HOKUGA 北海学園学術情報リポジトリ
学校法人北海学園
北海学園大学
北海商科大学

タイトル	北海道主要都市の観光入込客数に関する基礎的研究 (ベイズ統計の導入)
著者	上浦，正樹; KAMIURA，Masaki
引用	開発論集(108): 1-17
発行日	2021-09-30

Hokkai-Gakuen Organization of Knowledge Ubiquitous through Gaining Archives.

開発論集　第108号　1-17(2021年 9 月)

北海道主要都市の観光入込客数に関する基礎的研究（ベイズ統計の導入）

上　浦　正　樹*

1　はじめに

本研究は「北海道における持続可能な地域社会の形成方策に関する総合的研究」において「JR 北海道における「単独で維持困難線区」に属する富良，北見，稚内の地方創生に資する交通インフラに関する研究」を対象としている。本論文は，観光入込客数や日平均気温の月平均値など現時点で入手可能なデータに基づき，北見，稚内の各都市に加え北海道の主要な 4 都市（札幌，旭川，函館，帯広）の合計 6 都市の統計データに基づき都市の観光規模について検討を加える。その目的は，持続可能な地域の形成にあたり観光は重要な要素にあたるとの考えから，観光入込客数から都市の観光規模を推定することにある。加えて一般的に北海道では気温が高い時期に観光客が増えるとされていることから，観光入込客数と日平均気温の関係を検討することとする。これらの検討では少ない情報から精度よく他の情報を推定する方法を確立することが前提である。これらのデータの解析にあたり，従来から定量的な解析に必要な統計上の検定において従来の統計学で使われる概念（p 値）が，2016 年に世界的な統計学の団体であるアメリカ統計学会（ASA）から標本数に影響を受けるなどの欠陥[1)] があるとの指摘を受けたことから，最近になって注目を集めているベイズ統計学を導入することした。

2　従来の統計学とベイズ統計学

2.1　p 値の否定

近年の統計学は科学に発展と関係しながら急速に進歩し，PC の性能向上と相まって統計学ユーザーである我々一般の研究者にも統計的検定に使われる概念が利用しやすいものとなってきた（補足 A)。ところが，2016 年 3 月に世界的な統計学の団体であるアメリカ統計学会（ASA）が従来の統計的検定に使われる概念（p 値）に否定的な声明を出した[1), 2)]。また 2019 年 3 月に発行された権威のある科学雑誌 *Nature* には「統計的有意性を引退させよう」とのコメント論文に 800 人以上の科学者が署名していた。

以上のように現在では 19 世紀後半から約 150 年以上に亘って使われてきた p 値の概念が否

*（かみうら　まさき）北海学園大学開発研究所特別研究員

定されている。この理由として p 値が標本数に影響を受けて変化し一定の値にならないこと，標本の分布を正規分布など理論曲線分布に合わせている点がある。補足 A の図 A2 でもわかるようにこれは鉄まくらぎや木まくらぎにおいてその分布形状を理論曲線に合わせると誤差が生じる可能性がある。次に p 値の根拠が明確でないことがある。この理論を完成させたフィシャーは例として p 値を 0.1 としているが，その根拠は示していない。とはいえ，100 年以上の年数の間で p 値は多くの研究部門などで活用され，統計学ユーザーにとっては，統計的検定に標準正規分布表や p 値の目安を 0.1 している。それほど p 値が求やすいというメリットがあるものと考えられる。このような背景から現在は，統計学ユーザーに利用しやすい新たな統計手法の開発が望まれている。

2.2　ベイズ統計学の導入

補足 A で示したように，フィッシャー流やネイマン流などの従来法では，統計モデルに対して平均値などのパラメータが定数であることを前提に観測結果 D からパラメータ Θ を推定し，これから母集団の母数（平均値，標準偏差）を推定している。これに対してベイズ統計学では，パラメータを確率分布として捉える立場をとっている。この場合に確率分布の種類が分かればそれを採用するが，情報がない場合ではある範囲内での一様分布か，分散が大きい場合は正規分布とすることが多い。本論文では，この確率分布を事前確率分布とし，一様分布を採用する。次に，ある事象から得られた観測結果 D からみると D は確定していることから定数と見なし，パラメータは 1 つの値とするパラメータ Θ ではなく D の条件付きの確率変数としてのパラメータ θ_i とする。よって補足 B の式（B5）から

$$P(\theta_i|D)=\frac{P(D|\theta_i)P(\theta_i)}{\sum_{1}^{n}P(D|\theta_i)P(\theta_i)} \qquad (1)$$

となる。これから式(1)の分母を連続関数として，補足 C の式（C4）から

$$P(\theta|D)\propto P(D|\theta)P(\theta) \qquad (2)$$

と示す。

この式で $P(\theta)$ はデータの発生以前からパラメータと存在しているので前述したように事前確率 $P(\theta)$ とし，その後取得されたデータの条件でパラメータが決まるので $P(\theta|D)$ をパラメータの事後確率とする。また $P(D|\theta)$ はパラメータ θ がデータに与える影響の確率（関数）である。よって統計モデルに尤（もっと）も当てはまりのよいパラメータとデータの程度を尤度と呼んでいる。この尤度関数 $P(D|\theta)$ には正規分布やポアソン分布などが考えられる。

以上から，ベイズ統計学ではデータを用いて尤度関数から，確率分布している統計モデルのパラメータを確定する。これをパラメータ推定とする。なお，本論文において計算ソフト R を使用してパラメータ推定にマルコフ連鎖モンテカルロ（MCMC：Markov chain Monte

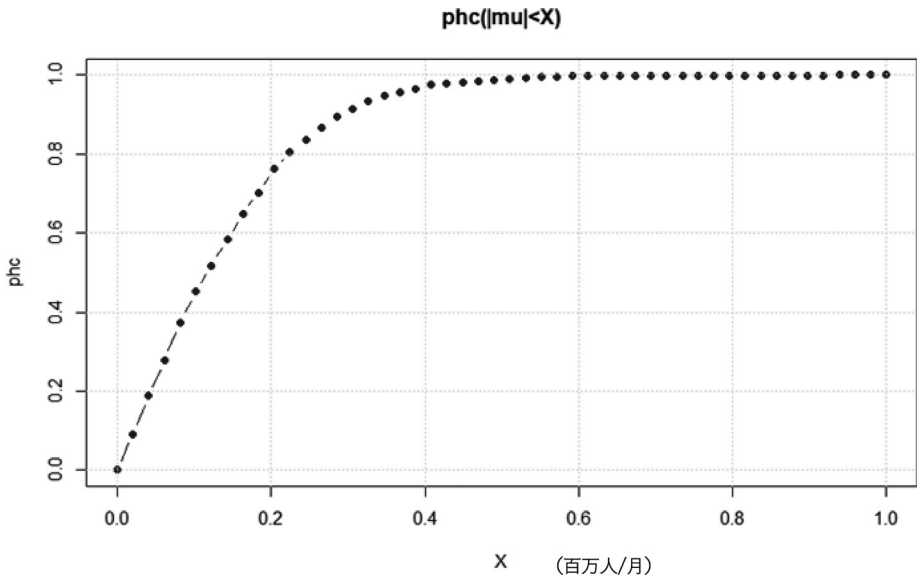

図１　月ごとの入れ込み観光者数の差と phc 値

Carlo）法を用いる。補足 D に計算ソフト R について説明している。これによりパラメータ推定の手順を補足 E に示す。

2.3　従来の統計学とベイズ統計学の比較

平均，分散，正規分布などはベイズ統計学では従来の統計学と同じように用いる。一方，その違いはベイズ統計学では信用区間と従来の統計学の信頼区間の違いに見られる。例えば，図 1 は補足 E により求められたもので，H28 年度と H31 年度の入れ込み観光者数の月ごとの差（x）とその的中確率（phc：Probability that Hypothesis is Correct）を示している。

例えば，その差を信用区間の最大値 c とすると，c＝0.6 百万人/月まで認めるとその確率はほぼ 1 になる。c＝0.43 百万人/月まで認めるとその確率を 90％となる。将来の予測に活用する場合に，この差を用いるときは信用区間として妥当な c 値を定めることになる。一方，このような phc 値における c の値はフィシャー流では信頼区間の p 値に対応している。この p 値は 2.1 で示した 0.1 または 90％の確率で一般に使用されており，固定化されている。その意味で一般の統計学では信頼区間の変更が可能な構成になっていない。

さらに，一般の統計学ではデータから標本の平均と標準偏差とデータ数から図表を用いることで統計ユーザーにとって求めやすい方法である。一方，ベイズ統計学では MCMC 法のプログラムを使用するなどで統計ユーザーにとっては，まだ習熟が必要な方法である。

3　北海道主要都市の観光入込客数

3.1　観光入込客統計[3)]

北見，稚内の各都市に加え北海道の主要都市における観光入込客は公表されているデータとした。国土交通省観光庁のホームページ[3)]によると，観光入込客統計は平成 21 年 12 月に策定した「観光入込客統計に関する共通基準」に基づき都道府県が調査を実施しているものである。それまでは各都道府県独自の手法により観光入込客を調査・集計されていたが，この基準により全国で統一したものとなり，さらに公表されるに至り，各都道府県の「観光消費額単価」「観光消費額」等への対比が可能となった。本論文では，北海道の各市で公表されている観光入込客統計を各市にホームページより入手した。

3.2　研究対象の年度の選定

本論文ではコロナ収束後に平常に戻る前提で，過去の平常状態のデータを基本とする。最近の 5 年間の平成 28 年～令和 3 年の北海道の観光を振り返ってみると，この 1 年あまりはコロナ感染による自粛があったことから対象とはしない。その前年の平成 31 年では特別な異変が見られなかったので平常状態と見なす。これと比較する年度としてその前年の平成 30 年には胆振東部地震が発生したことから，この年を外し，さらに遡り平成 28 年度のデータを用いることとした。以上から，観光入込客統計の平成 31 年度と平成 28 年度のデータを用いることとした。なお観光入込客統計では月ごとのデータまで入手できるので，これらのデータを評価することとした。また，研究対象の都市は，観光入込客統計をホームページから入手可能であることを前提とした。その中から人口，支庁のある都市などから選択し，北見[4)]，帯広[5)]，札幌[6)]，稚内[7)]，旭川[8)]，函館[9)]とした。なお，釧路は観光入込客統計がホームページに掲載されているが，月ごとの原データではなく 3 か月分として加工されていたため用いないこととした。

3.3　平成 31 年度と平成 28 年度の観光入込客数

選定した 6 都市のおける平成 28 年度（以下 H28 とする）と平成 31 年度（以下 H31 とする）の年間の観光入込客統計の比較を**図 2** に示す。この図から全体としては直線で近似した線のこう配は 1.1 であった。これによって H28 と H31 では年間の観光入込客数にあまり変化はないと判断した。また相関係数 R≒1.0 であることから，各都市間の相関がよいことが明らかになった。

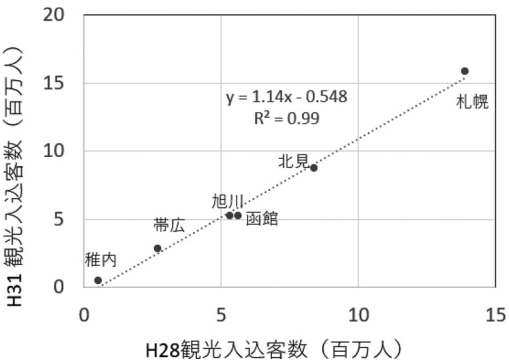

図 2　年間観光入込客数の比較（H28 と H31）

以上の結果を活用することで，これらの 6

都市以外の都市であっても，都市の観光規模を位置づけることが可能と思われる。

3.4　月ごとの観光入込客数の比較

図 3 は北見市における月ごとの観光入込客数の推移を示している。ピークは H28 と H31 とも 8 月で，4 月から 12 月まで非対称の山の形となっている。また，年間の合計は H31 年が約 40 万人多いものの，H28 と H31 の各月の傾向はほぼ同じと考えられる。**図 4** は月ごとの観光入込客数を H28 と H31 の散布図にまとめたものである。この結果から直線近似（$y=ax+b$）において $a\cong 1$ でかつ $b\cong 0$ かつ $R\cong 1$ として H28 と H31 の月ごとの分布はほぼ同じ傾向と見なせることが分かる。

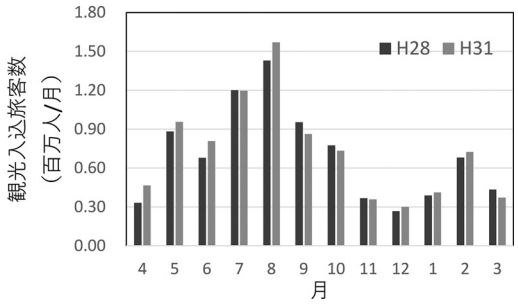

図 3　月別観光入込客数（北見）

表 1 は**図 4** の散布図を 6 都市のそれぞれ作成し，直線近似で得た勾配 a，切片 b，相関係数 R をまとめたものである。その結果，函館を除き全体として $b\cong 0$ かつ $R\cong 1$ と見なせると考える。$a=1$ の北見，帯広では H28 と H31 で変化はないが，$a<1$ である札幌，旭川，函館では観光入込客数の月ごとの変化が H28 よりも H31 で小さくなっていることが推定できる。

このように従来の方法では，確率のような定量的な評価はできないが，定性的な状況を示すことができる。

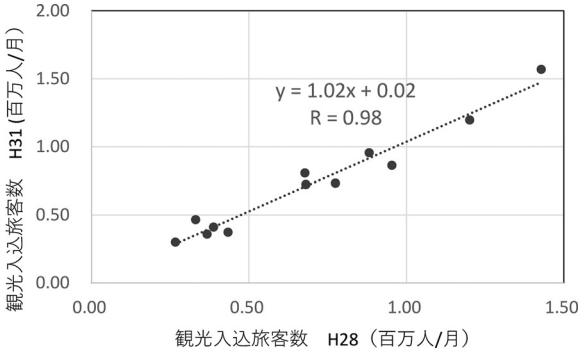

図 4　月別観光入込客数の相関（北見）

表 1　近似直線の勾配 a，切片 b，相関係数 R

	a	b	R
北見	1.00	0.00	0.98
帯広	1.00	0.00	1.00
札幌	0.76	0.20	0.92
稚内	1.14	0.00	0.95
旭川	0.76	0.10	0.91
函館	0.76	0.20	0.86

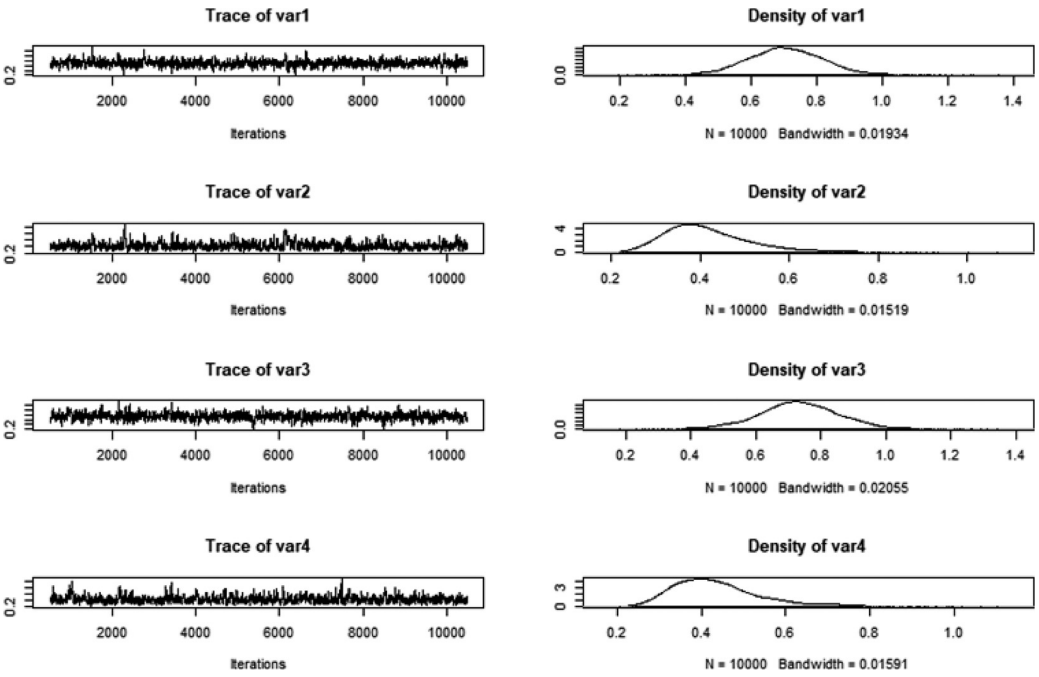

図 5　乱数の安定性の検証

3.5　ベイズ統計による検定

北見の H28 と H31 の観光引込旅客数を例にベイズ統計による検定を示す。**図 5** は北見において補足 E と補足 F に示した MCMC 法のプログラムで，正規乱数を用いて繰り返し回数を 10000 回の行ったときの H28 と H31 の平均と密度関数をまとめたものである。これから MCMC 法における乱数の安定性が確認できる。Trace は繰り返し回数ごとの経緯を示し Density は 10000 回の分布を示す。ここで var1 と var2 は H28 の平均と標準偏差，var3 と var4 は H31 の平均と標準偏差を示す。これらを概観すると，Trace では 4 項目とも波形の傾向に特段の変化がなく，繰り返し計算が正常に行われていることが分かる。また，Density では 4 項目とも正規分布に近い形状をしており，正規乱数が正常に機能していることが分かる。

これより補足 F により求めた phc 値と H28 と H31 の月ごとの差が含まれる範囲（信用区間）との関係を**図 6** に示す。ここで phc 値が 0.9 とは，確率が 0.9 の精度を意味している。また横軸 x は phc 値に対応する観光入込客数の誤差を示している。よって確率 0.9 である phc 値では観光入込客数の誤差は 0.42（百万人/月）であり，phc 値 0.5 では差 0.26（百万人/月）であることがわかる。

同様の方法で残りの 5 都市についても，phc 値が 0.9 と 0.5 において月ごとの誤差を求めた。

これらを**図 7** にまとめた。この図は**図 2** の観光規模から同じような都市の序列になっている。よって信用区間（phc 値）を定めればその月ごとの差を求めることができる。

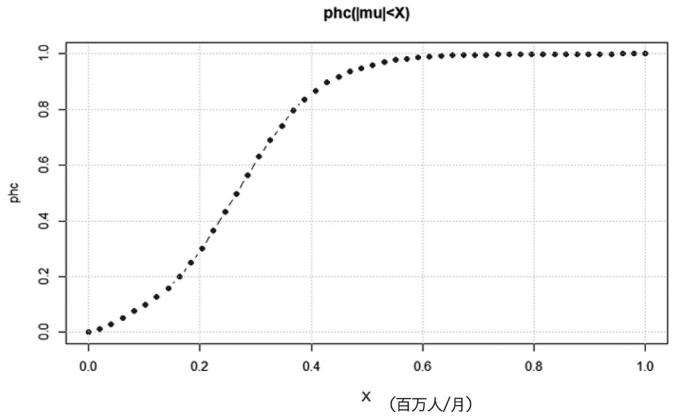

図 6　phc 値と H28 と H31 の信用区間

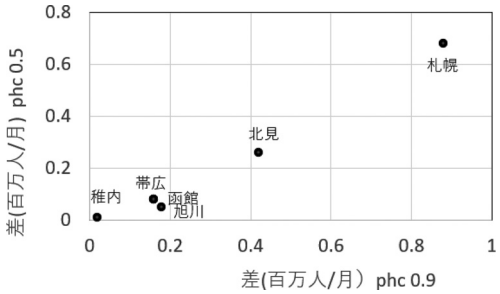

図 7　信用区間（phc 0.9 と 0.5）と差

4　月ごとの平均気温と観光引込旅客数の関係

北海道では一般的に気温が高くなると観光客が増える傾向とのイメージがある。気象庁のホームページから各都市の 30 年月別平均気温を入手した[10]。これらに基づき，北見における 30 年月別平均気温と H31 月別観光引込旅客数の関係を求めた。その結果を**図 8** に示す。この図から相関係数は 0.85 であり，相関関係が認められる。加えて他の 5 都市でも**図 8** と同様の方法で相関係数を求めた。また，そのなかで平均気温が 0 度超える 4 月から 10 月に相当する春・夏・秋の期間での相関係数も求めた（**表 2**）。

表 2 から 6 都市全体から，年間と 4 月～10 月とでは相関係数に大きな差がみられないが，気温が高くなると観光客が増える傾向が認められる。

5　ま　と　め

本論文では，持続可能な地域の形成にあたり観光は重要な要素にあたると考え，観光入込客数から都市の観光規模を推定する方法を検討した。その前提として，世界的な統計学の団体で

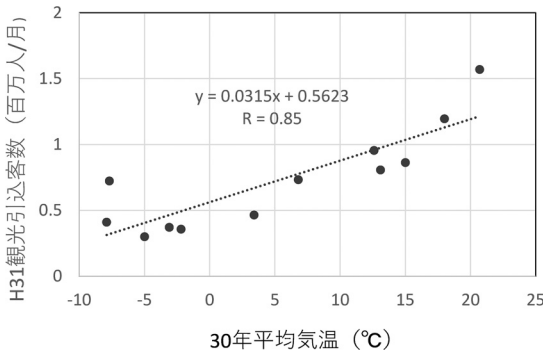

図 8　30 年月別平均気温と月別観光引込旅客数

表 2　30 年月別平均気温と月別観光引込旅客数の相関係数

	年間	4 月～10 月
北見	0.85	0.92
帯広	0.77	0.77
札幌	0.85	0.82
稚内	0.93	0.84
旭川	0.64	0.96
函館	0.84	0.40

あるアメリカ統計学会（ASA）から欠陥があると指摘を受けた p 値に代えて，ベイズ統計の信用区間を示す phc 値を用いて少ない情報から精度よく他の情報を推定する方法を確立した。これにより北見，帯広，札幌，稚内，旭川，函館の 6 都市の月別における観光入込客数から観光規模を推定できる結果をえることができた。また，30 年月別平均気温と H31 月別観光引込旅客数の関係から気温が高くなると観光客が増える傾向が認められた。なお，補足 A から補足 F には，本論文の検討に必要な概念や数式の展開，プログラムなどを示した。

参考文献

1）豊田秀樹：瀕死の統計学を救え！，朝倉書店，p72-73，2020.3
2）朝野熙彦，土田尚弘，小野滋：ベイズ統計学，朝倉書店，p95，2017.2
3）国土交通省ホームページ：https://www.mlit.go.jp/kankocho/siryou/toukei/irikomi.html
4）北見市ホームページ：https://www.city.kitami.lg.jp/administration/town/detail.php?content=9018&preview=true
5）十勝総合振興局ホームページ：https://www.tokachi.pref.hokkaido.lg.jp/ss/srk/shoko/05kankou/irikomi.html
6）札幌市観光統計データ：https://www.city.sapporo.jp/keizai/kanko/statistics/statistics.html
7）稚内市　観光調査：https://www.city.wakkanai.hokkaido.jp/sangyo/kanko/tyosa/
8）旭川市　観光客入れ込み状況：https://www.city.asahikawa.hokkaido.jp/700/701/705/d055229.html
9）函館市　来函観光入込客数推計：https://www.city.hakodate.hokkaido.jp/docs/2015062500021/

10）気象庁ホームページ（過去の気象データ検索）：http://www.data.jma.go.jp/obd/stats/etrn/view/nml_sfc_ym.php?prec_no=14&block_no=47412

補足 A

A-1　従来の統計学の黎明期（ピアソン，ゴセット，フィッシャー）

統計学を利用するユーザー（統計学ユーザーとする）の立場で統計学の基礎を作ったピアソン，ゴセット，フィッシャーについて述べる[1)]。18 世紀のころ，例えば重力加速度 g の値を測定する場合，ニュートン力学の（$F=\frac{m_1}{r^2}\times m_2=g\times m_2$）より二つの物質の質量（$m_1$，$m_2$）とその質点距離（$r$）にこれらの物質が引き合う力（$F$）を測定することで正しい重力加速度 g が得られるとして，理論に近い実験結果が得られるように実験器具の改良が行われていた。しかし，いくら測定機器や測定条件を整えても理論値に近づくものの一定の値とはならなかった。これに対し 19 世紀当初のころに統計の革命を起こしたと言われるピアソンは，大標本に分類（標本数は 100 以上）される数値データでは，一定の条件下であっても測定値はばらつくことを前提とした。これに基づき測定結果を平均値，標準偏差，対象度，尖度で評価すべきとし，これを母数と称した。このような観測値を母数で評価する方法をピアソン流とする。また，19 世紀の統計学の進歩に寄与したものに「中心極限定理」がある。これは大標本の平均の分布は正規分布（ガウス分布）に近づくとするものである。一般的に最小値や最大値が生じる確率が非常に小さな事象おいて，大標本ではほとんどが正規分布に近い分布を示す。ここで正規分布と見なすと，正規分布では母数の対象度と尖度は 0 となり，母数は平均値 μ と標準偏差 σ だけとなる。

次に少ない数の標本を小標本として，その統計処理についてゴセットは勤めていた会社から論文発表を禁じられていたのでスチューデントの名で研究成果を収めた。それは標本数に対応して分布形状を数式で示したものでｔ分布と呼ばれる。図 A1 はこの分布の確率密度関数である。標本数から求める自由度 k に対するｔ分布の確率密度関数 $f(x)$ と横軸との間の面積は x での確率を示す。よって，この関数が左右対称であることからある値 α より大きい確率は，$\alpha/2$ 以上の面積（図の黒塗りの部分）となる。このｔ分布は標本数が大標本に近づくに従い正規分布に収束していく理論構成であった。これをゴセット流とする。

次に重要な概念として，フィッシャーは理論値（真値）が存在する前提で，これと観測値（統計量）とは区別これに対する観測値は生ずる誤差を少なくためには，偏り（バイアス）を無くすことと無作為であることの重要性に着目した。またフィッシャーは有意性検定を生み出した。ここで十分な数の観測データの集合を母集団とし，その一部の観測データの集合を標本集団とすると，標本集団における偏り（バイアス）がなく無作為で得られた平均値 x は母集団の平均値 μ と同じである。また，母集団の標準偏差 σ とし標本数 n の標本集団の標準偏差 SE

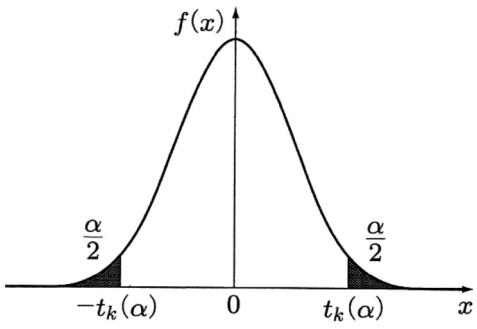

図 A1　t 分布の確率密度関数[2)]

では，SE $\sigma/\sqrt{n}$ の関係にあり，これらを不偏推定量（平均値 $\mu(=x)$，標準偏差 $\sigma(SE\times\sqrt{n})$）とした。標本でのデータ X とし平均値 $\mu(=x)$ とすると確率変数 Z は $Z=\dfrac{X-\mu}{\sigma}$ として定義される。この X から Z への変換は標準化と呼ばれる。

これにより**図 A1** が標準正規分布と仮定すると，$-\infty\leq x\leq\infty$ で $f(x)$ と x 軸の間の全体の面積が 1 となり，確率を示す。また，$-t_k(\alpha)\leq x\leq t_k(\alpha)$ は信頼区間と呼ばれている。

以上により，ある現象の発生を正規分布などの理論曲線に仮定した場合に，**図 1** の α を p 値とした。なお，フィシャーは p 値を確定していないが，例として p 値が 0.1 を示している。これをフィッシャー流とする。

A-2　仮 説 検 定

例えば，最近の話題でワクチンの効果において接種した人としない人の比較やある駅で月単位の乗降者数における年ごとの比較などでは，個別にフィッシャー流の理論値（真値）を見出すことは現実的ではない。このような二種類の比較を 2 者問題とする。19 世紀前半において 2 者問題を統計に取り組み動きが出てきた。ネイマン・ピアソンの共同研究（1928 年）では，検定されるべき仮説（帰無仮説）とそれ以外の仮説（対立仮説）の 2 つの仮説を設けて p 値を用いて検定する理論を生み出した。例えば，指標を用いて新薬のワクチンを接種した人（A）としないひと（B）に差がないとする帰無仮説と差があるとする対立仮説を設定する。ここで指標の母集団（不偏推定値）における平均値，標準偏差，標本数をそれぞれ A(μ_A，σ_A，m)と B(μ_B，σ_B，n)とすると帰無仮説 H_0：$\mu_A=\mu_B$ であり，対立仮説 H_1：$\mu_A\neq\mu_B$ となる。

次に標本集団の平均を x_A，x_B とし母集団の標準偏差は既知として σ_A，σ_B とする。ネイマン・ピアソンの理論では標本集団の標準偏差 SE を標本から求めず，1 つのケースを扱うフィッシャー流を拡大して 2 つのケースでは $SE\sqrt{\sigma_A^2/m+\sigma_B^2/n}$ としている。

ここで帰無仮説 $\mu_A=\mu_B$ より $\mu_A-\mu_B=0$ の仮説となる。標本集団の平均値の差 $x_A-x_B(=\mu_A-\mu_B)$のばらつきの程度を調べる。このばらつきに関係する確率変数 $Z=\dfrac{x_A-x_B}{SE}$ を求める。こ

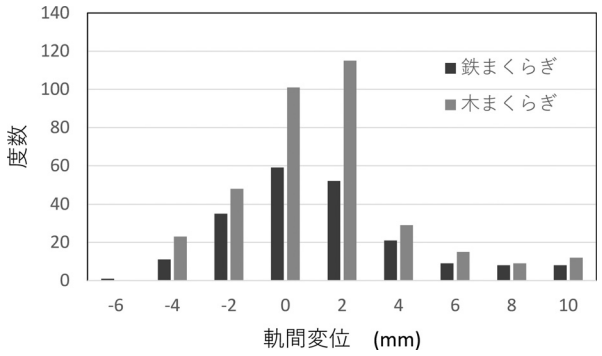

図 A2　木まくらぎと鉄まくらぎの比較（軌間変位）

こで確率変数 Z は図 A1 の t_k（α）に相当すると考える。統計に関する本に出ている標準正規分布表から Z に対応する p 値（図 A1 の黒い部分の面積）を求める。この p 値が 0.1 より小さい場合とは，このばらつきが大きいと判断し $\mu_A - \mu_B \neq 0$ として帰無仮説（H_0）は棄却される。これをネイマン・ピアソン流とする。

A-3　統計学ユーザー

統計学を専門領域としないで統計を活用する者を統計学ユーザーとする。統計学ユーザーが論文を作成する場合に，実験室においては理論解を導くことができれば，母数（平均値と標準偏差）を求めるのに大標本を用いてピアソン流で正規分布を求めることもあれば，小標本ではゴセット流にて t 分布を求めることがある。

さらに，統計学を理論の証明ではなく，もっと広い分野で活用する場合もある。その例を次に示す。鉄道工学を研究分野としている著者は腐朽する木まくらぎに代わって再生利用のできる鉄まくらぎを敷設する研究をしている。ここで左右レールの間隔の狂いである軌間変位が所定の値よりも大きい場合を（＋），小さい場合を（－）とする。図 A2 はある駅構内の鉄まくらぎと木まくらぎが混在している区間を対象に，フィッシャー流の普遍性を持たせることと偏り（バイアス）を無くす目的で，連続して 1 線区全部で鉄まくらぎ 204 本と木まくらぎ 352 本の軌間変位の測定結果である。両まくらぎにおける標本数が 100 を超えていることから，大標本による正規分布と仮定できる。その結果からそれぞれの不偏推定量を求めた。この測定結果から各まくらぎの母数を求めたが，これに対するフィッシャー流の理論値（真値）を見出すことは現実的ではない。

それは，それぞれのまくらぎを通過する車輪荷重が変動することや各まくらぎを支えるバラストの状態が異なることなど多くの要素が関係していることによる。ここで必要なのは木まくらぎと鉄まくらぎの違いの検討である。このような二種類の比較を 2 者問題として扱う。なお，ネイマン・ピアソン流では鉄まくらぎと木まくらぎで軌間変位に差がある場合に，帰無仮説 H_0 をこの差がないことで棄却することになる。一方，一般的に鉄まくらぎと木まくらぎが

混在している区間でまくらぎの違いによる軌間変位が見受けられないことから，フィッシャー流の有意性検定を採用してその範囲を信頼区間として求めることとする。

前述のように標本数が 100 を超えているので，両まくらぎとも正規分布と仮定する。この分布の平均値は両まくらぎとも 0.2 mm であり，標準偏差は鉄まくらぎが 3.2 mm，木まくらぎが 3.0 mm である。鉄まくらぎ（XG）と木まくらぎ（YG）とし，の差（XG－YG）を用いて検討した。その結果，不偏推定値の差（XG－YG）の標準偏差 SE は 0.156 であった[3]。

図 A1 は小標本における t 分布の確率密度関数であるが，t 分布の大標本では正規分布のなることから信頼区間の考え方は，$t_\infty(\alpha)$として正規分布にも用いることができる。ここで $\alpha=0.05$ を採用した（p 値が 0.05 に相当）。よって信頼区間は p 値が 0.01 よりも拡大するためのリスクが生じることになる。標準正規分布表から $t_\infty(0.05)$より，不偏推定値の差（XG－YG）が入る範囲は－0.4 mm～0.3 mm が求められた。ここで軌間変位は 1 mm 以下を無視していることを考慮すると，信頼区間を拡大するリスクがあっても，この範囲は 1 mm 以内である。以上から鉄まくらぎと木まくらぎの軌間変位の差は殆どないものと考えられる。このように p 値は統計学ユーザーにとって重要な判断手法となっている。

補足 B　ベイズの定理

一般的に A が原因で結果の B が生じるとすると，高校で習う「ベイズの定理」では結果 B から原因 A を推定することとしている。これを統計として扱うと**図 B1-1** に示す二重線の B の範囲で A が起こる確率 P と考えることができる。これを P(A|B)とすると，P(AB)を A と B が同時に起こる確率として

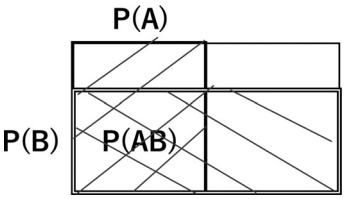

図 B1-1　A と B の確率

$$P(A\,|\,B)=\frac{P(AB)}{P(B)} \qquad (B1)$$

がなりたつ。

次に**図 B1-2** で A_1～A_i～A_n に分解できる A をデータと見なし B をパラメータとして B が起こる確率は

$P(B\,|\,A_i)=\dfrac{P(BA_i)}{P(A_i)}$ である。

よって，P(BA_i) = P(A_iB)であるので

$$P(A_iB)=P(B\,|\,A_i)P(A_i) \qquad (B2)$$

ここで $\sum_1^n P(BA_i)=P(B)$ であることから

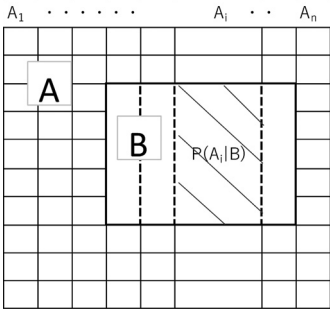

図 B1-2　データとパラメータ

$$P(B)=\sum_{1}^{n}P(A_iB)=\sum_{1}^{n}P(B|A_i)P(A_i) \qquad (B3)$$

となる。

以上の（B1）を

$$P(A_i|B)=\frac{P(A_iB)}{P(B)} \qquad (B4)$$

から（A4）へ（A2）と（A3）を代入すると（B5）が導ける。

$$P(A_i|B)=\frac{P(B|A_i)P(A_i)}{\sum_{1}^{n}P(B|A_i)P(A_i)} \qquad (B5)$$

補足 C

式（B5）の分布は離散関数として扱っているが，A の分割幅を細かくすることで連続関数と見なすことが可能である。

$$P(\theta|D)=\frac{P(D|\theta)P(\theta)}{\int P(D|\theta)P(\theta)d\theta} \qquad (C1)$$

となる。

ここで P(θD)が確率分布であるためには式（C2）の条件が必要となる。

$$\int p(\theta|D)d\theta=\int\frac{P(D|\theta)P(\theta)}{P(D)}d\theta=1 \qquad (C2)$$

式 B2 の 2 番目と 3 番目の両辺に定数である P(D)をかけると次式となる。

$$P(D)=\int P(D|\theta)P(\theta)d\theta \qquad (C3)$$

以上から式（C1）の分母は定数 P（D）となる。
よって，式（B1）の分子を比例式として次式に示すことができる。

$$P(\theta|D)\propto P(D|\theta)P(\theta) \qquad (C4)$$

補足 D　計算ソフト R と RStudio

検索サイト「R CRAN」から「Download R for windows」→「base」→「Download Rxxx for windows」（xxx はバージョン）により無料でダウンロードが可能である。R だけでは使いづらいので無料の RStudio もダウンロードをお勧めする（http//www.rstudio.com）。ともに文献 4 で詳しく説明されている。

補足Ｅ　パラメータ推定における R を用いた MCMC の解析プログラムの例(1)

ここでは北見市を対象に確率 90％で H28 年度と H31 年度の月ごとの入れ込み観光客数の差の範囲を求める。なおこの解析プログラムは[5] https://sturgeon.hatenablog.com/entry/phc を参考にしている。

プログラムの概略を説明すると次になる。

①ではマルコフ連鎖モンテカルロ法のパッケージを呼び出している。

②は乱数をセットしている。

③では H28 年度の月ごとの入れ込み観光客数を百万人単位で 4 月から翌年の 3 月まで示している。

④では H31 年度の月ごとの入れ込み観光客数を同様に示している。

本文では事後確率 P(θ|D)，尤度関数 P(D|θ)，事前確率 P(θ)としているが，本プログラムでは D=X としている。

⑤では，尤度関数 P(X|θ)における θ を母数として H28 年度と H31 年度のそれぞれで平均と標準偏差を定義している。

表 E1　MCMC の解析プログラムの例(1)

```
#対象：北見市　月ごとの入れ込み観光客数の差（H28とH31）

#マルコフ連鎖モンテルロ法のパッケージ
①  library(MCMCpack)
#乱数のセット
②  set.seed(1)
#H28のデータ(百万人/月)
③  treat <- c(0.33,0.88,0.68,1.20,1.43,0.95,0.78,0.37,0.27,0.39,0.68,0.43)
#H31のデータ(百万人/月)
④  control <- c( 0.47,0.96,0.81,1.20,1.57,0.86,0.73,0.36,0.30,0.41,0.72,0.37)

⑤  #尤度関数 theta：推定する4つの母数（H28の平均θ1・標準偏差θ2
                    およびH31の平均θ 3 ・標準偏差θ4）
#分布には正規分布を仮定
llnormfun <- function(theta, x) {
 a <- x[1:12]
 b <- x[13:24]
 if(theta[2] < 0.0 || treat[4] < 0){
  l <- -Inf
 }
⑥ else{
  l <- sum(log(dnorm(a, mean = theta[1], sd=theta[2])))+
    sum(log(dnorm(b, mean = theta[3], sd=theta[4])))
 }
 return(l)
}

#事後分布f(theta|x)
post.diff <- MCMCmetrop1R(llnormfun, theta.init=c(1,0.5,1,0.5),
                  x=c(treat,control),
⑦                 thin=1, mcmc=10000, burnin=500,
                  tune=1.0,
                  verbose=500, logfun=TRUE)
⑧  #事後分布のプロット
plot(post.diff)
```

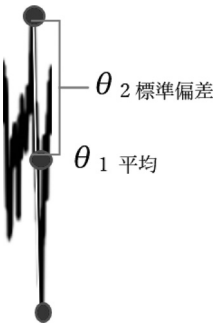

図 E1　θ1 とθ2 の組み合わせ

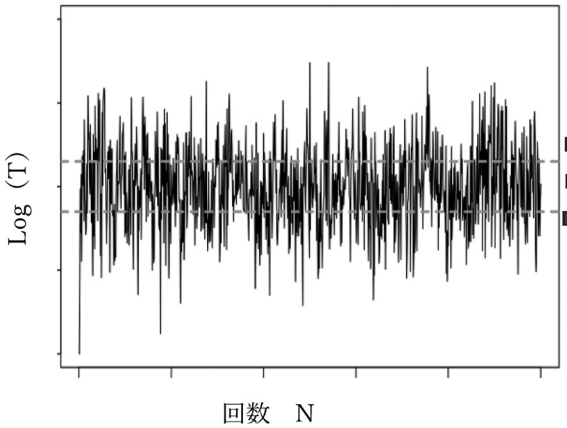

図 E2　回数と Log(T) の関係

図 E3　確率密度

⑥では，MCMC 推定に使用する関数 function(theta, x)としている。ここで x に入る値は a と b とし，a は H28 のデータの 1 ～12 までで，b は H31 のデータの 23～24 とする。また theta[1]は H28 の平均，theta[2]は H28 の標準偏差であり，theta[3]は H31 の平均，theta[4]が H31 の標準偏差とする。以上の 4 つの theta をそれぞれ計算する。ここで標準偏差は非負なので，theta[2]と theta[4]負になったときは対数尤度をマイナス無限大（-inf）とする。

次に Sum 以下のプログラムを Sum(log(T))，T=dnorm()で表すと，dnorm()は正規分布の確率密度を算出する関数で dnorm(x, mean, sd)は，x における正規分布の確率密度を求める命令文である。例えば H28 年度 4 月のデータ（入れ込み観光客数）0.23 百万人（D または X に相当）に対し，ランダム正規分布を用いて $\theta 1$ と $\theta 2$ の組み合わせを計算する（図 E1）。これを 10000 回繰り返すと概念図（図 E2）が得られる。これらの 10000 回の平均などから θ_1 と θ_2 を定める。ここで MCMC のプログラムでは θ_1 と θ_2 の最大範囲（θ_{1max} と θ_{2max}）を入力するが，これにより正規化（$(D-\theta_{1max})/\theta_{2max}$）して確率分布へ変換する。これを尤度関数 P(D|θ)に対して尤も当てはまりのよいものとする。

H31 年度のデータも H28 年度のデータのときと同じような手順で，ランダム正規分布を用いて θ_3 と θ_4 の組み合わせを計算し，尤度関数 P(D|θ)を求める。なお，図 E3 は縦軸から見た頻度を示す。

⑦MCMC を実行し $\theta 1$，$\theta 2$，$\theta 3$，$\theta 4$ を求め，事後分布 P(θ|D)を確定する。

ここで初期値（init）を平均 1，標準偏差 0.5 としている。X は③と④を入力する。Thin は区切りのステップを 1 とし，mcmc とは繰り返し 10000 回を示している。birmin は初期の不安定さを回避するために採用しない数を 500 としている。tune は乱数発生法のメトロポリスのサンプリング間隔で 1.0 としている。verbose は計算された結果を画面に表示する間隔を示している。

⑧事後分布 P（θ|D）に必要なパラメータ（θ_1，θ_2，θ_3，θ_4）を出力し，パラメータの推定が完了する。

補足 F　パラメータ推定における R を用いた MCMC の解析プログラムの例(2)

表 F1 の⑨は補足 D でそれぞれ 10000 回繰り返し計算された各パラメータ（θ_1，θ_2，θ_3，θ_4）のうち平均（θ_1 と θ_3）の差を計算している。**図 E3** に示す方法でこの結果を発生頻度に整理して phc 値とする。

⑩は以上の平均の差（X）と phc 値の関係としてグラフ化する命令文である。

表 F1　MCMC の解析プログラムの例(2)

```
⑨ #MCMCで生成したデータを利用
  #H28とH31の平均値の差
  diff_of_mean <- post.diff[,3]-post.diff[,1]

  #phc curve
  phc <- function(v, hypo="larger", xlim=c(0,1), title=NULL){
   n <- 50
   X <- numeric(n)
   phc <- numeric(n)
   a <- seq(xlim[1],xlim[2],length=n)
   for (i in (1:n)) {
    X[i] <- a[i]
    if (hypo == "larger") {
     phc[i] <- length(v[v>a[i]])/length(v)
    }
⑩   else if (hypo == "smaller") {
     phc[i] <- length(v[abs(v)<a[i]])/length(v)
    }
    else if (hypo == "non-overlap") {
     phc[i] <- length(v[abs(v)>a[i]])/length(v)
    }
   }
   plot(X, phc, type = "b", col="blue", pch=16)
   title(title)
   grid()
  }
```

補足の参考文献

1）ディヴィット・ザルツブルグ（竹内恵行・熊谷悦生訳）：統計学を拓いた異才たち，日本経済新聞，2006.3
2）高田佳和：例題で学ぶ統計入門，森北出版，p.39，2013.1
3）上浦正樹 他：鉄まくらぎにおける周辺のバラスト移動の評価と木まくらぎとの比較，令和3年度土木学会第76回年次学術講演会，2021.9
4）朝野熙彦，土田尚弘，小野滋：ベイズ統計学，朝倉書店，pp.194–196，2017.2
5）Hatena 実験計画，phc 解析：https://sturgeon.hatenablog.com/entry/phc