仮説検定と区間推定の関連を重視した統計の授業について

On a statistics class emphasizing relations between hypothesis tests and interval estimations

中　村　宗　敬[1]
NAKAMURA Munetaka

要約：仮説検定，特にそのうちの棄却域の設定から区間推定への関連付けを重視した統計学の授業構想とその授業実践を報告する。この授業において重要な点が二つある。一点目は，仮説検定において対立仮説から棄却域を決定する過程である。そこでは，第１種の誤りを基準以下に抑えると同時に，第２種の誤りを最小化するということが核心部分になる。そこで決定的な役割を果たすネイマン・ピアソンの補題を，グラフを用いながら素朴な形で提示し，棄却域を決定していく。これにより素朴な感覚として極端な領域として理解される，棄却域の意味づけを確実にできる。二点目は区間推定の仮説検定との関連付けである。棄却域の設定時に現れる不等式の反転として，未知パラメータの信頼区間を設定する。検定と区間推定において同じような計算の意味理解と両者の関連が明確になる。

キーワード：ネイマン・ピアソンの補題，仮説検定，検定の反転，区間推定

Ⅰ　はじめに

まず，本論の対象とした授業について概要とその意図を述べる。対象授業は本学・教育学部において筆者が担当している専門科目の「数理統計学」である。例年数学教育系２年生15名程度（今年度は若干少ない13名）が受講している。なお，前期においてこの授業科目の数学的基礎となる確率論を全員が履修済みである。確率論を除いて高校時に学習した統計関係の内容は数学Ⅰの「データの分析」に留まり，ほとんど初めての統計学習と言ってもよい。

授業を始めるにあたって考えた点がいくつかあるが，それについて順次述べていく。まず，統計の初めての学習時には，この場合はこういう計算をせよ，次の場合はこうだ，ということを連綿と与える方法論に終始しがちである。これは流通している多くの大学１，２年次で必修科目用に使用されて教科書を見てもそれはうかがえる。例えば，決定的に重要なネイマン・ピアソンの補題を取り上げた教科書はレベルの高い［2］，［10］，［14］，［16］のようなものに限られる。後者は数学を専門とする３，４年次学生向けと思われるので，上に書いた大学１，２年次教科書という意味では範囲外だろう。このような状況から，極言すれば方法論のみの教授を一歩前進させたいと企図した。

具体的にまず一点目は，仮説検定において対立仮説から棄却域を決定する過程における棄却域決定の意味づけを明確にする，ということである。そこでは，第１種の誤りを基準以下に抑えると同時に，第２種の誤りを最小化するということがネイマン・ピアソン（[6]）の意図したところになる。この箇所において決定的な役割を果たすネイマン・ピアソンの補題を，確率分布のグラフを用いながら素朴な形で提示し，棄却域を決定していく。これにより素朴な感覚として極端な領域として理解される棄却域の意味づけが，数学的裏付けを基盤として確実になる。

[1] 教育実践創成専攻

次の二点目は区間推定の仮説検定との関連付けである。棄却域の設定時に現れる不等式の反転として，未知パラメータの信頼区間を設定する。検定と区間推定において同じような計算の意味理解と両者の関連が明確になる。

以上が事前に考えたことであるが，これを授業化するために各回の内容を次のようにした。

第１回　統計の考え方と導入問題
第２回　導入問題の考察
第３回　仮説検定の考え方
第４回　正規分布の平均の検定
第５回　t 分布
第６回　t 分布による正規分布の平均の検定
第７回　区間推定
第８回　適合度検定とその他の検定

第９回以降は具体的な学生個々が興味をもった事項に関する統計調査とそのレポート作成にあてている。上記の８回の計画はその基礎付けとなる統計的考え方の基礎として位置付けている。内容は数少ないが，標準的コースで扱う他の推定，検定は９回以降に各自必要になった段階で，教科書（[11]）や他の文献を学生自身が検索して調べるように伝えた。もちろん，その際には筆者に相談してもよいとのことも周知してある。なお，本論執筆時点では第５回まで進んでおり，授業実践の報告としてはその手前の第４回まで，学生のアンケートは第３回までの内容に関するものである。

教科書執筆者の考えはいろいろあるだろうが，筆者が考えた上記２点に関しては不足しているので筆者作成の資料を授業前に配布することにした。次章にそれを掲載し，補足説明を次々章以降で行う。なお，[11] を教科書に選んだ理由は，

・現実問題に近い演習問題が多数取り入れてあるところ
・Excel の操作が詳しく記載されている

の２点を重視したからである。本音を言えば，R や Python を対象ソフト（言語）にしたかったのであるが，以前の経験から教育学部学生には敷居が高いことがわかっていた。その経緯から Excel を学生が使う処理ソフトとして想定している。

なお，先行研究についても付言しておく。結論を言えば，上記二点，すなわち，ネイマン・ピアソンの補題の素朴な形での提示と検定の反転としての信頼区間の導出についての授業実践の先行研究については見つけることができなかった。論説としては古いもので [5]，[9]，教科書の形では [2]，[13] に検定の反転に類する記述がみられる。

Ⅱ　授業資料

この章では上述した筆者作成の授業資料を提示する（以下の枠内）。数年にわたって同内容の授業をしたので，それらの資料を今回向けに改定した。「第２章・導入問題」から「第３章・補足説明（正規分布における）ネイマン・ピアソンの補題」，までである（資料ページ番号 pp. 3–10）。なお，青字は授業後に追加した部分（授業時の口頭説明を主にして説明を意図的に省略した部分がある），赤字部分は授業後の修正部分である。なお一部にメモ用の余白部分があるがそのまま掲載している。

また些細なことであるが，図を左寄せにしてあるのもメモ用の余白を確保するためである。

2　検定の基本的な考え方

2.1　導入問題

2.1.1　予想する猫

統計の 2 つの柱として「推測」と「検定」があり，通常は推測を先に行って次に検定に進むという手順を取る（教科書もその順になっている）のだが，逆の方が理解しやすいのでこの授業では「検定」→「推測」の順に学習する。まずは高校の現行教科書の次の問題をいとぐちとして取り上げてみる。

試合結果を当てる猫

サッカーの試合の勝敗予想がよく当たるという猫に，あるトーナメント戦の勝敗を予想させたところ，30 試合中 19 試合が的中した。この結果から，この猫の予想は本当によく当たると判断してよいだろうか。

[ヒント]　仮に，当てずっぽうで予想したとしても，30 試合中 19 試合かそれ以上の試合で的中することは起こり得る。そこで，この猫が当てずっぽうで予想をしていると仮定したとき，30 試合中 19 試合以上で的中するという事象がどの程度起こりやすいか調べてみよう。

数学 I Advanced，令和 4 年度発行 (p.190 5 章「データの分析」4 節「仮説検定の考え方」より)

2.1.2　教科書の解答・解説の概要

当てずっぽうの予想が的中するかどうかを 30 回コイン投げに置き換えて考える。このコイン投げげ実験をコンピュータを用いて 10000 セット繰り返したところ，次のような結果が得られたられた。

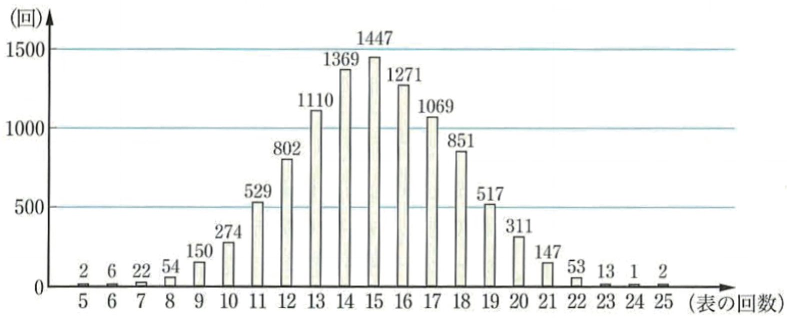

図 2.　コイン投げ 30 回を 1 セットとし，10000 セット実験した結果

上の結果から表が 19 回以上出たセット数は $517+311+147+53+13+1+2=1044$ であるから，この確率はおよそ 10.4% といえる。これでは猫の予想が当てずっぽうであったことを否定するのは難しい。(以上)

2.2　疑問

上記教科書の該当部分のを参照するとわかりやすい解説が加えてあり，十分に納得ができ何ら問題はないと思うが，あえて疑問をあげておく。あら探しをするのが目的ではなく，批判的に見直しつつ，統計的検定について掘り下げていければよい。

まず，まず,「確率」と「頻度 (相対度数)」を同一しているような表現が見られるが，これは本筋から離れ，かつこの授業ではすでに解決しているのでここでは取り上げない。学校によってはこの時点で確率は未習の可能性があるため，上の実験の記述になっているのである。本来は確率がわかれば，そちらを用いるところではある。それ以外では次の点が気になる。

疑問 1　なぜ 19 回そのもの，あるいはその近くではなく，19 回以上を調べるのか。

疑問 2　確率 10.4% では当てずっぽうを否定できないとしているが，普通の感覚では確率 1/10 程度は低いのではないか。

開き直ったような子供じみた質問だが，確たる理由をもって答えるのはなかなか難しいのではないだろうか。実際それにはかなりの準備が必要で，遠回りでも原点である J. ネイマンと E. ピアソンの検定理論に立ち返る。

以下では，青文字の部分は授業後に追加した説明である。

2.3　ネイマン・ピアソンの検定理論

2.3.1　帰無仮説，対立仮説，有意水準

ネイマン・ピアソン検定の一般的記法として，教科書問題中の当てずっぽうという仮定を帰無仮説と呼び H_0 と書き，観測された一見もっともらしい「猫の予想はよく当たる」を仮定化して対立仮説と呼び H_1 と書く。さらにほとんど起こりえない事象だと判断する基準の確率を有意水準と呼び，α と書くことにする。すなわち，猫の的中確率を p と書くと，

$$H_0 \ : \ p = 0.5, \quad H_1 \ : \ p = p_0, \quad \alpha = 0.05 \tag{S}$$

ここで，p_0 は 0.5 より大きい一つの定数である。これは一般的な設定の条件である。

疑問 2 の答え　ここで疑問 2 に関係することであるが，(S) で $\alpha = 0.05$ とかなり小さくしているのはなぜだろう。答えは手っ取り早く言ってしまえば，仮説の真偽判断において慎重を期すためである。合理的・科学的判断に正当性をもたせるための設定だと思えばよい。一般的な感覚より判断基準の確率を小さくするということでは確かにそれでよいのだが，$\alpha = 0.05$ に特別な意味があるわけではない。はっきり言うと，検定を行う者の恣意的な設定にすぎない。

究極は検定者の主観に依存するようで納得しがたい部分も残るが，疑問 2 はこれで一応済みとし後は疑問 1 に解答すべく考察していく。設定 (S) に戻って話をはっきりとわかりやすくするため，例えば $p_0 = 0.8$ として

$$H_0 \ : \ p = 0.5, \quad H_1 \ : \ p = 0.8, \quad \alpha = 0.05 \tag{$S_{0.8}$}$$

としてしばらく考察してみる。H_0 の $p = 0.5$ と H_1 の $p = 0.8$ のどちらがもっともらしいかを判断することになる。

2.3.2　棄却域と 2 種類の過誤，最強力検定

この猫が一般に 30 回予想をしたときの的中回数を X とする。今の場合は $X = 19$ であるが，別の機会には値が変わる。次に全体集合 $U = \{0, 1, \ldots, 30\}$ 内に，(α に依存する)H_0 の棄却域 R を作り，

- $X \in R$ ならば H_0 を棄却 (し，H_1 を採択) する
- $X \notin R$ ならば H_0 を棄却しない

と判断基準を決めておく。念のため言い添えておくと，R は U の部分集合である。さて，上記基準によってなされた判断は当然間違うこともあるわけで，H_0, H_1 の真偽と棄却・採択による正誤の状況に下の表のように名前をつけて分類しておく。同じ誤りではあるが，E_1 と E_2 では現実的な意味合いがちがう。

表 1. 2 種類の誤り

	H_0 を棄却しない	H_0 を棄却する
H_0 が真	正	第 1 種の誤り E_1
H_1 が真	第 2 種の誤り E_2	正

今の猫の予想に即して述べると，2 種類の誤り E_1, E_2 は具体的に次のようになる。

E_1 ：本当は猫に能力がないのに，よく当たるという誤判断を下す

E_2 ：本当はよく当たるのに，猫に能力がないという誤判断を下す

どちらを重視するかは立場によるだろう。

2.3.3　最強力検定

そこで，二項分布 $B(30,p)$ の確率関数を $p_r(x)$ と書くことにして，あらかじめ定めた α により，

$$P(E_1) = E_1\text{が起こる確率} = \sum_{x\in R} p_{0.5}(x) \leq \alpha \tag{I}$$

としつつ，同時に

$$P(E_2) = E_2\text{が起こる確率} = \sum_{x\notin R} p_{0.8}(x)\ \text{を最小にする} \tag{II}$$

ように R の構成を試みる。一般にこれらを同時に小さくすることはできないが，

(I) E_1 を優先的に抑えつつ　(II) E_2 も可能な限り抑え込むように R を選ぶ，

という発想である。H_0, H_1, α の下でこのように選ばれた R による検定を最強力検定と呼ぶ。

猫の飼い主が，この猫の能力を見極めたいと思っているとしよう。この人が慎重ならば E_1 を重要視して避けようとするだろうし，例えば猫の関係する者 E_2 を避けようとするだろう。疑問 2 の答えと同じく，ここでは前者の立場だと思えばよい。別の例でたとえると，裁判の判決だと思えばよい。帰無仮説を無罪，対立仮説を有罪と当てはめると，無罪の人を有罪にしないことを最優先し，有罪の人を無罪にすることはいたしかたないが最小限に抑える，ということである。

注：「$P(E_2)$ の最小化」というよりも，「検出力 $P(E_2^c) = 1 - P(E_2)$ の最大化」の概念の方がよく言及される。ここでは誤りの確率ということで共通化して扱いやすいので，前者を用いることにする。

2.3.4　ネイマン・ピアソンの補題による棄却域の構成

最強力検定を与える R は，はたして具体的には何だろうか。まずは H_0, H_1 に関連して B(30,0.5) の確率関数 $y = p_{0:5}(x)$ と $B(30, 0.8)$ の $y = p_{0.8}(x)$ のグラフを見ておこう。

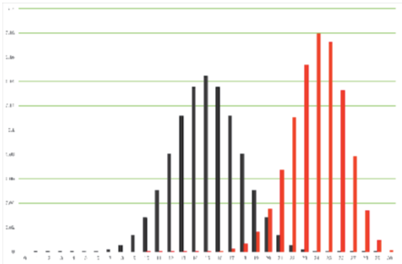

図 3.　$B(30, 0.5)$ と $B(30, 0.8)$　(Excel により描画)

当然のことながら，x が大きい部分では $p_{0.8}(x)$ の方が大きくなっている。もっとくわしく見ると，確率の比 (尤度比) $\dfrac{p_{0.8}(x)}{p_{0.5}(x)}$ は x の単調増加である。なぜならば $n = 30$ として一般に $0 < r < q < 1$ ($r = 0.5, q = 0.8$ を想定する) のとき

$$\frac{p_q(x)}{p_r(x)} = \frac{{}_n\mathrm{C}_x q^x (1-q)^{n-x}}{{}_n\mathrm{C}_x r^x (1-r)^{n-x}} = \left(\frac{1-q}{1-r}\right)^n \left(\frac{f(q)}{f(r)}\right)^x, \quad f(t) = \frac{t}{1-t} \quad (\ 0 < t < 1 \text{ において } f(t) > 0 \text{ かつ } t \text{ に関して単調増加})$$

であることに注意する。

そこで，棄却域 R　〜 そこでは H_1 が有利であるはずであり，そのためには $\dfrac{p_{0.8}(x)}{p_{0.5}(x)}$ が大きいほどよい 〜　は，

x が大きい方　($\dfrac{p_{0.8}(x)}{p_{0.5}(x)}$ が大きい，といってもよい　) を見て　$\displaystyle\sum_{x=t}^{30} p_{0.5}(x) \leq \alpha\ = 0.05$　を満たす最小の t を定め，$R = \{t, t+1, \ldots, 30\}$ とする　(ネイマン・ピアソンの補題 〜 数学的に上の (I), (II) が証明される 〜)

ことで得られる。実際に計算すると，$t = 20, R = \{20, 21, \ldots, 30\}$ になる。

したがって $(S_{0.8})$ H_0 : $p=0.5$, H_1 : $p=0.8$, $\alpha=0.05$ に関して最強力検定の際に起こりうる確率を計算してまとめると，次のようになる。棄却域 $R=\{20,\ldots,30\}$ に基づいて計算していることに注意する。

表 2. 最強力検定の際の各状況の確率（状況名の後の数値）

H_0 が真の場合は $B(30,0.5)$ から，H_1 が真の場合は $B(30,0.8)$ に基づいて計算する。

	H_0 を棄却しない	H_0 を棄却する
H_0 が真	正　0.9506	第 1 種の誤り E_1　0.0494
H_1 が真	第 2 種の誤り E_2　0.0256	正　0.9744

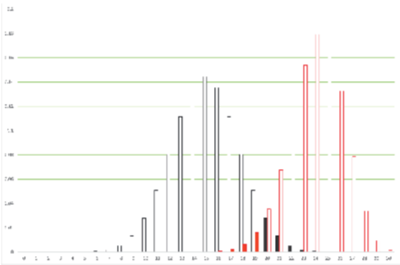

図 4.　第 1 種の誤り E_1 の確率 0.0494 (黒) と第 2 種の誤り E_2 の確率 0.0256 (赤)

2.3.5　一様最強力検定

他の $p(>0.5)$ の値の場合はどうだろうか。$0.5<p$ のとき，$\dfrac{p_p(x)}{p_{0.5}(x)}$ は x に関して単調増加であったから，前と全く同じ構成法を用いて棄却域 R は，

H_1 が有利である x が大きい方を見て，$\displaystyle\sum_{x=t}^{30} p_{0.5}(x) \leq \alpha\ (=0.05)$ となる最小の t を定め，$R=\{t,t+1,\ldots,30\}$ とすることにより得られる。したがって

p が 0.5 より大きければ何であっても，$R=\{20,21,\ldots,30\}$

である。

注：$p_p(x)$ はよくない記号であるが，意味はとれるだろう。下付きの p は数値としての p, 大きい p は確率関数 (probability function) の頭文字の p の意味である。記号は自由に作れるといっても，さすがにこれは良くなかった。

これより設定を H_0 : $p=0.5$, H_1 : $p>0.5$, $\alpha=0.05$ (S) としても，棄却域は一様に $R=\{20,21,\ldots,30\}$ となることがわかる。この意味でこの $R=\{20,21,\ldots,30\}$ による検定は一様最強力検定と呼ばれる。

ただし，(S) の下では H_1 の p が何かはわからないから，場合によっては第 2 種の誤りはそれほど小さくできない (図 5 を参照のこと)。それでもそれぞれ最小に抑えこんでいる。$(S_{0.8})$ の下では第 1 種，第 2 種のあやまりともにそれらが生じることを低確率で抑えることができ，かつ正しい場合を高確率で検出できたのだが，それは $p=0.5$ と $p=0.8$ の違いが大きかったことによる。

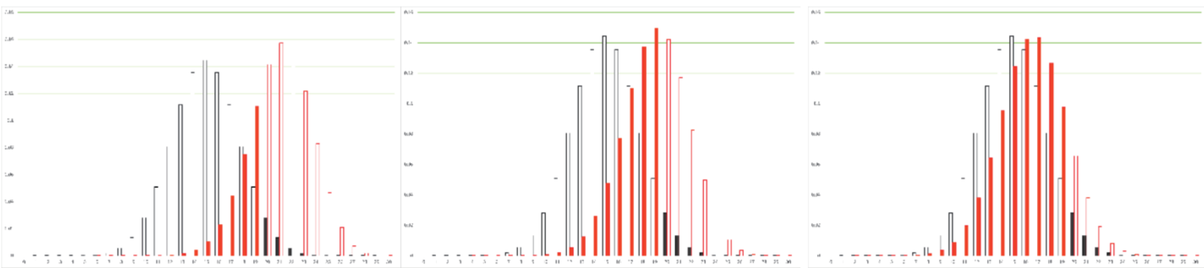

図 5.　$R=\{20,21,\ldots,30\}$ による第 1 種の誤り (黒) と第 2 種の誤り (赤) : 左 $H_1: p=0.70$, 中央 $H_1: p=0.63$, 右 $H_1: p=0.55$. それぞれ赤の部分の確率 (第 2 種の誤り確率) を計算しておいてほしい。

2.3.6 疑問への答え

ここでようやく疑問 1 に答えられるようになった。その前に留意事項を述べておく。建前としては帰無仮説，対立仮説，有意水準の設定は標本調査前に行うべきものである。その際には $p > 0.5$ が当然関心事であるにしても，前もって一つの値に特定できるわけではない。したがって，対立仮説 H_1 は $H_1 : p = 0.8$ のような単純仮説ではなく，$p > 0.5$ のうちのどれかという意味で，複合的な $H_1 : p > 0.5$ が望ましく，かつ，このように設定しても優れた性質の一様最強力検定を行うことが可能，すなわち，この最良の検定を与える棄却域 $R = \{20, 21, \ldots, 30\}$ を作ることができたのである。

[疑問 1 への答え]

今までの議論を振り返って，まず，上に述べたように，$H_1 : p > 0.5$ から一様最強力検定を与える棄却域 R が定まる。そして，その作り方から H_0 の仮定下で推論していることに注意して

$$H_0\text{を棄却する} \quad \Longleftrightarrow X = x \in R \quad \Longleftrightarrow P(X \geq x) = \sum_{t=x}^{30} p_{0.5(t)} < \alpha$$

となるから，結局，猫の的中回数である確率変数 X の実現値 x 以上になる確率を求め，それがあらかじめ定めたより小さいかどうかを見ればよいことになる。猫の例でいうと $x = 19$ だったので $P(X \leq 19) = 0.1002$ (シミュレーション結果とほぼ同じ) なるので最後の不等式を満たさないので，当てずっぽうを否定できないことになる。$P(X \leq x)$ は p 値と呼ばれ，一般的によく言及される。

(ネイマン・ピアソン理論を前提とすると，) このような背景が「以上の」という言葉に隠されているのである。

ただし，棄却域が今までに述べたように単純になるのは二項分布という単純で質の良い分布を用いているからである。前提とする分布によっては棄却域が別れたり複雑な形になりうる。もちろんそのような場合は,「何々以上」という形にならないことも留意するべきである。また，p 値に関しては，検定に際してこれのみでは実質が伴っていない，(実験) 内容がわからないとして，後で学ぶ信頼区間も提示することが求められている。

ところで，この猫の予想的中が 30 回中 20 回だとするとどのように結論を下せばよいのだろうか。単純な統計的な判断では $20 \in R$ だから，H_0 棄却，H_1 採択で,「この猫の予想はよく当たる」となってしまう。現実に即して冷静に考えてみて，はたしてこのような結論にしてよいのだろうか。

練習問題 1

あるコインを投げていたら，表が出ることが多いような気がした。そこで 200 回投げて調べたところ，114 回表が出た。このコインは表が出やすいのだろうか。Excel を利用して結論を出せ。

3　正規分布を用いた平均の検定

3.1　平均身長は小さいか

今度も具体的問題をきっかけにして考えていこう。この節で扱うのはいずれも大標本の場合である。

> 問題 2. 2019 年度学校保健統計調査によると，全国 17 歳女子高校生の平均身長は 157.9cm であり，大分県の 17 歳の女子高校生全体から無作為抽出された 450 人の身長の平均は $\bar{x} = 156.8$cm, 標準偏差は $u = 5.33$cm であった。2019 年に 17 歳であった大分県の女子高校生全体の平均身長を μ とするとき，μ が全国平均より小さいかどうかの仮説 $H_0 : \mu = 157.9,\ H_1 : \mu < 157.9$ を有意水準 5 % で検定せよ。(教科書 p.140, 問 9.3 改題，p.139 の例題 9.2 も参照のこと)

上の問題のように全体の中から一部分を抜き出して調べる調査をサンプリングという。他に，ランダムサンプリング，標本調査（教科書はこれを使用している），標本抽出などとも呼ばれる。対象となる集団全体を母集団といい，母集団から選び出された一部をサンプル（データ，標本）という。

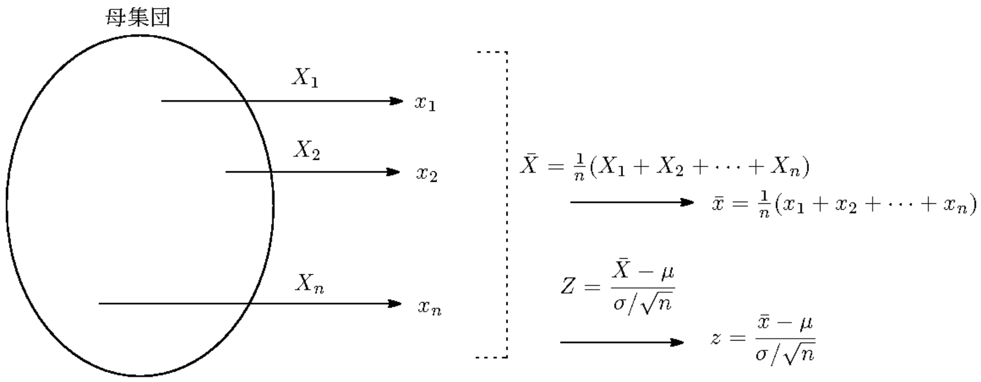

図 6.　母集団からのサンプリングと統計量

[問題の解説]　$\mu_0 = 157.9$ とする。問題に記されているように，この検定の仮説および有意水準を

$$H_0 : \mu = \mu_0, \quad H_1 : \mu < \mu_0, \quad \alpha = 0.05 \tag{3.1}$$

と設定する。すると，H_0 の下で大分県の女子高生 $n = 450$ 人の身長の標本平均 $\bar{X}$ は，n が大きいので

$$\bar{X} \sim N\left(\mu_0, \tau^2\right), \quad \tau = \frac{\sigma}{\sqrt{n}} \fallingdotseq \frac{s}{\sqrt{n}} \tag{3.2}$$

を満たし (教科書 p.105, 中心極限定理)，標準化すると，

$$Z := \frac{\bar{X} - \mu_0}{\tau} \fallingdotseq \frac{\sqrt{n}(\bar{X} - \mu_0)}{s} \sim N(0, 1) \quad (\tau \text{ は上記の通り}) \tag{3.3}$$

となる (教科書 p.107, 母分散が既知のときに $\cdots$)。以下 (3.2), (3.3) を用いたそれぞれの解答，および p 値を用いた解答を紹介する。

[(3.2) を用いた解答]　前節を振り返ると，猫の予想の場合は，$H_0 : p = 0.5,\ H_1 : p > 0.5,\ \alpha = 0.05$ の設定の下で，ネイマン・ピアソンの基本定理から棄却域は $B(30, 0.5)$ の分布の上側の $\{20, 21, \ldots, 30\}$ と決まった。それとまったく同じく，設定 (3.1) の H_0 の下では上記分布 (3.2) の $N\left(\mu_0, \tau^2\right)$ の下側 α 点を $x_-(\alpha)$ とすると，棄却域 R は $R = (-\infty, x_-(\alpha)]$ となる (くわしくは次小節の補足説明を参照のこと)。Excel を使って計算すると

$x_-(\alpha)$ の値：= NORM.INV(0.05,157.9,5.33/SQRT(450)) = 158.48667

[(3.3) を用いた解答] (3.3) を使った場合も，$N(0,1)$ の下側 α 点を $z_-(\alpha)$ として棄却域は $R_0 = (-\infty, z_-(\alpha)]$ になる。
Excel を使って計算すると

$z_-(\alpha)$ の値：= NORM.S.INV(0.05) =-1.64485

z 値：= SQRT(450)*(156.8-157.9)/5.33 =-4.37796

以上で双方とも，実現値 $\bar{x}, z$ がそれぞれ棄却域に入ることがわかる。下の注も参照のこと。

[p 値を用いた解答]
Excel を使って計算すると

p 値：= NORM.S.DIST(SQRT(450)*(156.8-157.9)/5.33, 1) =5.98978E-06 (5.98978×10^{-6})

または = NORM.DIST(SQRT(450)*(156.8-157.9)/5.33,157.9, 5.33/SQRT(450) 1)= =0
Excel の仕様上このようになるのかもしれないが，両者は本来同じ値である。正規分布において確率が 0 となることはありえない。

この値は $\alpha = 0.05$ よりはるかに小さい。

注：$z_-(\alpha)$ を $N(0,1)$ の下側 α 点とすると，

$$\alpha = P(Z < z_-(\alpha)) = P(\frac{\bar{X}-\mu_0}{\tau} < z_-(\alpha)) = P(\bar{X} < \mu_0 + z_-(\alpha)\tau) = P(\bar{X} < \mu_0 + z_-(\alpha)\cdot\frac{s}{\sqrt{n}})$$

となるので，$x_-(\alpha), z_-(\alpha)$ の間には $x_-(\alpha) = \mu_0 - z(\alpha)\cdot\frac{s}{\sqrt{n}}$ という関係がある。

以上 3 つの方法のいずれによっても H_0 は棄却，H_1 が採択され，大分県の女子高生の平均身長は全国より低いと結論される。
結局同じことを違う面から見ているだけだからどれでもよいのであるが，2 番目の標準化の計算方法が古くから一般的であろう。また，これは後で出てくる t 検定とも接続が良い。

3.2 補足説明：ネイマン・ピアソンの適用

一般に正規分布 $N(\mu,\sigma^2)$ の確率密度関数を g_{μ,σ^2} とする。すなわち，$g_{\mu,\sigma^2}(t) = \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(t-\mu)^2}{2\sigma^2}}$ である。このとき，身長の分布は正規分布としてよく，分散は同じで平均のみが異なる可能性 (この場合大分県の平均が小さい可能性) があるとする。すなわち，$\mu_0 = 157.9$ として全国の分布は $N(\mu_0,\sigma^2)$, 大分県の分布は $N(\mu,\sigma^2)$，かつ $n = 450$ が大きいので $\sigma^2 = s^2 = 5.33^2$ とするのがよさそうに見える。しかし，これは一人ずつを見た場合であって，大分県の方は 450 人の無作為標本の平均を考えているので，それに全国もあわせて 450 人平均を考える必要がある。すなわち双方とも 450 人の無作為標本の標本平均の分布を考え

$$\text{全国の分布は } N(\mu_0,\tfrac{\sigma^2}{n}),\ \text{大分県の分布は } N(\mu,\tfrac{\sigma^2}{n}),\quad n \text{ は上と同じく } n = 450 \tag{3.4}$$

として比較するべきである。 これは間違えやすいところなので注意を要する。

注：身長の分布なので上のように個人ごとの分布が正規分布としてよい。したがって (3.4) は正確な分布導出であり，中心極限定理による近似ではない (教科書 p.104, (3 行目) 母集団分布 $N(\mu,\sigma^2)$ からの $\cdots$)。

μ を $\mu < \mu_0$ となる定数とする。また，n(=450) 人の標本平均の分散を簡単のために τ^2 と書く：$\tau^2 = \frac{\sigma^2}{n}$. ここで猫のときに考えたように (3.4) の 2 つの分布の確率密度関数（起こりやすさ，もっともらしさ）の比を考えると，上記のように $\tau^2 = \frac{\sigma^2}{n} = \frac{5.33^2}{450}$ として

$$\frac{g_{\mu,\tau^2}(x)}{g_{\mu_0,\tau^2}(x)} = \frac{\frac{1}{\sqrt{2\pi}\tau}e^{-\frac{(x-\mu)^2}{2\tau^2}}}{\frac{1}{\sqrt{2\pi}\tau}e^{-\frac{(x-\mu_0)^2}{2\tau^2}}} = e^{-\frac{1}{2\tau^2}((x-\mu)^2-(x-\mu_0)^2)} = e^{-\frac{1}{2\tau^2}(\mu_0-\mu)(2x-\mu-\mu_0)}$$

9

は $\mu_0 - \mu > 0$ なので x に関して減少関数となる。

したがって猫の場合とは逆に，x が小さいほど H_1 にとっては有利となる。これより，棄却域は $N(\mu_0, \tau^2)$ の下側 α 点を $x_-(\alpha)$ とし，$(-\infty, x_-(\alpha)]$ となることがわかる。しかも，ここも μ に依らないので，この棄却域は一様最強力検定を与える。

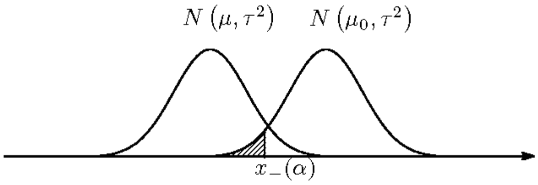

図 $N(\mu_0, \tau^2)$(全国と同じ) における対立仮説が H_1 のときの棄却域

3.3　両側検定：等しいか，等しくないか

今までの検定は棄却域が分布の片方だけにある片側検定と呼ばれる種類のものだが，次に両側検定と呼ばれる種類のものを扱っておこう。やり方はほとんど同じである。厳密なことをいうと付加条件が加わり長くなってしまうので，ネイマン・ピアソン理論に基づく理由付けの説明は省略する。

問題 3. 2019 年度学校保健統計調査によると，全国 17 歳男子高校生の平均身長は 170.6cm であり，沖縄県の 17 歳の男子高校生全体から無作為抽出された 500 人の身長の平均は $\bar{x} = 168.6$cm, 標準偏差は $s = 5.98$cm であった。2019 年に 17 歳であった沖縄県の女子高校生全体の平均身長を μ とするとき，μ が全国平均と異なるか (赤字部分は修正) どうかの仮説 $H_0 : \mu = 170.6$, $H_1 : \mu \neq 170.6$ を有意水準 1 % で検定せよ。(教科書 p.139 の例題 9.2 とほとんど同じ)

前と同じく $\mu_0 = 170.6, \sigma = 5.98, n = 500, \alpha = 0.01$ として，H_0 の分布 $N(\mu_0, \dfrac{\sigma^2}{n})$ の両端に等確率 $\frac{\alpha}{2}$ ずつの棄却域を作り，それらのどちらかに $\bar{x} = 168.6$ が入るかどうかを見るのであるが，前に述べたように一般的な標準化によって考えることにする。なお，教科書の u は厳密にいうと不偏分散であるが，s とほとんど同じであると思ってよい。ここでは s に書き換えた。u については次節の小標本の t 検定の際には重要になる。

問題は教科書と同じであるから解答は書かなくてもよいのだが，確認する意味で今までと同じ記号を使って述べてみる。

[解答 (解説つき)]　H_0 の下で，沖縄県の 17 歳の男子高校 500 人の身長の標本平均は $\bar{X}$ の分布は $N(\mu_0, \dfrac{s^2}{n})$ である。$n = 500$ が大きいので $\sigma = s$ と見なせるからである。正規化した $Z = \dfrac{\bar{X} - \mu_0}{\frac{s}{\sqrt{n}}} = \dfrac{\sqrt{n}(\bar{X} - \mu_0)}{s}$ は標準正規分布に従うが，H_1 による棄却域は両側に作り，$R = (-\infty, -z(\frac{\alpha}{2})] \cup [z(\frac{\alpha}{2}), \infty)$ となる。ただし，$z(\frac{\alpha}{2})$ は $N(0, 1)$ の上側 $\frac{\alpha}{2}$ 点である。

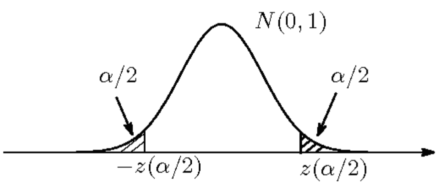

図　標準正規分布 $N(0, 1)$ における上側 $z(\frac{\alpha}{2})$ 点による両側棄却域

$\alpha = 0.01$ として Excel で計算すると $z(\frac{\alpha}{2}) = 2.58$ である：=NORM.S.INV(0.01/2) と入力すればよい。一方で Z の実現値は $z = \dfrac{\sqrt{500}(168.6 - 170.6)}{5.98} = -7.478$ となる。これらより，$z \in R$ がわかるので，H_0 を棄却し H_1 を採択する。言い換えると，沖縄の高校生の平均身長は全国より低い。

10

Ⅲ　授業資料の意図

1　導入問題設定

この導入問題は上述の通り高等学校「数学Ⅰ」にある検定に関する問題であるが，主な目的は，検定を先にし，検定から区間推定へという学習経路を取りたかったからである。これについては後で詳しく述べる。ここではなぜ，主に高校の教科書からとってきたのか，なぜこの猫の問題なのかについて書くことにする。

2018年（公式には元号なのだろうが，平成から令和にまたがり，わかりにくくなるので西暦表記にする）告示の高等学校学習指導要領，高等学校，数学Ⅰ「データの分析」，数学B「統計的な推測」の内容が改められ，両者ともに仮説検定の項目が加わった。もっとも，数学Bについては，それがほとんど履修されてなかったとはいえ，前回削除されていたものが復活したというべきかもしれない。そこで，導入問題はなるべく素朴な形のものがよいと考えていたので，高校教科書のものを利用することにした。新指導要領は2022年より実施されているので，当然ながら学生たちはその内容については知らず，彼らに高校で扱っている内容の紹介をするという意味も込めている。

問題設定前に参照した教科書は東京書籍，数研出版，啓林館のものである。このうち，後の２つについては仮説検定の定式化や手続きまで踏み込んで書いていたので，残りの素朴な東京書籍の問題を使うことにした。帰無仮説，対立仮説，有意水準の設定とその意味するところは，導入問題の考察過程で筆者流に説明していきたいと考えたからである。特に，対立仮説の設定が形式的でその意味するところがわからず，天下りに「このようにするものだ」と強制するのは避けたかった。

なお，この問題には猫がサッカーの予想をするという突飛な状況を扱っているが，それについて，「猫の予想はよく当たる」という肯定的結論に至る場合もありうる。単に統計的判断を下すことが現実的にはありえない結論に至った場合どのように判断したらよいのかも学生が考えることができるという状況も想定した。もっとも，そのようなもっともらしい意図だけでなく，猫が主役として登場する数学の問題というのも授業の場で心が和むだろう，と考えた面もある（このような場面設定を，多くの人が目にし，文科省の検定にもかかる教科書に，よく取り入れることができたと個人的には思っている）。

2　疑問について

資料の具体的内容を説明・検討することにしよう。この導入問題を初回に提示した。授業内で答えてもらうには，検討時間が不十分だろうと考え，その解答を次回授業までにLMS（Moodle）上に解答してもらうことにした。当該の教科書の説明を調べ，さまざまな確率を設定して計算する試行錯誤の時間が必要だったからである。ここでは，まったくの独力で解答するということは求めていない。そのようにすると，まったく手につかない学生が少なからずいると考えたからである。

なお，資料は問題部分のみを渡し，誘導と見られなくもないが，問題文後半の［ヒント］を付したのもそのためであるし，出典元の教科書を記したのも同様である。ただ，これらにある程度従うとしても，自分がそれにどのように考える，納得がいかない部分があればそれを明確にしておくようにとの注文はつけておいた。

上記資料中の問題文以降2.3.4まではその次回授業で配布したものであるが，あらかじめ想定していた

疑問１　なぜ19回そのもの，あるいはその近くではなく，19回以上を調べるのか。

疑問２　確率10.4¥%では当てずっぽうを否定できないとしているが，普通の感覚では確率1/10程度は低いのではないか。

に相当する解答は多数見られた。

疑問１に関しては，19回のみの確率を考え「0.05だから難しいことができた」として，猫に「能力あり」と認めるものが代表的なものである。疑問2に関しては，「0.1だからあまり大きくない」とした者と「0.1は小さい」とした者に分かれた。

双方半々といったところ（受講数が少数のため以降も定量的な分析は避ける）だが，前者は出典元かあるいはなんらか参考文献，情報を参考にしていると思われる。後者に関しては，上記のように普通の感覚であれば0.1は小さい確率であろう。判断の基準となる確率（有意水準）の設定はあくまで恣意的なものにすぎないから，授業の場でそれについて議論しても結論が出ないことは明らかであるので，資料中にあるような「慎重さ」に帰すことでこの疑問については早々に引きあげることにした。

この後，疑問１に絞って考察することにし，ネイマン・ピアソンの基本定理に進みそれにより棄却域を設定し，そこに猫の予想的中19回がそこに属するか否かということから，19回以上という枠組みが現れることを論じることになる。

さいころを１回投げて１の目が出ても誰もあやしまないが，そのあとにまた続けて１の目が出たら，多くの者がさいころに細工がしてあると疑いをもつだろう。そこで，その境界値は１回目と２回目の間にあると考えて，１回目の確率1/6と２回目の確率1/36の調和平均をとって0.048がそれであり，0.05はこれに近いとの話が[8]に載っているが，あくまでもこれは余談だろうし，書き方もそれを思わせるものである。この話の出所もわからないが，0.05が一般的になった経緯は筆者にはわからない。

３　ネイマン・ピアソンの補題の提示について

さて棄却域の設定であるが，ここではネイマン・ピアソンの補題に言及した。ネイマン・ピアソンの補題は対立仮説に有利な値から棄却域に組み込むという，少なくない数学の定理がそうであるように，言ってみればごく当たり前のことを示したに過ぎない。この重要結果は，補題というその名に反して，検定の基礎となる重要定理であるので，一部の文献でそうしているように「基本定理」と呼ぶべきものであるかもしれない。それは授業時に口頭でも伝えている。

それはともかく，これは授業においてはもっと重要なことなのだが，なるべく直観に訴える理解が望ましい。上に述べたように，ネイマン・ピアソンの補題の構造自体は単純ではある。しかし，初学者が数式のまま理解するのは困難であろう。

数式の証明を避け，かつ直観的理解をうながすためにはどうしたらよいかを考えて，前期に学習した各種の確率分布の確率関数，確率密度関数のグラフを活用することにした。すなわち，数式は用いずに単純仮説の場合の2つの二項分布のグラフを用いて説明を試みた。なお，第２種の誤りよりも検出力に言及するのが普通であろうが，一方は（第１種の）誤りの確率，他方で正しく判定する確率を考えるのは，初学者に混乱が生じやすいので，双方とも誤りの確率を比較している。

なお，「誤り」というよりも「過誤」ということば使いが一般的だろうが，固い表現は避けたかったのでこのようにした。やまとことばで全面的に置き換えるのは造語力の面で難しいだろうが，少なくともこの授業内においてはなるべく平易な表現を用いることにしている。このようにしていても，学生が授業以外の場で漢語にあわせるのは困難ではないだろう。

４　推定と検定の順序について

この先，t 検定を終えた後，学生たちは信頼区間の計算に望むことになる。通常，信頼区間の構成を終えた後，検定へと向かうのが広く行われ，ほとんどの教科書もこの順序で記述してある。

しかし，よく似た計算を扱いながら，両者がどのような関係にあるか触れられていない。これは

消化不良感を学習者に残すだろうことは容易に想像できる。その点を勘案して，検定を先にした。このようにすると，信頼区間を「検定方式の反転」として構成できるからである。

実際，R の t.test では，信頼区間も同時に出力される。ただし，片側検定に対応させる場合は$\pm\infty$を含む片側の信頼区間が出てくる。片側信頼区間は一般的ではないが，わかった上で用いれば何の問題もないだろう。実際 R では t.test 等でオプションとして alternative = "greater" または，"less" とすると，それに応じて片側の信頼区間が出力される。

とはいうものの，このような片側信頼区間を持ち出してくるのはある程度のレベルに達した後の話であり，通常用いる有界区間としての信頼区間は，両側検定の反転として算出される場合を主に述べ，片側信頼区間は注意程度にとどめる。

事前知識が全くない状態からの両側検定とそれに対応する通常の有界信頼区間と，予想も込めたある程度の見込みを持って臨む片側検定と片側信頼区間の意味の違いを強調できるだろう。もちろん，推定に属する信頼区間算出は，何の見込みも持たないという前者が重要視されるということを学生には詳しく説明する。今のところ，過去に準じて次のページのような説明資料を授業で使用する予定である，反面，信頼区間に対応する両側検定の理論的側面を述べきれていないところが弱い点である。

ここでは，単なる検定の反転としているので，検定から区間推定への接続が非常になめらかである。両者が密接に関係していることも見やすい。さかのぼって検定を見てみると，帰無仮説で仮定した値が信頼区間中にあるかどうかで，棄却するか否かの判断ができる。

なお，実験・調査結果を報告する場合，検定結果のp値のみでは不十分あるという批判が長らく続いている（[1], [3]）。検定を先に学習したが，実質的な内容を知らせるには信頼区間（もしくは効果量）を併記する必要があることを，十分に注意喚起したいと考えている。

一方で，難しく，かつ（ネイマン・ピアソンに従えば）重要なのは，上で確率を規定する不等式の集合条件から個別の値の不等式として取り出す部分であろう。[6] では以下のようなコメントが与えられている。

> We are inclined to think that as far as a particular hypothesis is concerned, no test based upon the theory of probability can by itself provide any valuable evidence of the truth or falsehood of that hypothesis.
>
> But we may look at the purpose of tests from another view-point. Without hoping to know whether each separate hypothesis is true or false, we may search for rules to govern our behaviour with regard to them, in following which we insure that, in the long run of experience, we shall not be too often wrong.

この考え方が，ネイマン・ピアソン理論が，帰納的推論ではなく，帰納的行動理論を指向するものであると言われる所以である。これは彼ら，特にネイマンの頻度主義思考が端的に現れているところだろう。筆者は全面的にこれに同意するわけではないが，こうした考えも首肯できるところはあるし，それを十分咀嚼して従っているかは別にして，実際に広く受け入れられてきた（[10]，[15]）。それゆえ，現在のような統計教育，あるいは初めて学ぶ統計の主流をなしてきたという経緯もある。今後，より進んでいかなる統計理論に進むにせよ土台となるものであると考えているし，ネイマン・ピアソン理論を意識せずに，仮説検定，信頼区間を応用で多用しているのが実際のところであろう。

以上のような，統計の基礎，有用性という面では，筆者もそれを大いに評価するし，それが学部２年次生の段階で統計を学ぶ意義があると考えている。

以下に信頼区間導出の資料を提示しておく。ただし，大標本の正規分布を用いた両側検定の反転

としての有界信頼区間の導出部分である。これが土台として t 分布を用いた信頼区間導出へと進む。

これからいよいよ区間推定に入る。以下ではサンプルサイズ n は十分大きいとする。$X_1, X_2, \ldots, X_n$ を平均 μ, 分散 σ^2 をもつ母集団からの無作為標本とするとき，定理４より小さい正の数 α に対して $Z = \dfrac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}}$ は,

$$P\left(-z(\frac{\alpha}{2}) \leq Z \leq z(\frac{\alpha}{2})\right) \fallingdotseq 1 - \frac{\alpha}{2} - \frac{\alpha}{2} = 1-\alpha$$

を満たすはずである。ただし，$z(\frac{\alpha}{2})$ は $N(0,1)$ の上側 $\dfrac{\alpha}{2}$ 点，すなわち $\displaystyle\int_{-\infty}^{z(\frac{\alpha}{2})} g_{0,1}(t)dt = 1-\frac{\alpha}{2}$ を満たす点である。

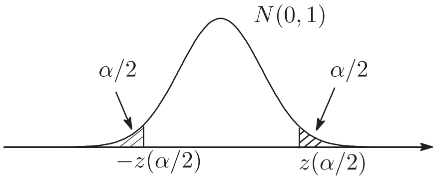

図 5. 標準正規分布 $N(0,1)$ における $z(\frac{\alpha}{2})$

元の統計量に戻していいかえると，確率 $1-\alpha$ で $-z(\frac{\alpha}{2}) \leq Z = \dfrac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \leq z(\frac{\alpha}{2})$ を満たすことがわかるのだが，$\bar{X}$ の実現値 $\bar{x}$ に対してはどのように見ればよいのだろう。観測者はその真偽を知る由もないが，未知の μ, σ に対してであってもこのサンプルに関して,

$$-z(\frac{\alpha}{2}) \leq z = \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}} \leq z(\frac{\alpha}{2}) \tag{1}$$

は満たされるかそうでないかのいずれかに既に決まっているはずである。それが観測者にはまったくわからないのは事実だが，今行ったサンプリングと同様のことを多数回繰り返せば,「確率」を逆手にとって実現値 $\bar{x}$ に関して (1) は $1-\alpha$ の頻度で満たされるだろう。このことを，信頼度 $1-\alpha$ で (2) が成り立つ，という言い方をする。

注：信頼度は，信頼係数とか信頼率とか呼ばれることもある。ここも有意水準と同じく% 表記の場合が多い。

(1) を μ に対する不等式

$$\bar{x} - z(\frac{\alpha}{2}) \cdot \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + z(\frac{\alpha}{2}) \cdot \frac{\sigma}{\sqrt{n}} \tag{2}$$

と書き換えて，信頼度 $1-\alpha$ で (2) が成り立つ，と言ってもよい。

この (2) に現れた下限および上限には未知の σ が残っているので，サンプルから推測した標本標準偏差 $s = \sqrt{\dfrac{1}{n}\displaystyle\sum_{k=1}^{n}(x_k-\bar{x})^2}$ を σ の代替として用いると,

$$\bar{x} - z(\frac{\alpha}{2}) \cdot \frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} + z(\frac{\alpha}{2}) \cdot \frac{s}{\sqrt{n}} \tag{3}$$

としてよい。というのは，n が大きいので，大数の法則から s も σ に近いはずだからである。

(3) で得られた下限および上限からなる区間

$$\left[\bar{x} - z(\frac{\alpha}{2}) \cdot \frac{s}{\sqrt{n}},\ \bar{x} + z(\frac{\alpha}{2}) \cdot \frac{s}{\sqrt{n}}\right] \tag{4}$$

を信頼度 $1-\alpha$ の信頼区間と呼ぶ。もちろん，すべてサンプルのデータから計算できることに注意しておく。信頼度 $1-\alpha$ の部分は % 表記されることが多いようである。

Ⅳ　まとめと今後の展望

第３回目終了時点でのアンケートからすると，当然のことながら学生のネイマン・ピアソンの補題に対する理解度はさまざまで，その受け止め方もさまざまであった。ただ回答提出が任意であったため，回答者数が少なく全面的に把握できているわけではない。学生の理解度については，一通り（統計調査を含めた）授業を終えた段階で再度問いかけて，その結果については別の機会に稿を改めて分析したい。

ネイマン・ピアソンとともに気になっていたのが，猫の予想結果が20回と棄却域に含まれていたときの判断である。統計的には帰無仮説棄却になるが，はたして現実的な理解として妥当なのか。かつてのサッカーワールドカップで，タコのパウルの予想的中率が高かったことがニュースになったことがあった。これはタコが横縞のデザインを識別できそれを好むからであり，横縞国旗のスペイン（優勝国），ドイツ，セルビアといったサッカー強国を選びやすかったということのようである([7])。猫がサッカーの予想を当てるという結果には，にわかには賛同しがたい，疑った方がよいという学生が（回答者中では）多かったのは，猫をおとしめるわけではないが，ものの見方として，統計リテラシーとして学生の多くが健全なものを持っていると考えている。

参考文献

[1] 南風原朝和，続・心理統計学の基礎 有斐閣（有斐閣アルマ），2014.

[2] 久保川達也，現代数理統計学の基礎 共立出版（共立講座 数学の魅力），2017.

[3] A.ラインハート，ダメな統計学：悲惨なほど完全なる手引書，勁草書房，2017.

[4] 俣野博，河野俊丈ほか，数学Ⅰ Advanced（文部科学省検定済高校教科書），東京書籍，2021.

[5] 松本今朝七，高等学校における推定・検定の扱い，日本数学教育会誌，52巻1号，1970.

[6] Neyman, J., & Pearson, E. S., On the Problem of the most Efficient Tests of Statistical Hypotheses, Philosophical Transactions of the Royal Society of London, 231A, pp.289-338, 1933.

[7] G.スミス，データは騙る　改竄・捏造・不正を見抜く統計学，早川書房，2019.

[8] 岡本雅典，鈴木義一郎，杉山公一，基本統計学，実教出版，2012.

[9] 奥野忠一，t-分布を用いる検定と推定，化学と生物7巻8号，1969.

[10] 大塚淳，統計学を哲学する，名古屋大学出版会，2020

[11] 大内俊二，データサイエンス指向の統計学，学術図書出版，2021.

[12] 竹村彰通，検定と標本の大きさ（「自然科学の統計学」の第6章），東京大学出版会，1992.

[13] 竹村彰通，統計（共立講座　21世紀の数学 14），共立出版，1997.

[14] 竹内啓，数理統計学―データ解析の方法，東洋経済新報社，1963.

[15] 竹内啓，歴史と統計学：人・時代・思想，本経済新聞出版，2018.

[16] 柳川尭，統計数学，近代科学社，1990.