判定水準以上の実力のある受験者を誤って却下する確率とその要因に関するノオト*

A note on the probability of rejecting the students who satisfy the requirement of the examiner

矢内浩文 (茨城大学工学部メディア通信工学科)
http://mu.dmt.ibaraki.ac.jp/yanai/

1　はじめに

試験や評価については，無数の理論や，実践で得られた経験則が存在するでしょうから，大学教師となって 13 年に過ぎない者の思いつきの考察は，ベテラン教師から見たら周知の事実であるという恐れが十分にあると思われますが，せめて確率の授業での演習問題の素材になるくらいの価値はあるだろうと判断し，敢えてここに数理的考察を紹介します．

試験問題を作成するたびに頭をよぎるのは，僕の大学教師経験の最初である今とは別の大学で，新任教師に配布された小冊子の一節です (今ではどこかへしまい込んだのか，なくしてしまったのか，見つけ出すことができません)．教育の大家 (外国人) の後進に向けた助言を日本語に訳したものだったと記憶しています．文言は覚えていないのですが,「試験は 0 点が出てもいけないが，満点が出てもいけない．0 点を取った者は自分がどれくらい身に付いたのか判断できないし，満点を取った者は，自分がどれくらい不足しているかが分からないからである」という風な内容でした．これまでは幸い，0 点というのはありませんでしたが，満点というのは何度か経験し，その度に「しまった」と思いました．これが現在の混沌とした教育体制の中で実効的な意味を持つのかどうか確信はありませんが，個人的なひとつのよりどころとなっていることは確かです．

さて，本題に入りましょう．

個人的には，論述式で部分点を設ける問題を出すことがほとんどなので，これまであまり真剣に考えてこなかったのですが，最近，数理的考察の演習問題としての興味から，正解か不正解のどちらかと判定する問題が多数ある試験の得点について考察してみたところ，予想外の結果が得られました．それは，他の教師のみなさんにも，少なからず参考になるかもしれない思いました．

試験の判定は，合格と不合格の場合や A+, A, B, C, D, E の 6 段階にするなどいくつかの方法がありますが，そのどれにも共通しているのは，ある得点以上かどうかに応じて判定を決めることです．ですから，以下に述べる「誤って却下する」とは，本来合格の実力のある受験者を不合格とすることはもちろん，本来 A 以上の評価に相当する実力のある受験者を B 以下と判定することも含みます．

考察の観点は，

- 誤って却下する確率はどの要因にどう依存するのか？
- 誤って却下する確率を小さくするためにはどうすればよいか？

*これは「矢内浩文: “判定水準以上の実力のある受験者を誤って却下する確率とその要因に関するノオト”, 大学教育研究開発センター年報, 第 8 号, pp. 59–62 (2004 年 3 月)」の印刷前原稿です．

です．2 番目については，できるだけ多くの問題を出すというのが最も単純な解決法ですが，現実的ではありません．問題を少数に抑えつつも誤って却下する確率を小さくするにはどうすればよいかを明らかにするのが課題です．

2　問題の設定と補足説明

このノオトの考察の前提は次の通りです．

1. 解答の判定は正解か不正解の何れかで，中間点はつけない．
2. 配点は均一である．
3. 受験者の実力は，ある問題に正解する確率 p である．

全問題数を n，正解数を m とすると，正解率は m/n です．各問題に対して正解となるか不正解となるかは問題が異なれば独立で，正解する確率は均一な p であるとすれば，n 問中 k 問が正解となる確率 $P_n(k)$ は二項分布に従い，

$$P_n(k) = {}_nC_k\, p^k(1-p)^{n-k}$$

となります．これを用いれば，正解率が T 以上 (ただし $0 \leq T \leq 1$) となる確率は

$$P\left(\frac{m}{n} \geq T\right) = \sum_{k=m'}^{n} {}_nC_k\, p^k(1-p)^{n-k}$$

となります．ただし，m' は $m' \geq T\,n$ を満たす整数のうち最小のものです．例えば，$T\,n = 21.6$ ならば $m' = 22$ で，$T\,n = 23.0$ なら $m' = 23$ です．

考察の前提 3 の「受験者の実力は，ある問題に正解する確率 p である」について補足しておきます．実力とは，十分に多数の同水準の問題に解答した場合の正解率と考えられます．だから，実力が出し切れなかったとは，運悪く苦手な分野の問題が多く出たということです．出題者から見て，A くんならある試験で 80 点取れるだろうと考えるとすれば，それは，その試験に対する A くん実力は 80%だと判断しているということであり，つまり，その試験に対する A くんの正解率は 0.8 であるとみなしているということになります．

すぐ下で見るように，ここにひとつの大きな誤解が潜んでいます．おおまかにいえば，実力が p の学生が試験全体で p 以上の正解率を得る確率は 50%前後です．つまり，判定水準を $T = p$ としてしまうと，実力が p の学生の半数が却下されてしまいます．簡単にいえば，実力が p の学生が合格するようにしたいなら，出題をもっとやさしくするか，判定水準を下げなければなりません (もちろん，判定水準を下げると新たな問題が発生します．実力のない学生を受け入れてしまう問題です)．

個人認証 (パターン認識) の分野では，受け入れるべき人を拒否することを false rejection，拒否すべき人を受け入れることを false acception と呼びます．そこでここでは，判定水準以上の実力のある受験者を誤って却下してしまう確率を，P_{FR} と書くことにします．

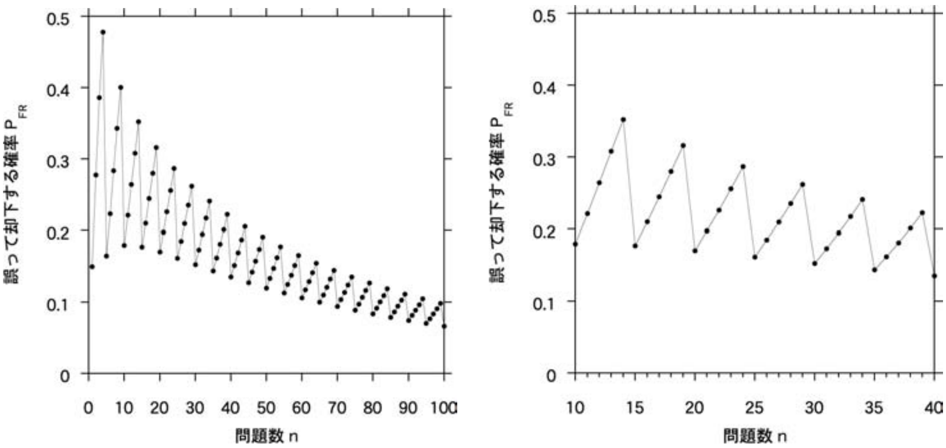
3　結果

まず特定の例を使って概要を説明しましょう．

- 実力：$p = 0.85$
- 判定水準：$T = 0.8$

の場合を見てみます．これは，十分に多数の，独立で同水準の問題が出題された場合に 85%に正解できる確率を持つ受験者を，限られた問題の評価で 80%以上の能力と判定できるかどうかという問題です．現在の茨城大学の評価でいえば A 評価となるかどうかの境界のところです．

図 1 に，判定水準以上の実力のある受験者を誤って却下してしまう確率を示します．これによれば，実力が判定水準より上回っていても，30 問程度では 20%前後の判定誤りが発生することが分かります．更に，問題数による変動が大きく，例えば $n = 25$ なら $P_{FR} \simeq 0.16(16\%)$ であるのに，$n = 29$ では $P_{FR} \simeq 0.26(26\%)$ にまで達します．



(a) (b)

図 1: P_{FR} と問題数 n の関係．実力 $p = 0.85$，判定水準 $T = 0.8$ の場合．(b) は (a) の一部を拡大した図．

念のため示しますが，実力 p が大きいほど，そして判定水準 T が小さいほど判定誤りが小さくなるというのは直感通りです (図 2)．

最後に，p と T の変化の影響を見るために，さまざまな場合について比較できるようにしたのが図 3 です．図から分かることをいくつか挙げると，図 3(a) によれば，判定水準が $T = 0.8$ の場合には，実力が $p = 0.9$ であっても，判定誤りを十分に減らすためには，問題数 n はおおむね 40 以上は必要であるといえるでしょう ($n = 40$ のとき $P_{FR} = 1.5\%$)．また，図 3(b) によれば，判定水準が $T = 0.6$ の場合には，実力が $p = 0.8$ であっても，判定誤り P_{FR} を 1% 以下にするには問題数 n が概ね 25 以上は必要です．

4　おわりに

最後に，重要な補足をしておきます．ここで考察したのは，「出題が独立」な場合でした．つまり，ある問題ができれば別の問題も確実にできるというような相関がない場合でした．もしも，問題に正の相関があると，実質上の問題数が少なくなります．よって，問題に正の相関がある場合にこの結果を適用する際には注意が必要です．例えば図 1 の場合で見ると，30 問出して判断誤りを小さく抑えたつもりが，一組の問題に強い正の相関があったとすると，実質上の問題数は 29 になり，現実の判断誤りは非常に大きくなってしまいます．

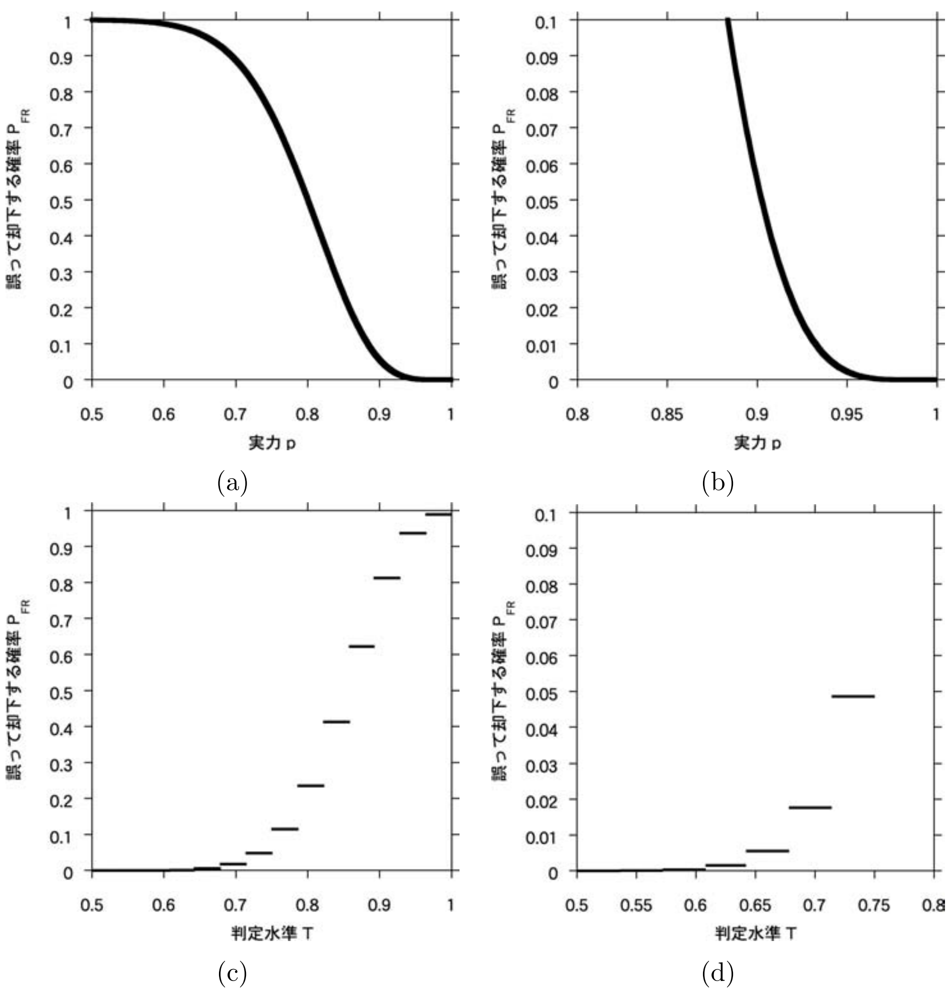


図 2: (a) P_{FR} と実力 p の関係，判定水準 $T = 0.8$．(b) は (a) の一部を拡大した図．(c) P_{FR} と判定水準 T の関係，実力 $p = 0.85$．(d) は (c) の一部を拡大した図．

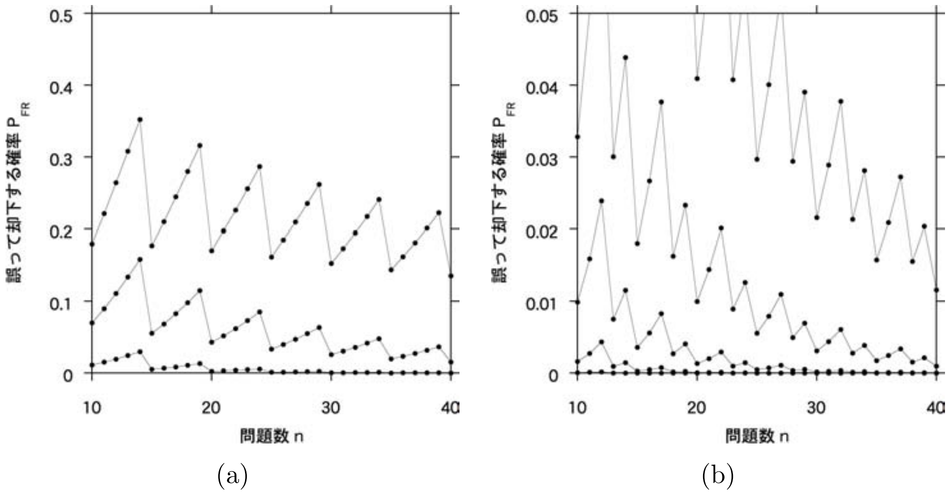


図 3: P_{FR} と問題数 n の関係．(a) は判定水準 $T = 0.8$ の場合で，実力 p は下から 0.95, 0.9, 0.85．(b) は判定水準 $T = 0.6$ の場合で，実力 p は下から 0.95, 0.9, 0.85,0.8,0.75 (0.95 はほとんど横軸に重なっている)．