統計同人誌をつくろう!

調べて、分析して、書きたい人のために

後藤和智事務所OffLine

統計同人誌をつくろう！

——調べて、分析して、書きたい人のために

著：後藤和智（後藤和智事務所 OffLine）
表紙イラスト：椹木（白地図と青写真）

まえがき

36 冊目の同人誌となります、後藤和智です。そして今回は、初めての同人誌製作指南書ということになります。同人誌の製作テクニックを紹介する同人誌は古くからあり、最近でもサークル「Circle's square」のフルカラーの同人誌だったり、あるいは LaTeX による同人誌の製作に特化した解説書を書いているサークル「PARRAREL ACT」のものだったりと、いろいろとあります。その中で私ができるものと言ったら、もう統計学しかないですね（苦笑）。というわけで統計学に基づく同人誌の製作ガイドというものを作ってみようと思いました。

私はコミケ 81 のあたりから、ROM 版のカタログを検索して、数学や統計学の同人誌を作っていると思われるサークルの紹介をツイッターで行っており、また統計同人誌についてもいろいろ集めてきました。統計を扱っているサークルには、弊サークルが統計学で活動する前から、成人向け漫画などの統計学的分析を行っているサークル「でいひま」や、東方の種々のデータを分析しているサークル「久幸繙文」、コミケのジャンルに関する統計を扱っている「Paradoxical Library」などいろいろありますが、自らの興味のある分野の大まかな傾向を示す際に、統計学の知見を活かすようになっているという傾向が見られます。

他方で、商業評論の世界においては、未だにデータによる分析よりも、特定の作品から若い世代の「心性」を決めつけるような「批評」が幅をきかせています。2008 年に『おまえが若者を語るな！』（角川 One テーマ 21）を上梓したときよりも幾分かはマシになったとは言え、作品の読み込みさえ疑わしいような「文化論」が出されている現状があります（その点、アニメや漫画に対する批評も同人誌のほうが充実しています）。

他方で「ビッグデータ」礼賛などのように、人々の行動をとにかくデータとして集めれば理想の社会ができるんだ！という意見も出ています。しかしこれも、社会評論における数学や統計学的なものへの忌避と合わせ鏡のものではないでしょうか。統計学というものは必然的に適用範囲が限定されるものであり、また複雑な分析であればあるほど、いろいろと恣意性も入ってくるものなのです。統計学は、その適用範囲や限界を知りながら使っていく必要があります。

本書では、同人誌製作という視点から、統計学に基づいた分析をいかに進めていくかということを述べていきたいと思います。第 1 章は、先ほど挙げたサークル「でいひま」の牧田翠氏との対談となっています。統計学をコンテンツとして提供する 2 つのサークルの語りを参考に、統計学を使った分析や同人誌の作成ってなんだろう、ということを学んでいただければ幸いです。

第 2 章では統計解析の手法を紹介しますが、前述の理由より、統計学に関する数理的、テクニカルな部分や数式よりも、統計分析の概念に重きを置いた記述になっています。本格的に数式を使って学びたいという方のためにブックガイドも作成しているので、本書で統計分析の概要をつかんだら、是非とも統計学を学んでみてください。

また本書は同人誌製作ガイドのため、統計学に基づいた同人誌作成に役立つソフトも挙げていきます。第 3 章では理系の論文やレポートで重用されているフリーの組版ソフト「LaTeX」について説明します。LaTeX は数式を綺麗に組むことができるほか、表やグラフなどを入れることもできて、統計学に基づいた同人誌を作成するには最適のソフトと言うことができます。本書では統計学に基づいた同人誌を作成するための最低限の機能をいくつか紹介します。

第 4 章では、統計学を用いた同人誌の製作に大きく貢献してくれるであろう、フリーの統計ソフト「R」の解説を行います。R は近年は良質な解説書も多く出版されているだけでなく、専門家からの支持も広く集めており、中にはいま最先端の統計ソフトとして見る人もいるくらいです。そのような R に関して、本書ではデータの扱い方を中心に解説します。解析手法についての解説は今回はあまり行いません。

第 5 章では、統計同人誌の製作という視点から見た、Adobe Creative Cloud や Microsoft Office についての解説を行います。

第 6 章は、弊サークルが統計学サークルとして作成してきた主要な統計学系の同人誌の作り方を紹介します。弊サークルの統計学同人誌は、統計学の解説書から、多変量解析やテキストマイニングなど様々なものがありますが、具体的な実例を示すことにより、読者の皆様が統計学を使った同人誌を作るためのヒントを得てくれたら幸いです。

なお本書は、2013 年の夏コミ（コミケ 84）で出した『R Maniax――フリーの統計ソフト「R」を使いこなす本』と同様、それぞれのトピックスを見開き 2 ～ 6 ページでまとめ、使いやすさを重視したものにしております（第 1 章除く）。

本書の表紙絵は、サークル「白地図と青写真」の椹木氏に担当していただきました。氏がコミケ 84 で製作されたオリジナル創作漫画同人誌『クローズドサークル』は、仲の良かったサークルがある事件をきっかけに崩壊していく様や救いのない結末が推理タッチで描かれていておすすめです。そして表紙絵を担当してほしいと依頼したところ、快諾してくださいました。こちらが「理系」ということをテーマとして提示し、氏からはいくつかラフをいただいたのですが、統計同人の入門書ということから今回の表紙を採用いたしました。あと、弊サークルの同人誌の表紙イラストに男性が出るのは初めてですね。

本書は第一義的には同人誌製作指南書ではありますが、サブタイトルにもある通り、「調べる」「分析する」そして「書く（発表する）」ことの大切さを追い求める本でもあります。統計学の魅力と問題点は自分で分析してこそわかるものですし、また思い込みや借り物の論理で、何らかの作品を若年層の心性を代表するものとして扱ったところで、新たな知見が生まれるのは難しいでしょう。だからこそ、今「調べる」「分析する」ということを改めて提示し、そこから「書く」というものが生まれれば、と思います。

統計同人誌をつくろう！ - 目次

第１章 統計同人はこんなに面白い
（ゲスト：牧田翠）

統計同人誌の製作ガイドブックを作成するにあたり、弊サークルの統計同人誌にも大きな影響を与えている同人誌「エロマンガ統計」シリーズの著者であり、近年はニコニコ超会議や講演などでも活躍している、サークル「でいひま」の牧田翠氏をお迎えして、統計同人誌の魅力や製作上での考えなどを語り合いました。
（収録：2013 年 8 月 11 日 ジョナサン五反田駅前店にて）

なぜ統計同人活動を始めたか

後藤：本日は、夏コミお疲れ様というのも兼ねまして、『統計同人誌をつくろう！』の対談企画と参ります。今回は、エロマンガ統計シリーズで有名なサークル「でいひま」の牧田翠様をお迎えして、ビッグデータ云々とか言われる中で統計同人について話していこうと思います。まず牧田さんは、「ニコニコ超会議 2」（注 4 月 27 ～ 28 日、幕張メッセ）のニコニコ学会（注 4 月 28 日）におきまして、なんとポスター発表で大賞という輝かしい成績を収められました。そのときのポスターには「ここに病院を建てよう」という紙が貼り付けられたという（笑）。まずは「エロマンガ統計」ということで、なぜそのような研究を始められたのかを伺いたいと思います。私が初めて「エロマンガ統計」シリーズを読んだのは 2009 年の夏コミなんですね。そのときは私も統計学の解説書を作り始めていたんですけど、その夏コミで牧田さんのスペースを偶然見かけて、こんなことを統計でやっているのかと思い、既刊全部くださいと言っていました。

牧田：正直に言うと覚えてない（笑）。

後藤：そうですよね（笑）。それはさておき。まずは牧田さんが漫画の統計分析のきっかけを始めたきっかけというものを説明していただきたいと思います。

牧田：きっかけとしてはいくつかあるんですけど、私の大学院の時の先生が、同じ時にやおいの研究をやっていて、それに惹かれたというのがあります。もうひとつは、マンガの研究というものが、例えば手塚の影響とかアシスタント筋の影響とかという、所謂歴史的な分析になってしまうと言う傾向があって、データとして何かを出していくというのがあまりないんです。数値として何かを出していくというのがなかなかできていないというものがあるんですね。

後藤：まあ漫画研究みたいなものって結構まあ文学研究的なものが主流であるようなものなんですけど、ただ、例えば文系でも心理学などは統計たくさん使ったりと、文系であっても統計は決して無縁な分野ではないわけです。それで、分析を進めていて、実際に得られたものとか面白かったりしたこととかありますか？

牧田：結構やっていると面白いことっていっぱい出

牧田氏が第４回のポスター発表で大賞を受賞した「ニコニコ学会β」は、統計学に基づいた研究発表も少なくない。
上：http://www.chokaigi.jp/2013/event/nicogakkai.html（ニコニコ超会議２公式サイト）
下：http://niconicogakkai.jp/nng4/nng4-poster（「ニコニコ学会β」サイトのアーカイブ）

てくると思いますが、何が一番面白いかと急に言われるとなかなか迷うところがありますが（笑）

後藤：まあ確かに分析する対象とか視点とか違うと何が面白いのかとか変わってくると思いますが、その中から敢えて共通点を探すとすればどういうものになるでしょうか

牧田：一つ自分のこれはあったことというか元々考

第2章 数式をあまり使わない統計解析の「概念」入門

2.1 単純集計とカイ二乗検定

本章では、統計同人誌を作成するにあたって、抑えておきたい統計学・統計解析のポイントについて述べていきたいと思います。ポイントについて述べるので、本章においては基本的に数式はあまり使いませんが、最初に注意しておくこととして、統計学を本気で学びたいのであれば数式は欠かせないものであるということを明言しておきたいと思います。統計解析を支える数理的な背景、すなわち数式を理解しないで統計的な分析を行うと、逆に統計的な手法に「使われる」ことになってしまうので、気をつけましょう。

また、統計解析に用いるデータを作成するためには、それなりに形が整っている必要があり、またあらかじめ何を分析したいのかということについても明らかにしておく必要があります。学術的な研究ではデザインとか設計とか言われていますが、統計学を用いた同人誌を作成するにあたって、分析の目的をある程度定めておくのはとても大事です。もしそれを行わないままデータの集計を行うと、自分は本当は何をしたいのかということを何度も考えてしまい、そこで製作が止まってしまうからです。

さてここから統計解析に入っていくわけですが、実は統計的な分析を行う上では、ただの単純集計・クロス集計も莫迦にすることはできません。量的なデータにしろ質的なデータにしろ、データがどのように分布しているかをあらかじめ掴んでおくことは、今後の統計分析を行う上でも極めて有利に働きます。

データの集計法としては、質的なデータであればそのまま集計すればいいですが、量的なデータの場合は、集計の際に一定の範囲を決めてその範囲に入るようなデータを作成する必要があります。そしてそれぞれの範囲に入るデータがどれくらいになるかを示したグラフをヒストグラムと言います。量的なデータであってもヒストグラムを使うと、データの分布をビジュアルで示すことができます。

量的なデータを取ってきたら、平均値を取ることが多いかもしれません。平均値のように、持っているデータの特徴を示す統計量のことを要約統計量と言います。ただ、確かに平均値は要約統計量の中でも重要なものですが、平均値だけではデータの分布を見るには情報量が少なすぎます。データの分布を見るために必要な要約統計量としては、分散や標準偏差、最大値･最小値、最頻値、中央値などがあり、これらを総合して見る必要があるでしょう。

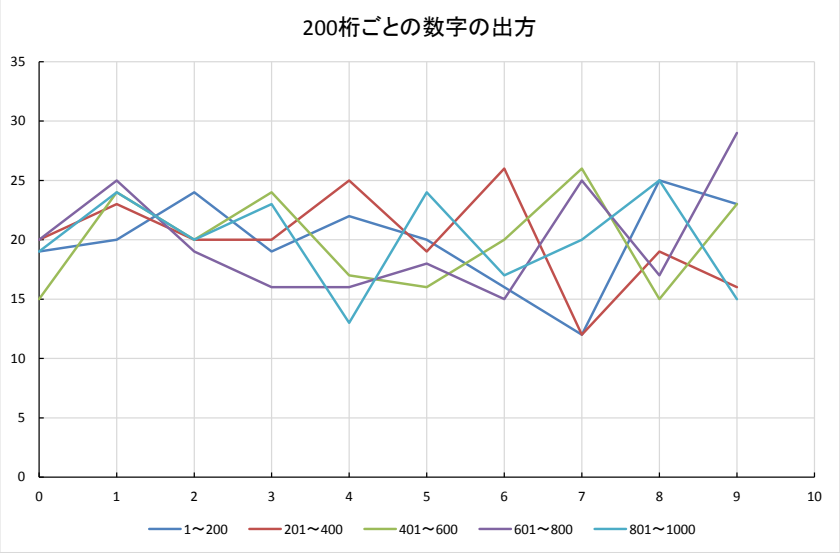
さて､分布を見る際には､母集団全体に対して行った悉皆調査（全数調査）ならともかく、母集団からいくつかのサンプルを抽出して行った標本調査の場合は、母集団において想定した分布になっているかを統計的に検討する必要があります。これを統計的（仮説）検定と言いますが、質的データの分布に関する統計的仮説検定では、持ってきた集計結果が予想していた分布であるかどうかの可能性を確率的な指標で計算します。使用するのはカイ二乗検定とい

例：円周率の小数第1～1000位における数字の出方
200桁ごとの出方に差はあるか？
円周率出典：http://www.geocities.jp/f9305710/PAI1000000.html / 集計：MeCab

数字	位[小数第X位] 1～200	201～400	401～600	601～800	801～1000	総数	200桁ごとの期待度数
0	19	20	15	20	19	93	18.6
1	20	23	24	25	24	116	23.2
2	24	20	20	19	20	103	20.6
3	19	20	24	16	23	102	20.4
4	22	25	17	16	13	93	18.6
5	20	19	16	18	24	97	19.4
6	16	26	20	15	17	94	18.8
7	12	12	26	25	20	95	19
8	25	19	15	17	25	101	20.2
9	23	16	23	29	15	106	21.2

Σ（実際の度数-期待度数）^2/期待度数					χ^2値
6.0367	9.0260	6.2576	7.8224	6.3406	35.4832
自由度=(10-1)*(5-1)=36				p値	0.5070 >0.1

数字の出方に差があるとは言い難い。



う手法です。

カイ二乗検定は、データを集めてきた群ごとに、データの分布が異なるかということにも使えます。群ごとの平均を想定分布として、それと違う分布を持つ群があるなら、群ごとの分布は違っていると見なします。

2.2 相関係数と回帰分析

2つの量的なパラメータの関係性を示す指標に相関係数があります。相関係数は、共分散と呼ばれる指標を計算したあと、パラメータに特有のばらつきの影響を取り除くためにそれぞれのパラメータの標準偏差で割って求めます。相関係数は1から-1までの実数になることが知られており（証明省略）、プラスになればなるほど、「片方のパラメータが増えれば（減れば）もう片方のパラメータも増える（減る）」という関係（正の相関）が強くなり、またマイナスになればなるほど「片方のパラメータが増えると（減ると）もう片方のパラメータは減る（増える）」という関係（負の相関）が強くなります。

回帰分析は、特定のパラメータに対して、それに対して影響を及ぼしていると思われるパラメータを用いたモデルを使い、具体的な影響の大きさを測る手法です。回帰分析においては、最終的な結果となるパラメータを従属変数（または被説明変数）、それに対して影響を及ぼしていると思われるデータを独立変数（または説明変数）と言います（本書では「被説明変数／説明変数」という言葉を使うこととします）。回帰分析は、そのモデルによって線形と非線形に、また説明変数に用いるパラメータの数によって単回帰分析と重回帰分析に分かれます。単回帰分析は説明変数の数が1つ、重回帰分析は2つ以上です。

回帰分析の中でも、線形単回帰分析は、右のページに示すとおり、そのモデルは極めてシンプルではありますが、重回帰分析や非線形の回帰分析、さらには因子分析や共分散構造分析（パス解析）などといった多変量解析（2.5参照）に至る、様々な分析の基礎となっています。また線形単回帰分析のモデルは、中学の数学で習う一次関数と似たような感じになっているので、説明する際にもわかりやすいと思います。この点で、線形単回帰分析は、説明変数と被説明変数の相関関係を具体的な数値で表したものと言うことができるでしょう。

回帰分析では、それぞれの説明変数が被説明変数に与える影響を数値として表すものですが、説明変数の影響の強さを示す指標には2つあります。第一に回帰係数で、これは影響の強さそのものを示します。第二に回帰係数の有意水準で、これは「どのくらいの確率で「回帰係数=0」になってしまうか」を示す指標です。一般にこれは0.05より低くなればその説明変数の影響は確かなものであると言うことができます。

またモデル全体の当てはまりの良さを示す指標もいくつかあります。例えば決定係数は、回帰モデルにおいて用いた説明変数によって被説明変数のどのくらいが説明されるかを示す指標です（計算方法は右のページ参照）。線形単回帰分析においては、決定係数は説明変数と被説明変数の相関係数の二乗に等しくなります（証明省略）。

例：東北新幹線の乗車人員に関する分析

乗車人員：http://www.jreast.co.jp/passenger/2012_shinkansen.html
（新幹線のみ。在来線を含まない）
営業距離：Wikipedia

駅名	2012年度 1日平均 乗車人員	（対数変換）	2013年 9月現在 上り 停車本数 （臨時除く）	東京 からの 営業距離 [km]
小山	4,996	8.5164	26	80.3
宇都宮	12,245	9.4129	55	109.0
那須塩原	3,212	8.0746	24	152.4
新白河	2,014	7.6079	19	178.4
郡山	8,597	9.0592	48	213.9
福島	7,301	8.8958	43	255.1
白石蔵王	893	6.7946	20	286.2
仙台	24,319	10.0990	67	325.4
古川	2,688	7.8966	19	363.8
くりこま高原	1,090	6.9939	18	385.7
一ノ関	2,451	7.8043	19	406.3
水沢江刺	1,028	6.9354	18	431.3
北上	1,284	7.1577	19	448.6
新花巻	993	6.9007	18	463.1
盛岡	7,235	8.8867	35	496.5
いわて沼宮内	88	4.4773	8	527.6
二戸	764	6.6386	12	562.2
八戸	3,248	8.0858	17	593.1
七戸十和田	638	6.4583	12	629.2
新青森	4,571	8.4275	19	674.9

予測された 乗車人員 （対数）	exp(予測値)
7.8164	2480.92
9.7570	17275.38
7.6709	2145.09
7.3315	1527.66
9.2711	10626.22
8.9293	7549.63
7.3817	1606.28
10.5280	37347.38
7.3024	1483.91
7.2319	1382.89
7.2958	1474.07
7.2248	1373.05
7.2892	1464.33
7.2198	1366.22
8.3548	4250.67
6.5390	691.56
6.8018	899.49
7.1324	1251.83
6.7913	890.10
7.2537	1413.31

乗車人員（対数）との相関係数　　0.8359　　-0.4434
モデル：log（乗車人員）=6.0850368-0.0001567*（営業距離）+0.0670742*（上り停車本数）
グラフはRコマンダーで作成