- omp_ set_ num_threads(int num)　　スレッド数を変更
- omp_ get_thread_num()　　スレッド ID を取得
- omp_ get_ max_threads()　　最大スレッド数を取得．
- omp_ get_ wtime()　　時刻を取得．(秒単位)

5.2　乱数検定

なんの規則性もない（見いだせない）数列である “乱数” は，情報セキュリティや暗号，モンテカルロ法などの各種アルゴリズムなど多様な領域で用いられている．一方で，実際に生成した数列が本当に乱数といえるかどうかは難しい問題であり，その問題をはっきりと解決する方法は実は（今のところ）存在しない．けれども，実用上 “乱数” が使用されている以上，多様な観点から何かしらの “評価” をしなければならない．

そのような評価方法の一つが**乱数検定**だ．やることは単純で，帰無仮説「与えられた数列は理想的な乱数である」のもとで仮説検定を行う，ただそれだけである．ここでは，前章で取り組んだ仮説検定の総合演習として乱数検定を取り扱う．

例を見てみよう．0 と 1 からなる 10-bit の数列 0010000100 が与えられたとしよう．理想的な乱数であれば 0 と 1 は半々の確率で出現するはずである．今の場合，0 が 8 個，1 が 2 個と偏っている．これを有意水準 5%の片側検定で評価しよう．

理想的な乱数の場合，1 の個数が 0 個になる確率は $1/2^{10} \sim 0.1\%$
理想的な乱数の場合，1 の個数が 1 個になる確率は ${}_{10}\mathrm{C}_1 1/2^{10} \sim 1.0\%$
理想的な乱数の場合，1 の個数が 2 個になる確率は ${}_{10}\mathrm{C}_2 1/2^{10} \sim 4.4\%$

まとめると，理想的な乱数の場合に 1 の個数が 2 個以下になる確率は約 5.5%となる．有意水準が 5%なので，この場合帰無仮説は棄却されない．(「理想的な乱数ではない」とは判定されない．)

一つ注意事項であるが，帰無仮説を棄却しなかったからと言って「いい乱数である」という保証にはならないということだ．世の中には「○○という乱数検定法をクリアしたからいい乱数です」などと謳う論文・商品があるが，そういったものを無邪気に信用してはいけない．

5.3 課題

1. OpenMP を用いたプログラムを作成・実行し，並列計算に慣れよ．
（1. については提出不要）

2. χ^2 検定を用いた乱数検定法を 1 つ考案し，0 と 1 からなる 100 万 bit の乱数に対して検定を行え．ただし，プログラムの途中で 1 箇所以上，OpenMP による並列化を行うこと．

3. 標準正規分布に従って数列を生成し，前章で取り扱った検定法を用いた乱数検定法を 1 つ考案して検定を行え．ただし，プログラムの途中で 1 箇所以上，OpenMP による並列化を行うこと．また，2. で考案した乱数検定法を使いまわさないこと．