棒グラフ、ヒストグラム　そして　統計学

--- ヒストグラムで語れるようになろう ---
--- ヒストグラムについて語れるようになろう ---

たぶん, 棒グラフを描いたことがない人はいないでしょう。しかしヒストグラムを正確に描ける人は少数派のように見えます。.そしてヒストグラムを描ける人もその多くはヒストグラムは本格的な統計解析に入る前に使う素朴な道具だと考えているようです。しかし、この講義を聞けばヒストグラムが統計学の中で中心的な位置を占める、強力な武器であることが分ります。

石黒真木夫@統計数理研究所名誉教授

2017. □. □ 15:40-17:10
2017. □. □ 15:40-17:10
At
□□県立□□高校：SS数学講座

ヒストグラム＝ヒスト＋グラム

ヒスト ＝ histo- ＝「組織」の意の連結形

Natural history ＝ 博物学 ＝ 自然の組織的記述

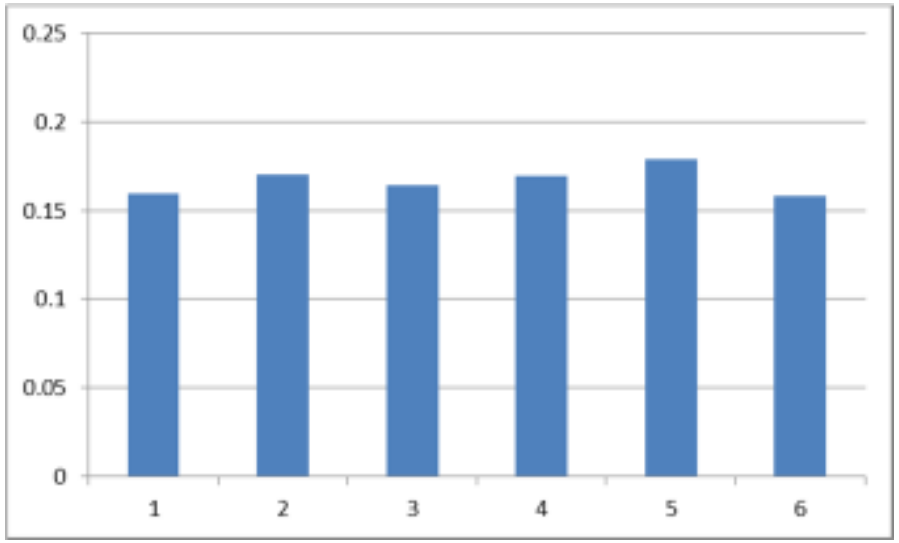
グラム ＝ -gram
＝「書いたもの」、「描いたもの」の名詞語尾

Tele- ＝ 「遠距離」の意の連結形
→ television, telescope

Telegram ＝ 電報（というものが昔、ありました）

ヒストグラム ＝ 組織的記述 → データの組織的記述

（サイコロの目の出方を表す）棒グラフの描き方



1. サイコロを何回も振って、それぞれの目が出た回数を数える。
2. 目 i が出た回数を n_i と表すことにする。
3. $n = n_1 + n_2 + n_3 + n_4 + n_5 + n_6$ を計算する。
4. 数学ではこの計算を
$$n = \sum_{i=1}^{6} n_i$$
と書く。
5. 目 i が出る確率 p_i の推定値
$$\hat{p}_i = \frac{n_i}{n} \quad (i = 1,2, \dots, K)$$
求める。サイコロの場合は $K = 6$。
6. K 本の長さ $\hat{p}_1, \hat{p}_2, \dots, \hat{p}_K$ の「棒」を描く。

(ザクロの実の重さの分布の) ヒストグラムの描き方（1～4）

1. ザクロの実を取り出す
2. 20粒づつとって目方を計る

3. 区間　[3.0, 3.4), [3.4, 3.8), [4.2, 4.6), … , [6.6, 7.0), [7.0, 7.4]　に入るデータの数を数える。
4. このとき、3.4 というデータがあったら、2 番目の区間に入れる。一般に、[a, b) という区間に　a は入れるが、b は入れない。3.3999999 は最初の区間に入れる。

(ザクロの実の重さの分布の) ヒストグラムの描き方(5〜9)

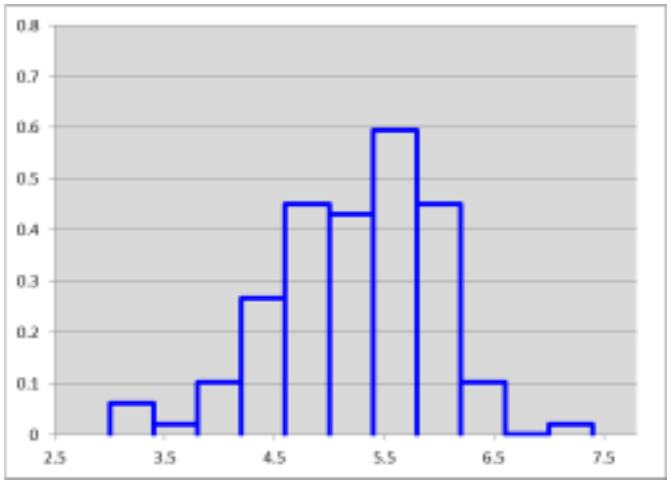
5. i 番目の区間に入ったデータの個数を n_i と表すことにする。
6. データ数

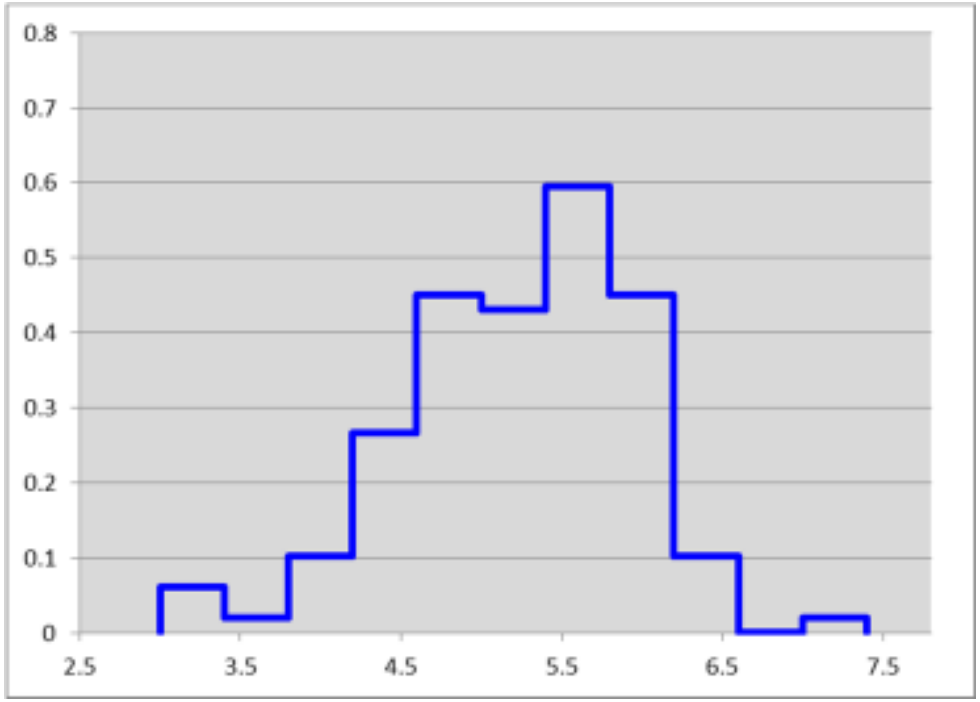
$$n = \sum_{i=1}^{K} n_i$$

を計算する。K は区間の数である。
7. データが i 番目の区間に入る確率 p_i の推定値

$$\hat{p}_i = \frac{n_i}{n} \quad (i = 1,2,\dots,K)$$

を求める。
8. i 番目の区間の上に、**面積(= 底辺の長さ × 高さ)** が $\hat{p}_i$ に等しい矩形を描く。
9. 余分な縦線を消す。



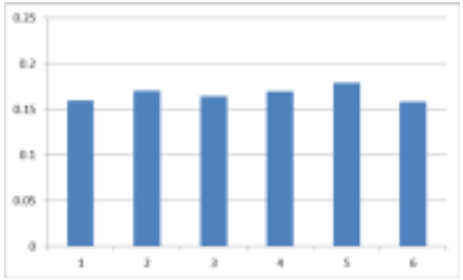


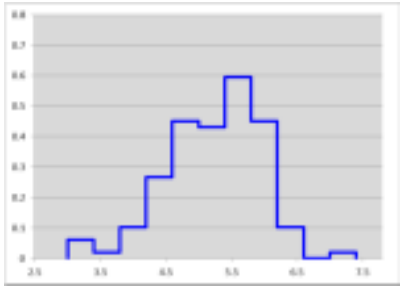
ヒストグラムの「区間」

1. $a < b$ という2つの数を両端とする閉区間　$[a,\ b]$　上のヒストグラムを作るものとする。
2. 閉区間　$[a,\ b]$　というのは、a 以上，b 以下の実数の全体である。a ，b　ともにこの区間に含まれる。
3. 「数学的」には閉区間上のヒストグラムでなくてもいいのだが、実用上はデータの最小値と最大値を端点とするヒストグラムを描きたいことが多いのでこの条件のもとで定義しておく。
4. 　$a = a_0 < a_1 < \cdots < a_{K-1} < a_K = b$　として、 k $(<$ K$)$番目の区間を半開区間 $[a_{k-1},\ a_k)$ 、最後の K 番目の区間を $[a_{K-1},\ a_K]$ とする。
5. 半開区間 $[a_{k-1,}\, a_k)$ というのは、a_{k-1} 以上，a_k 未満の実数の全体である。
6. 最初の区間を$[a_0,\ a_1]$ としてそれに続く $k > 1$ の区間を $(a_{k-1},\ a_k]$　で定義してもいいのだが、この資料では上で定義しておく。
7. $(a_{k-1},\ a_k]$ も半開区間で、a_{k-1} より大きく、a_k 以下の実数の全体である。
8. ヒストグラムは、本来、実数値データを分析するための道具であるが、100 点満点のテストの成績のような整数値データであっても、区間を作っておおまかな分析をする方が有利なことがあり、整数値データの分析にも使われる。
9. 整数値データの場合は、半開区間で区間を定義しなくてはならないということはないが、半開区間で定義して困ることもない。

ヒストグラムと棒グラフの違い、まとめ

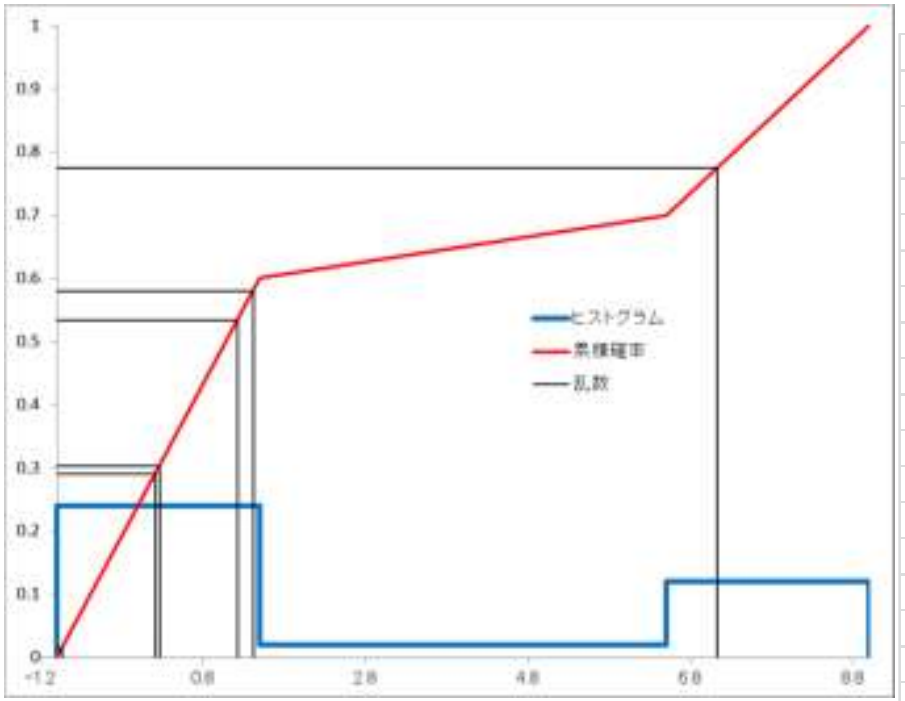
1. 棒グラフでは、いくつかカテゴリーに分けられる「結果」がそれぞれ、何回出たかを数える。サイコロの目とか、コインの裏表とか、晴・雨とか．．．
2. ヒストグラムでは、各区間に落ちる実数値データの数を数える。
3. ヒストグラムの場合の、区間のとりかたは研究目的に沿って決めるものである。
4. 「ざくろ」の例では、すべての区間が同じ「巾」だったが、区間ごとに巾を変えてもかまわない。
5. 「等幅区間」で作ったヒストグラムの場合、見かけは棒グラフに似たものになる。
6. ただし、棒グラフの場合、縦軸は確率であり、次元を持たない。「横軸の値」はデータの「カテゴリー」であり、これも次元を持たない。
7. ヒストグラムの場合、横軸が次元を持つことがある。縦軸は「1/横軸の次元」をもつ。
8. 「ザクロ」の場合の横軸は1グラム単位で計測された質量の次元をもち、縦軸の次元は「1/グラム」である。
9. 棒グラフの値が 1　を超えることはないが、ヒストグラムの値は 1 を超えることもある。
10. ヒストグラムでは、隣り合う区間の間に空白を置かない。棒グラフでは隣の棒との間に空白を置く。





ヒストグラムを用いたシミュレーション(1～7)

1. 区間 $[-1.\,1.5),[1.5,6.5),[6.5,9.0]$ の上で定義されたヒストグラム（右図青線）を例とする。
2. 3つの区間での確率密度は、それぞれ、0.24, 0.02, 0.12 とする。
3. $$h(s)=\begin{cases}0.24 & -1\le s<1.5\\ 0.02 & 1.5\le s<6.5\\ 0.12 & 6.5\le s\le 9.0\end{cases}$$
   とする。
4. 累積確率関数
   $$\mathrm{H(x)}=\int_{-1}^{x}h(s)ds$$
   を求める。
5. $H(x)$ は 区間 $[-1,x]$ と青線にはさまれた領域の面積にほかならず、積分を習っていなくても計算できる。
6. $H(-1)=0,$
   $H(1.5)=0.24\times 2.5=0.6$
   $H(6.5)=0.6+0.02\times 5.0=0.7$
   $H(9.0)=1$



7. $-1\le x\le 9$ に対する $H(x)$ は 4点 (-1, 0), (1.5, 0.6), (6.5, 0.7), (9.0, 1.0) を結ぶ赤の折れ線で表される関数となる。

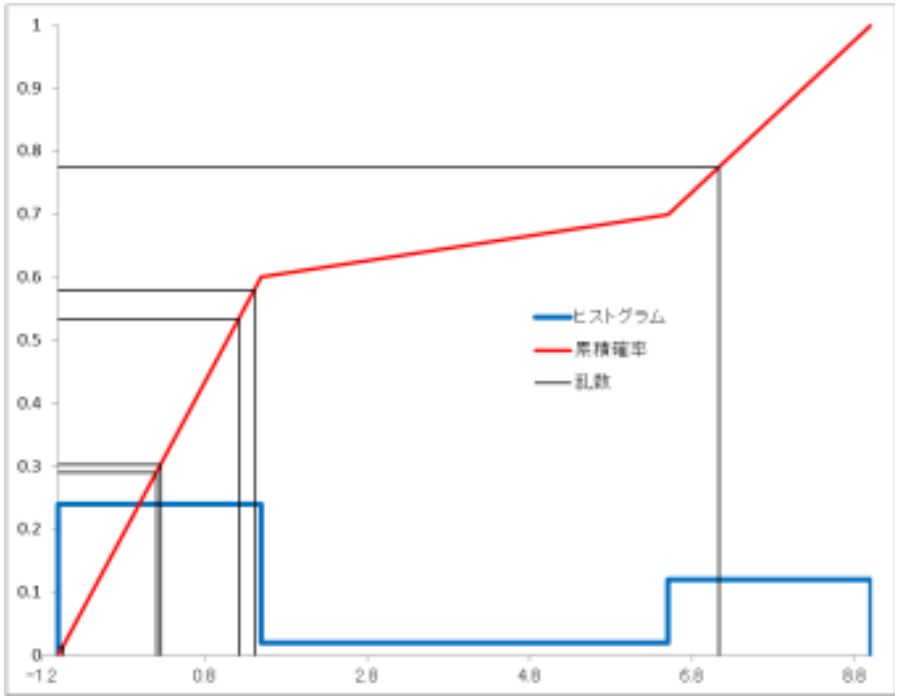
ヒストグラムを用いたシミュレーション(8～12)

8. 0 以上 1 以下の値をとる乱数 u_1 を発生させる。
9. 縦軸上の u_1 の位置から右に進み、赤い線 にぶつかったら、下に降りて横軸の値を読む。この値を$H^{-1}(u_1)$ で表すことにする。
10. 8,9 の手順を繰り返し、$H^{-1}(u_2)$, $H^{-1}(u_3)$, $H^{-1}(u_4)$, $H^{-1}(u_5)$, …. を発生させる。
11. 右の図から

$$H^{-1}(0.533424) = 1.2226$$
$$H^{-1}(0.579519) = 1.414661$$
$$H^{-1}(0.289562) = 0.20651$$
$$H^{-1}(0.301948) = 0.258117$$
$$H^{-1}(0.77474) = 7.122834$$
$$H^{-1}(0.014018) = -0.94159$$

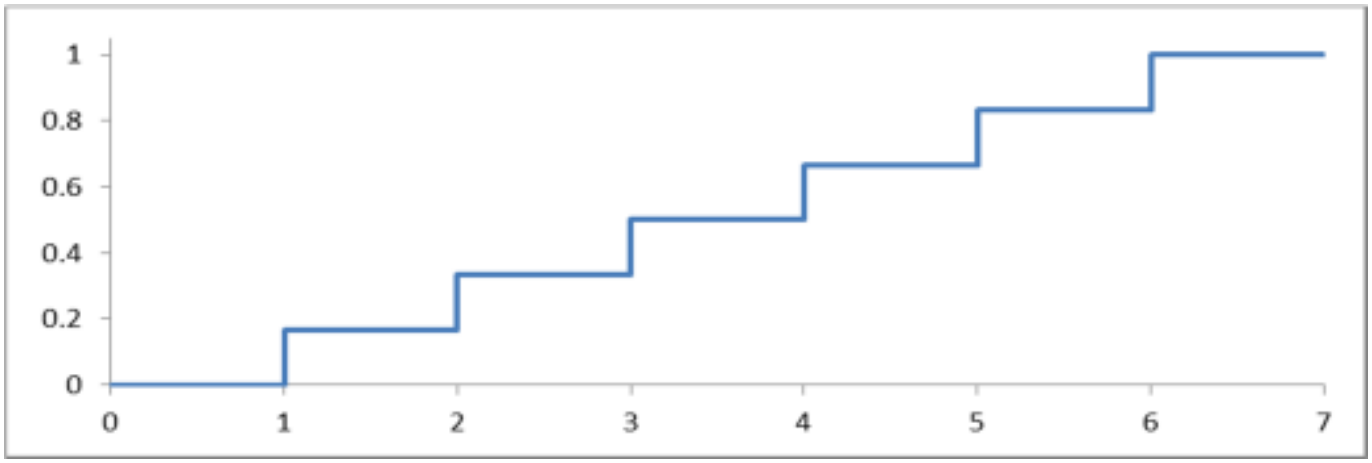
が読み取れる。

12. 1.5～6.5 の範囲の値をとるデータが出ていない理由は明らかだろう。

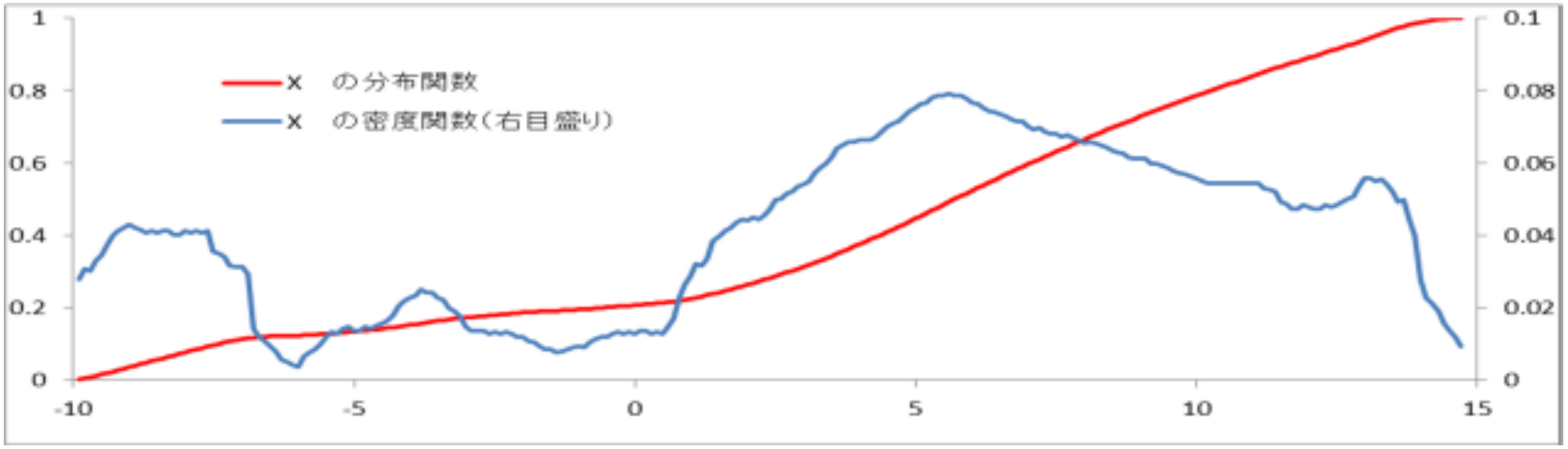


「離散データ」のシミュレーション

こんな「確率分布関数」を使えば、サイコロの目のようなデータを発生するシミュr-ションが可能である。



「神様」のシミュレーション

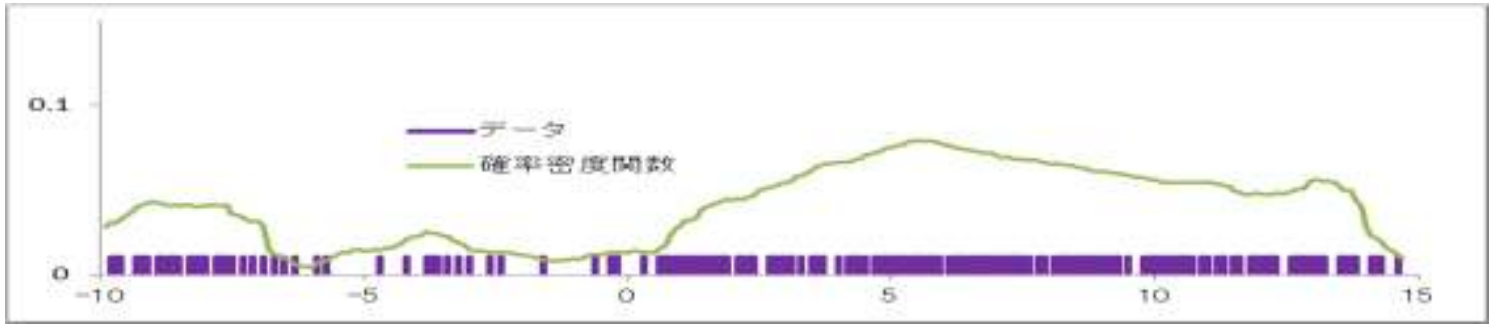


1. データというものは、「神様」が確率密度関数にもとづくシミュレーションによって発生させているのだと考えることにする。
2. ただし、神様が使っている確率密度関数は上の図のように複雑な形をしているものかもしれない。
3. こんな確率密度関数から作られる累積確率関数は、ヒストグラムの場合にように、折れ線をつなぐような形ではできず積分

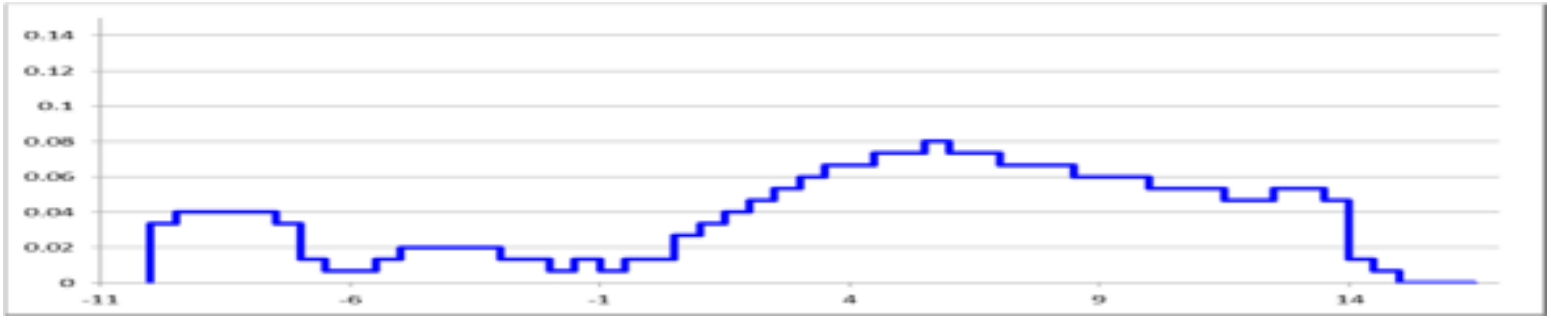
$$H(x) = \int_{-1}^{x} h(s)ds$$

を計算して求める必要があるが、神様は積分できるだろう。
4. 神様がデータ発生に使う累積積分関数を確率分布関数という。
5. 確率分布関数が出来てしまえば、シミュレーションは人間がやるのと同じ方法でできる。

統計学

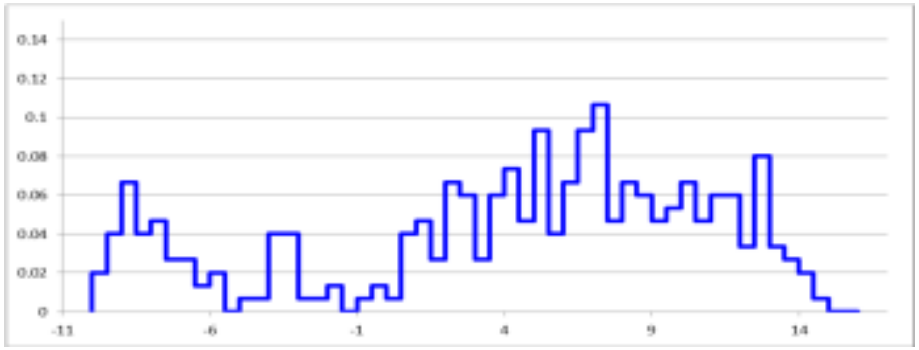


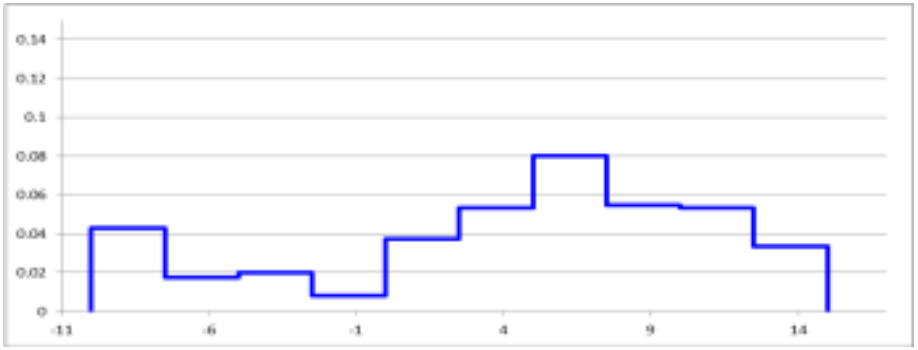
1. 世の中は不確実性に満ちているが、神様のすることがなんらかの確率密度関数にもとづくシミュレーションの結果だとすれば、神様が使っている確率密度関数を知ることで不確実性を軽減できるに違いない。
2. 統計学はデータにもとづいて、そのデータを生成している確率密度関数を推測することを目的とする学問であると考えることができる。
3. ヒストグラムは、どんな形の確率密度関数でも近似的に表現できる可能性をもった道具である。たとえば、上の確率密度関数をヒストグラムで近似することが可能である。

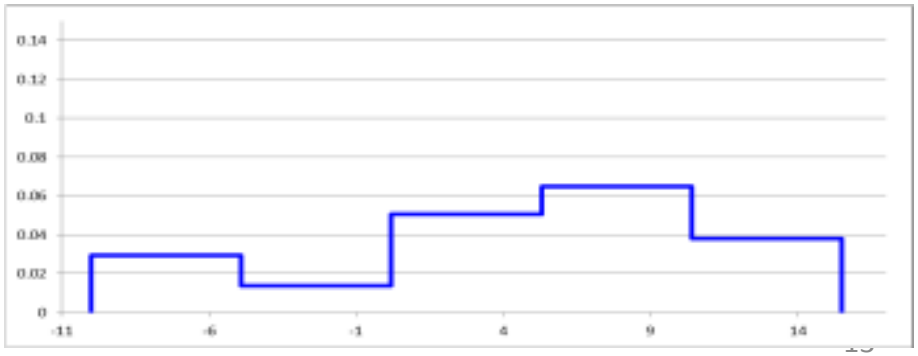


ヒストグラムのオーバーフィットとアンダーフィット

1. 区間の巾をどの程度にすべきかは、データを発生している確率密度関数の形に依存する。
2. しかし、確率密度関数が分かっているならそもそもヒストグラムを描く必要がない。
3. 区間の巾をせまくすれば、どんな確率密度関数のかたちでもなぞることはできるが、データ数が十分でないと、各区間における確率密度が正確に推定できなくなる。
4. 区間の巾が狭すぎて確率密度の推定の信頼度が落ちる状況をオーバーフィットという。
5. 逆に、区間の巾が広すぎて、確率密度関数の形をなぞれなくなっている状況をアンダーフィットという。
6. データ数に応じた適当な区間巾を決める方法があるが、残念ながら、今回の講義ではそこまで行けない。





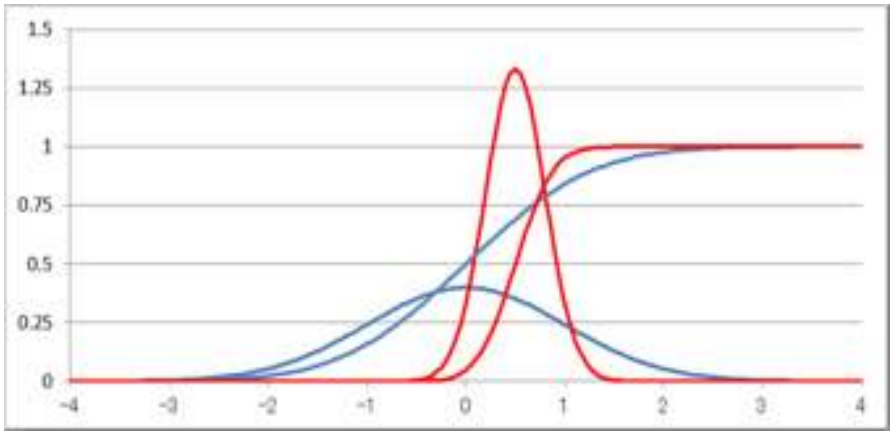


ヒストグラムと正規分布

1. 神様が正規分布を使うこともある。正規分布の確率密度関数は μ と σ^2、という2つのパラメータで形が決まる関数

$$g(x|\mu,\sigma^2)=\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

で定義される。



2. 図に、パラメータの設定がちがう2つの正規分布の密度関数と、対応する確率分布関数が描かれている。
3. データ $\{x_1,x_2,\ldots,x_n\}$ が与えられたときに、ヒストグラムで、確率密度関数を推定する場合には、区間群$\{[a_0,a_1),[a_1,a_2),\ldots,[a_{K-1},a_K]\}$に落ちるデータの数を数える。
4. 正規分布モデルのあてはめで推定する場合には、

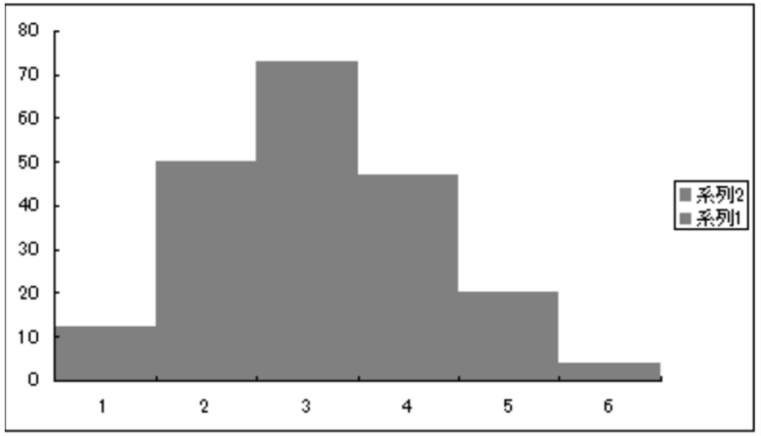
$$\hat{\mu}=\frac{1}{n}\left(\sum_{i=1}^{n}x_i\right),\qquad \widehat{\sigma^2}=\frac{1}{n}\left(\sum_{i=1}^{n}(x_i-\hat{\mu})^2\right)$$

と計算する。

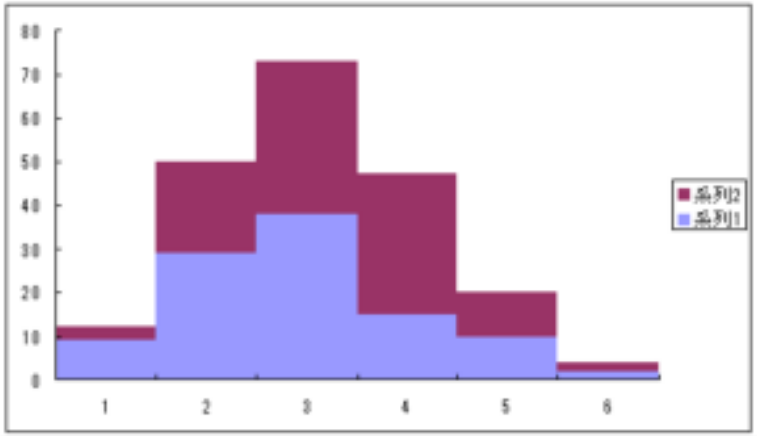
4. ヒストグラムは、「神様」が、実は、正規分布を使っていなかった場合にも確率密度関数をうまくなぞれるかもしれないが、データの数が十分でないとうまく働かない。
5. 正規分布は、かなり少ないデータでデータの概要を把握するが、その姿は「神様」の確率密度関数と似ても似つかないものかもしれない。

ヒストグラムの分解

1. ヒストグラムは、区間群を定義し、各区間に落ちるデータの数を勘定する方法で描かれる、
2. １次元直線上の１次元区間群でなく、２次元平面上の「２次元区間群」を用意して各区間に落ちるデータの数を勘定する方法でヒストグラムの分解が可能となる。
3. 多変量データにおいて、ある変量の出方が、他の変量の値とどう関係しているかをヒストグラムの分解で調べることができる。
4. 区間の次元を上げると、各区間に落ちるデータも数が減る。この場合にも区間巾の選択の場合と同じオーバーフィットとアンダーフィットの問題が生じる。
5. AICという統計量を使うことで適当な分解を探すことができるが、ここで話す時間がない。参考文献参照。
6. 区間の巾の選択にもAICが使われる。

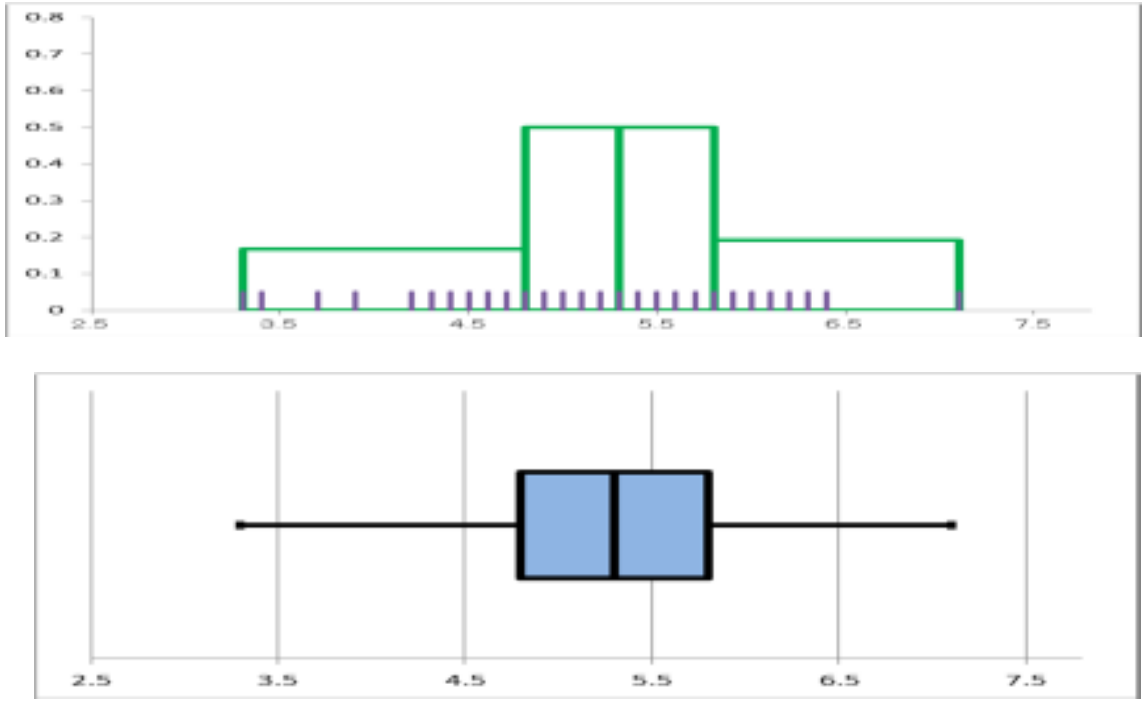


X1=1	9	29	38	15	10	2	AIC = −0.8
X1=2	3	21	35	32	10	2	
total	12	50	73	47	20	4	AIC = 0.0

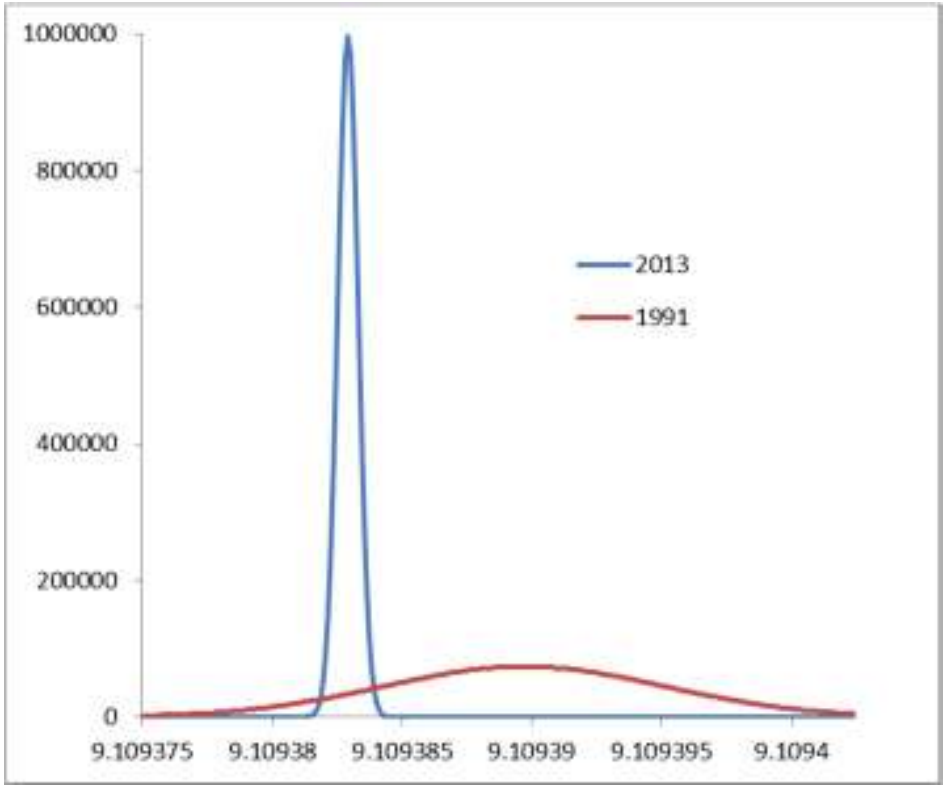


Box plot/箱ひげ図

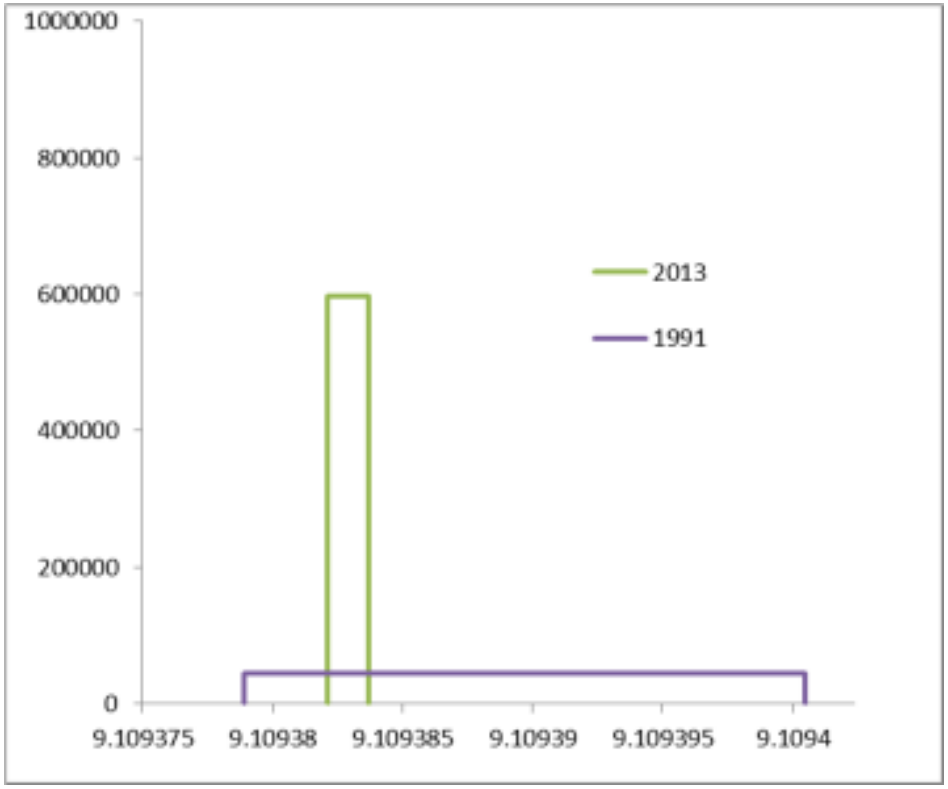
1. ヒストグラムの区間の巾を調節して、等面積の矩形が並んだ形にすることができます。
2. 「ザクロデータ」から面積 0.25 の矩形4個を並べた形のヒストグラムを作ると右のようなものが出来上がります。
3. 矩形の面積が分かっているので、矩形の位置が分かれば高さは計算できます。
4. 「箱ひげ図」という表示法がありますが、これは、4個の矩形からなるヒストグラムによるデータ表示と等価です。

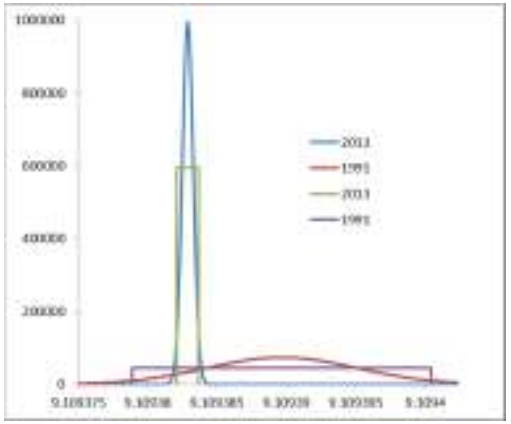


電子の質量の測定値の確率密度関数



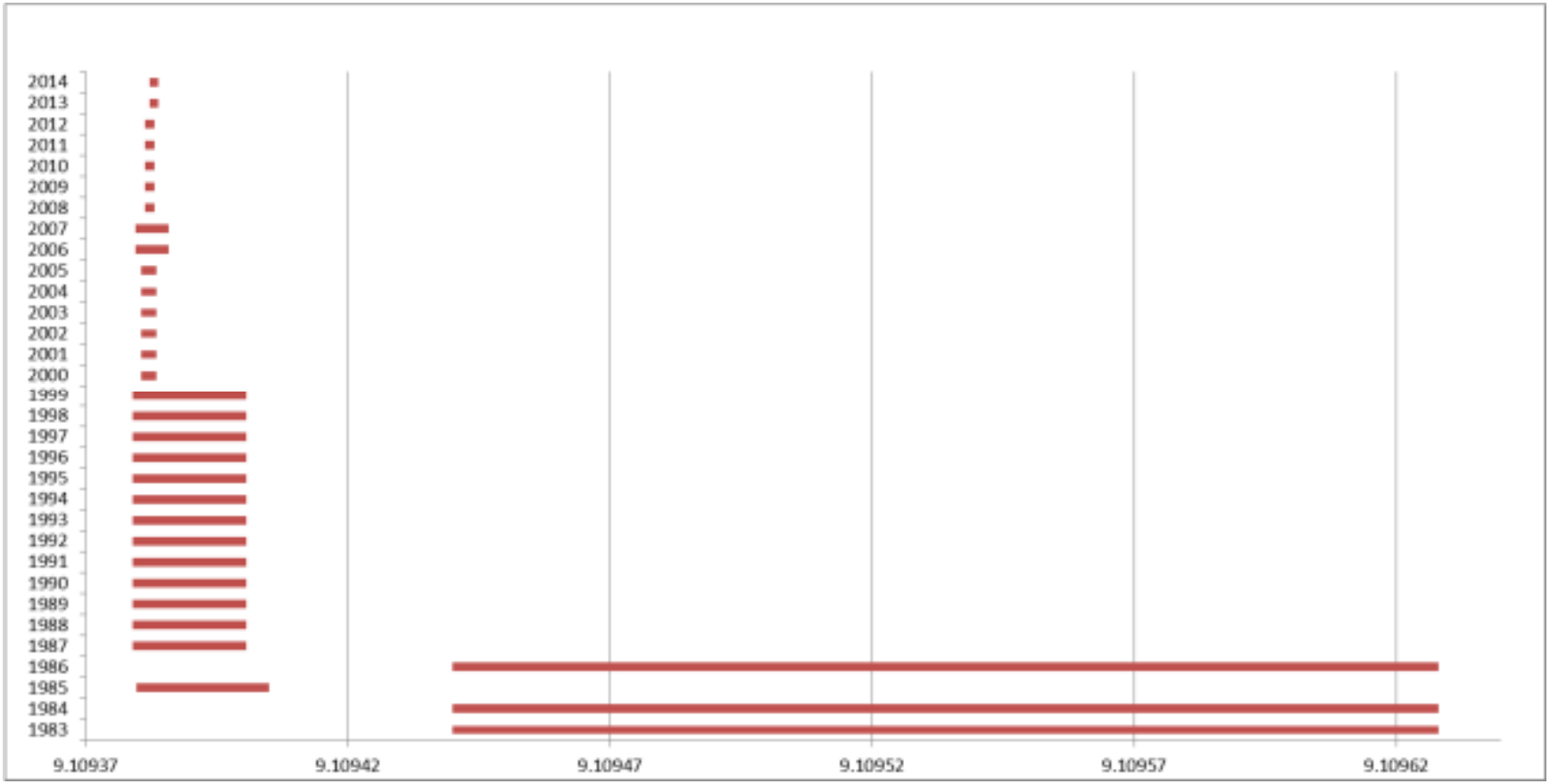
電子の質量の測定値の「ヒストグラム」



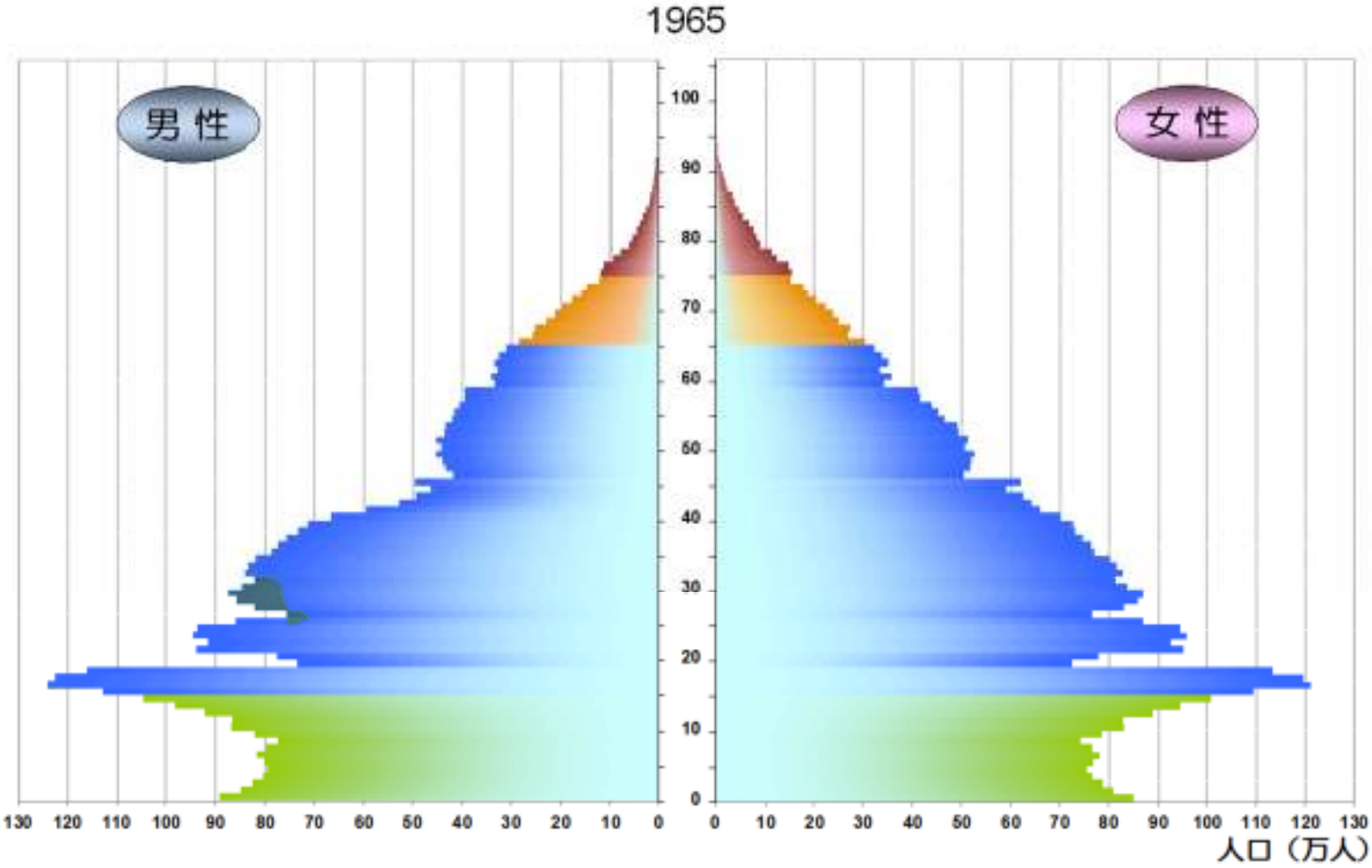


1. 誤差巾をつけたデータ表示は「矩形１個のヒストグラム」と見なすことができる。
2. ただし、「標準偏差の２倍で誤差巾を表示するヒストグラム」の面積は、約 0.95 である。

理科年表記載の電子の質量の歴史

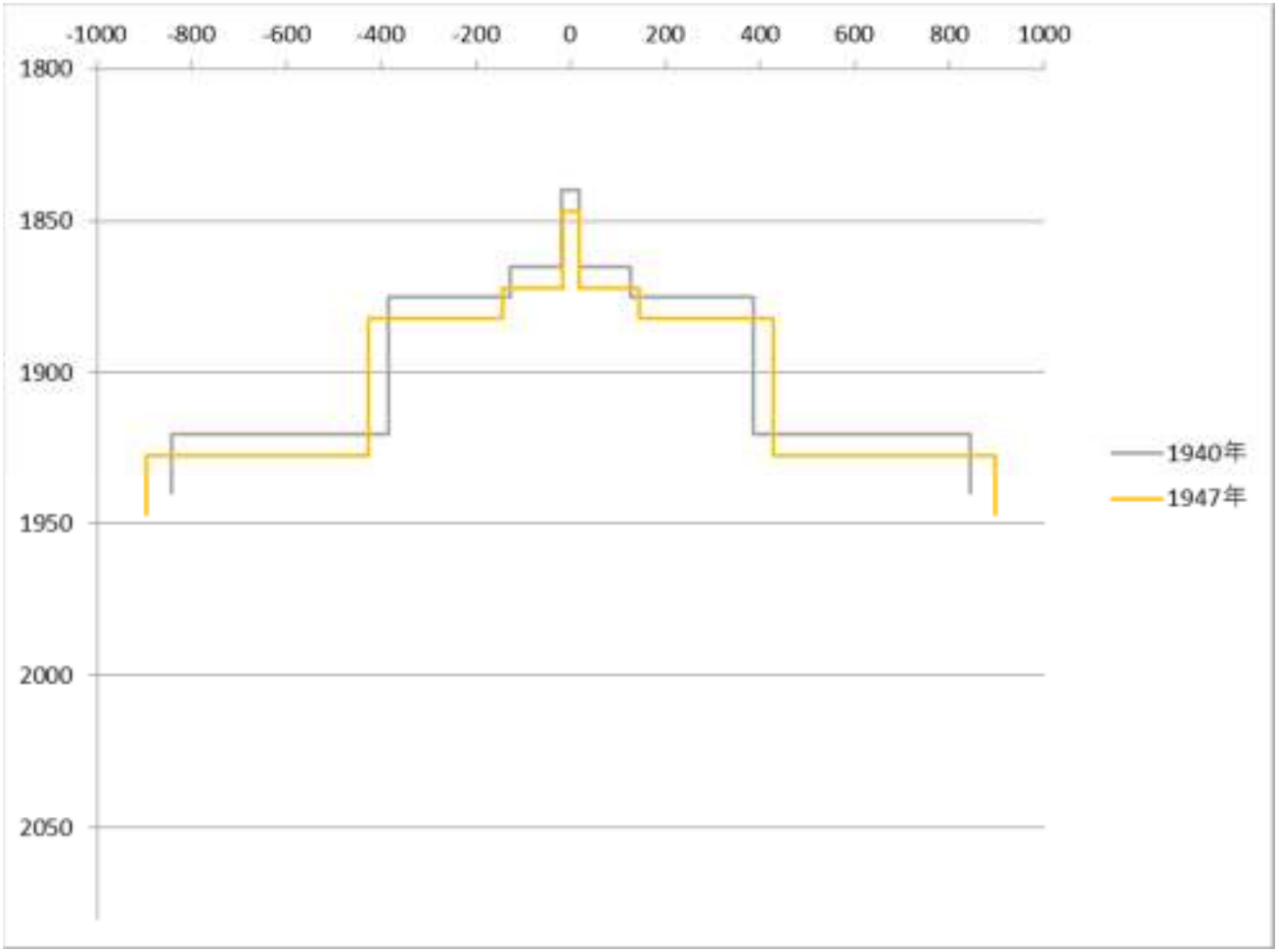


人口ピラミッドもヒストグラム

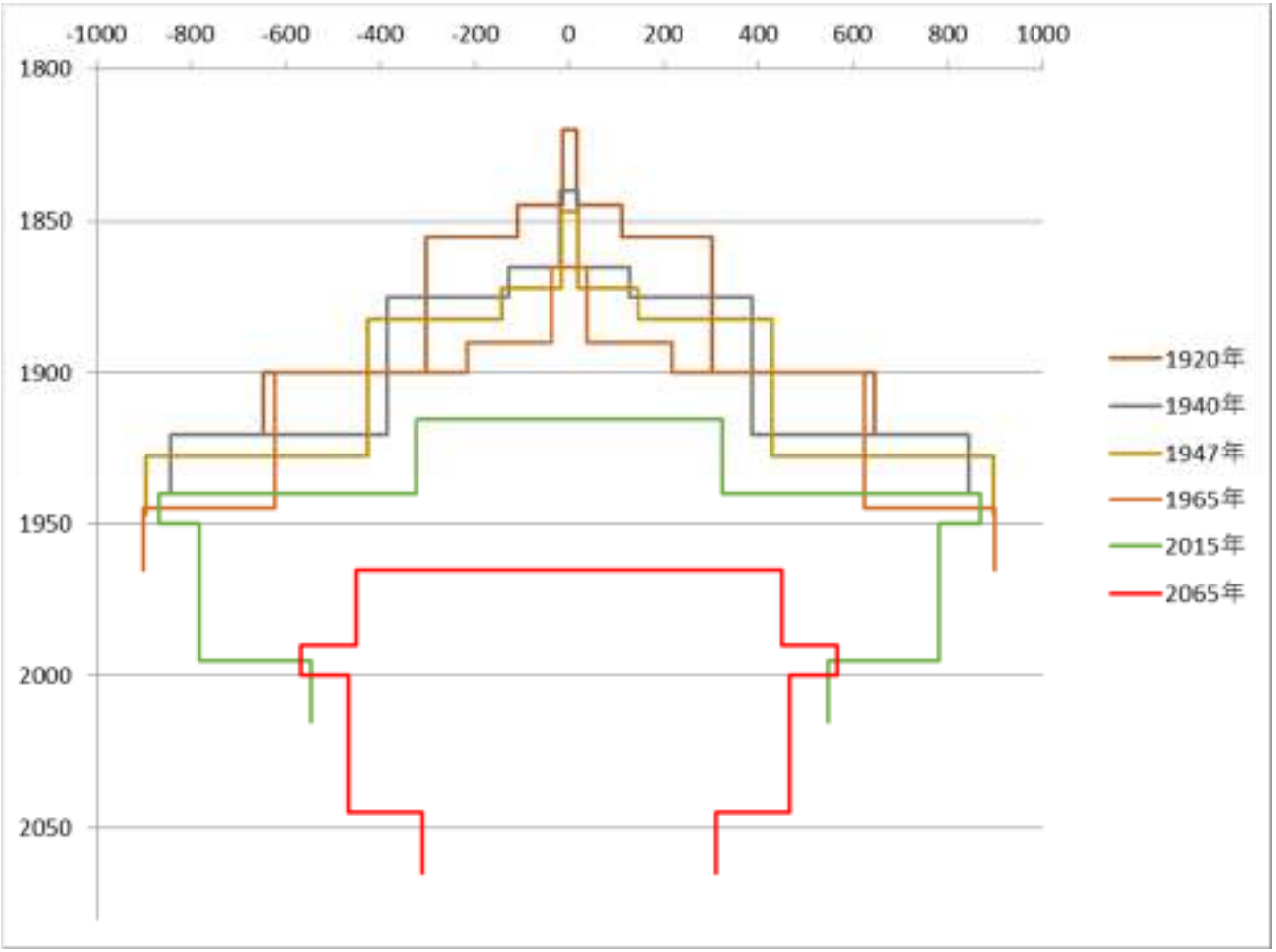


出典：国立社会保障・人口問題研究所ホームページ （http://www.ipss.go.jp/）

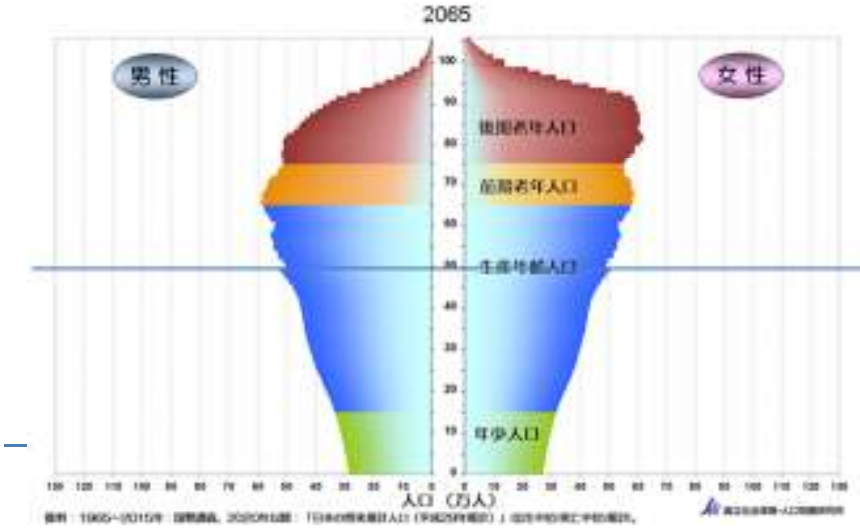
昔は本当にピラミッド型だった

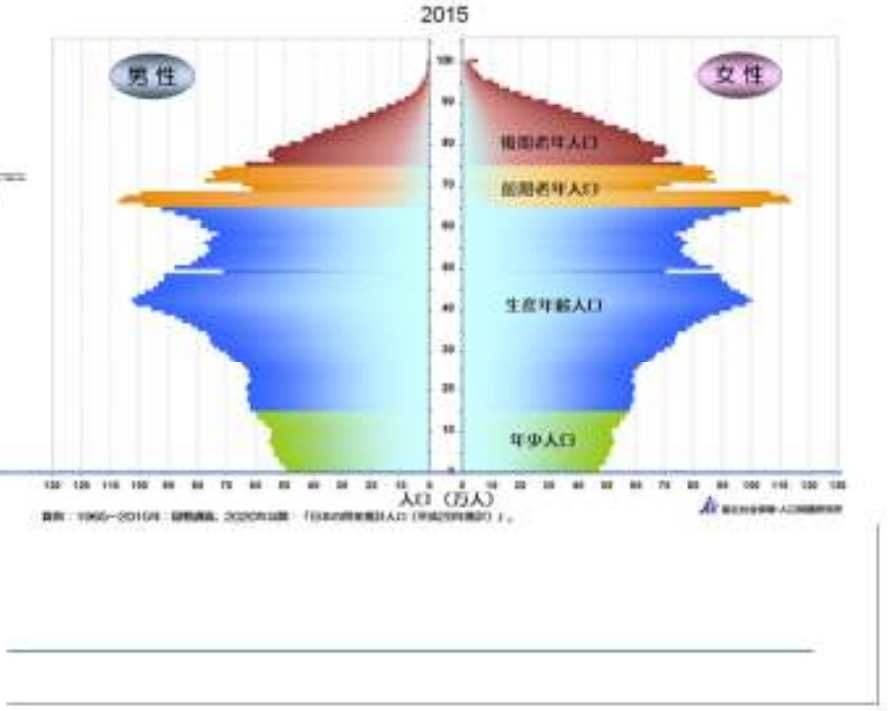


少子化！

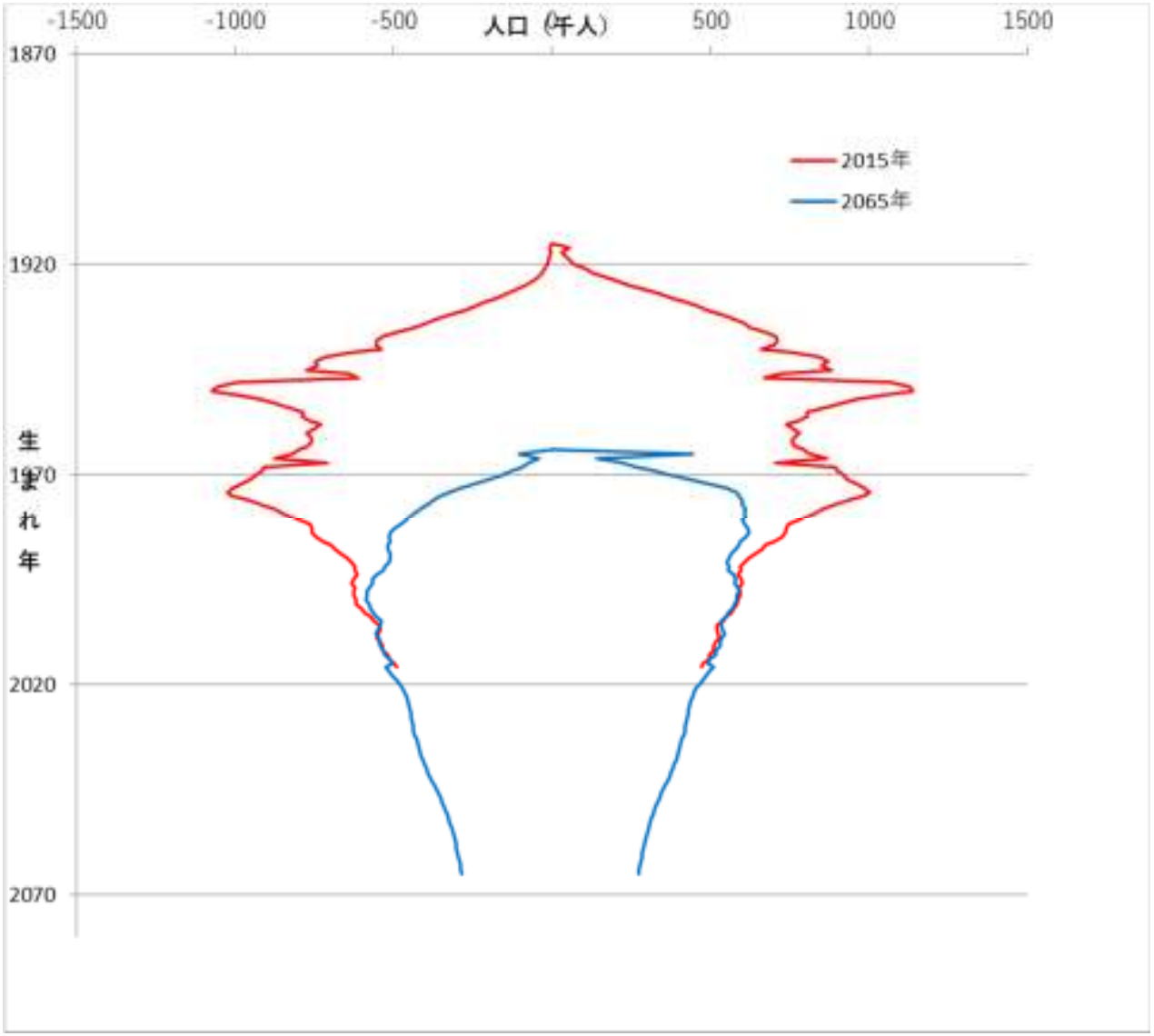

-1000
-800
-600
-400
-200
0
200
400
600
800
1000
1800
1850
1900
1950
2000
2050
1920年
1940年
1947年
1965年
2015年
2065年

2015年と2065年

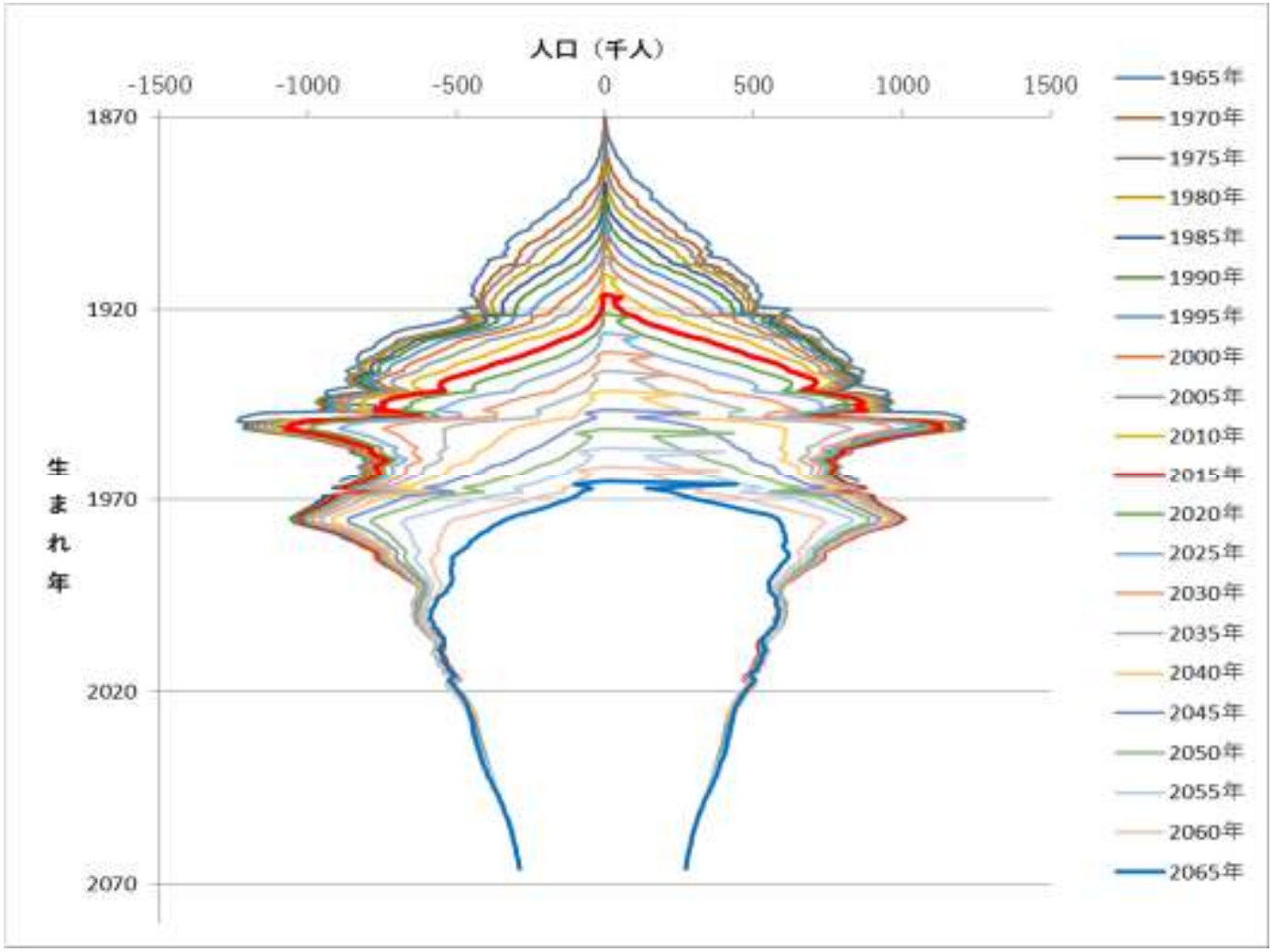




2015年と2065年

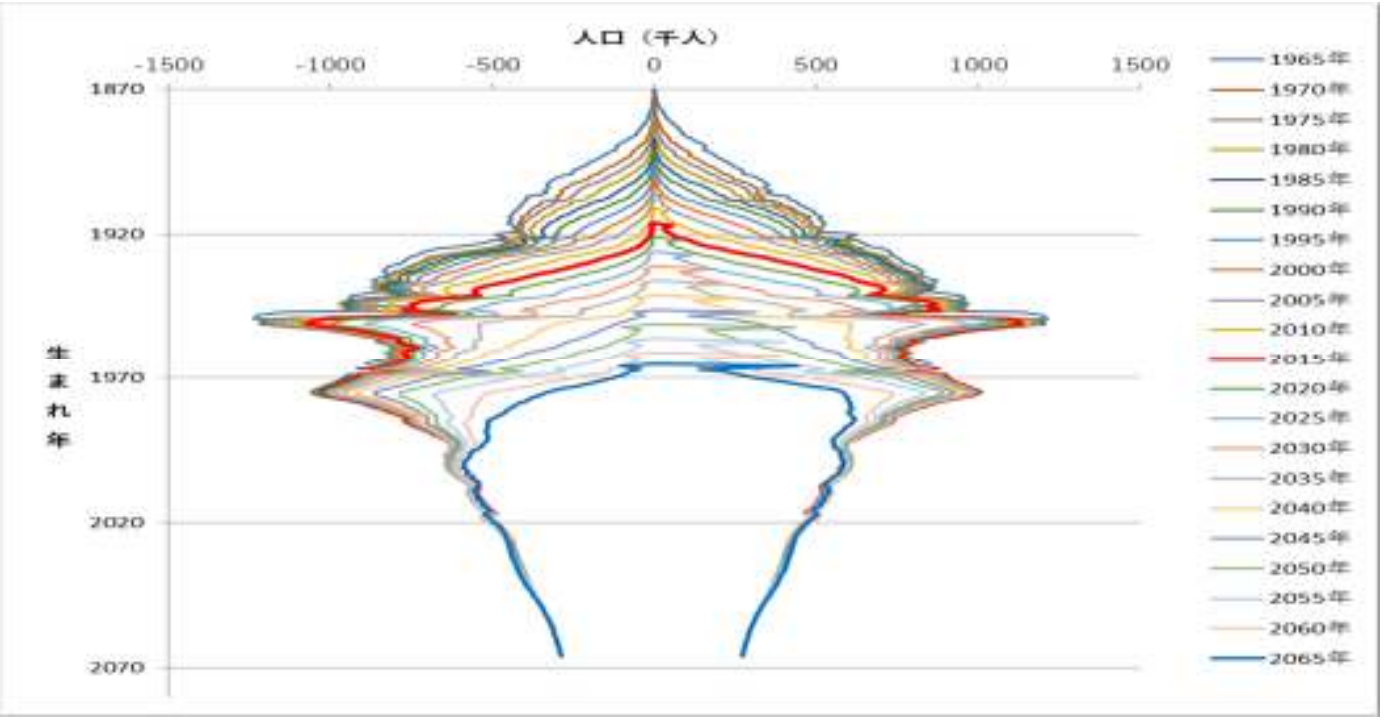


1965年→2065年



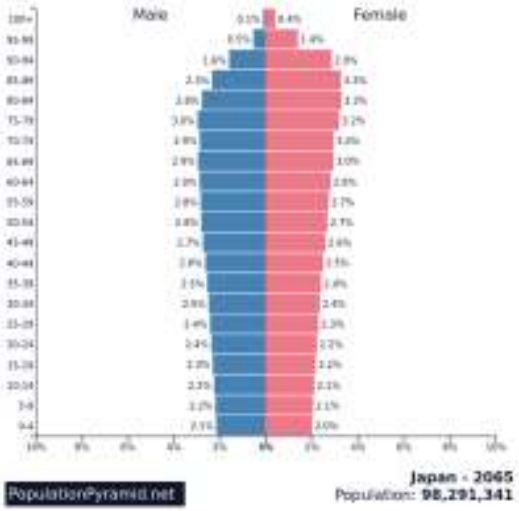
出典：国立社会保障・人口問題研究所ホームページ （http://www.ipss.go.jp/）

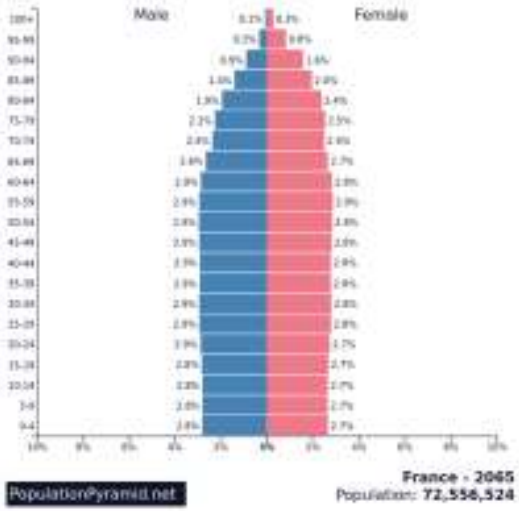
人口ピラミッドの見方

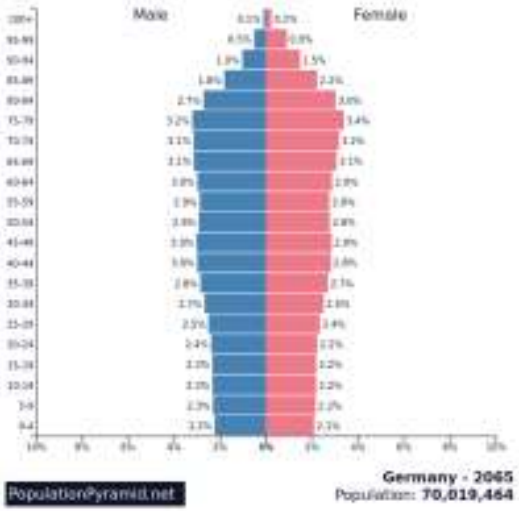


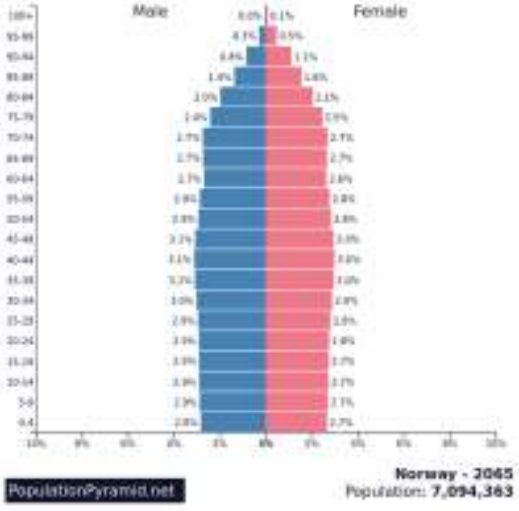
1. 縦軸は「生まれ年」を示す。
2. 図の左側は男性の生まれ年軸上の人口密度
3. 右側は女性の人口密度である。
4. この図には 1965 年現在の人口密度から 2065 年現在の予想人口密度までが描かれている。
5. 2015年以前の調査に登場していた1920年以前生まれの人が2015年以降の調査には登場しなくなっている。
6. 調査年が新しいほど線は内側にはいる。

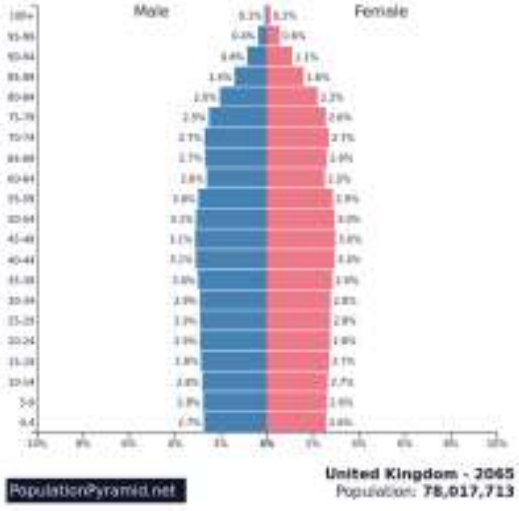
2065年の予想人口ピラミッド

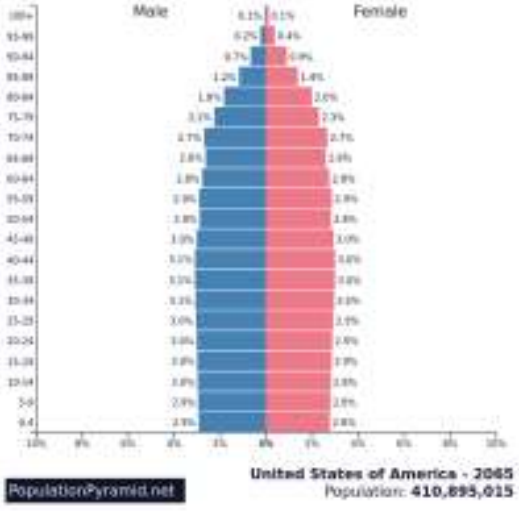












https://www.populationpyramid.net/

参考文献

1. このPowerPoint　ファイルのスライド 5, 8, 9 を描いたエクセルソフトをひとつにまとめて作ったファイル statcycle.xls　が統計数理研究所　学術研究リポジトリ（ https://ismrepo.ism.ac.jp/ ）にある。
2. CATDAPマニュアル：http://hdl.handle.net/10787/3821
3. 坂元・石黒・北川（１９８３）.情報量統計学、共立出版