

医学統計勉強会

東北大学病院循環器内科・東北大学臨床研究推進センター 共催

東北大学大学院医学系研究科 EBM 開発学寄附講座

宮田 敏

“Data! data! data!” he cried impatiently. “I can't make bricks without clay.”

From The Adventure of the Copper Beeches, The Adventure of Sherlock Holmes.

「データ！データ！データ！」ホームズはいらいらして叫んだ。「粘土が無ければレンガは作れない」

第5回 比率と分割表

1. 比と割合と率と Ratio, Proportion and Rate

ある疾患の発症率など、物事の**頻度 (frequency)** を議論するとき、「**比 (ratio)**」、「**割合 (proportion)**」、「**率 (rate)**」などの言葉が使われます。日常会話では、これらの概念は混同され多くの場合で区別なく用いられているようです。しかしデータ解析の場面では、これらは厳密に定義され、使い分けられています。

【定義】

- **比**：二つの量 A, B ($\neq 0$) が存在するとき、 A/B を比という。ただし、A と B は互いを含まないものとする。
 - 例：ある集団における男女の性比
 - 例：BMI (body mass index) = $\frac{\text{体重 (kg)}}{\text{身長 (m)}^2}$
 - 比と言うと“A:B”という書き方をよくするが、これは分数を見やすく書いたもの。
- **割合**：二つの量 A, B ($\neq 0$) が存在し、分子 A が分母 B に含まれるとき、 A/B を割合という。パーセンテージ (percentage) のこと。
 - 例：疾患の発症率 (= 発症した人の数 / 標本数)
 - $0 \leq \text{割合} \leq 1$.
- **率**：単位時間あたりの物事の発生頻度。
 - 率 = $\frac{\text{イベントの発生件数}}{\text{延べ観察時間}}$
 - 「率」には時間経過の概念が含まれており、単位時間あたりの「頻度、速度、強度」などの意味合いがある。

日本語には、さらに「比率」などと言う言葉もあつて複雑ですが、統計学の世界では、比率は上の定義で言う「割合」の意味で使われています。(母比率 = ある集団における、ある属性を持つ個体の割合)

医学統計で扱われるデータには、実数値で表される連続量の他に、カテゴリで表される離散量のデータもあります。本稿では、離散データの頻度に関する比率と分割表の、推定と検定について議論します。

2. 母比率の推定と検定

いま、 X_1, \dots, X_n をイベントの有無を表す確率変数とします。

$$X_i = \begin{cases} 1: \text{イベント有り} \\ 0: \text{イベントなし} \end{cases}, P(X_i = 1) = p, i = 1, \dots, n$$

ここで、 n は標本数を表し、イベント総数は $X = \sum_{i=1}^n X_i$ と書けます。このとき、母比率 p の推定値は、

$$\hat{p} = \frac{X}{n}$$

で与えられます。

標本比率 \hat{p} の、期待値、分散、標準誤差は以下の通り。

期待値： $E(\hat{p}) = p$

分散： $V(\hat{p}) = \frac{p(1-p)}{n}$

標準誤差： $S.E.(\hat{p}) = \sqrt{\frac{p(1-p)}{n}}$

信頼区間： 信頼水準 $100(1-\alpha)$ の信頼区間

$$\left(\hat{p} - z_{\alpha/2} \sqrt{\frac{p(1-p)}{n}}, \hat{p} + z_{\alpha/2} \sqrt{\frac{p(1-p)}{n}} \right)$$

ただし、 $z_{\alpha/2}$ は標準正規分布 $N(0,1)$ の上側 $100(\alpha/2)$ パーセント点。

仮説検定： p の仮説検定は、以下の通り。

$$H_0: p = p_0$$

$$H_1: p \neq p_0$$

有意水準 α の検定の棄却域 (rejection region) は、帰無仮説の元で

$$Z = \frac{|\hat{p} - p_0| - 1/2n}{\sqrt{\frac{p_0(1-p_0)}{n}}} > z_{\alpha/2}$$

Clopper-Pearson 正確信頼区間 (Clopper-Pearson exact confidence interval)

前ページの信頼区間と検定はよく用いられるものですが、標本数が十分大きいときに適用される近似を用いています。(通常 $np > 5$ 程度は必要とされています) さらに、母比率 p が 0 に近い (極めて希にしか起こらない) あるいは 1 に近い (ほとんど確実に起こる) 場合、 $\hat{p} = 0$ or 1 となって分散が 0 になってしまい信頼区間がうまく構成出来ないこととなります。

このようにサンプル数 n が小さい、もしくは p が 0 または 1 に近い場合に用いられるのが、以下の Clopper-Pearson 正確信頼区間になります。

Clopper-Pearson 正確信頼区間

$$\left(\frac{x}{(n-x+1)F_1+x}, \frac{(x+1)F_2}{(x+1)F_2+n-x} \right)$$

ただし、 $F_1 = F(\alpha/2, 2(n-x+1), 2x)$: 自由度 $2(n-x+1)$, $2x$ の F 分布の上側 $100(\alpha/2)$ パーセント点、 $F_2 = F(\alpha/2, 2(x+1), 2(n-x))$: 自由度 $2(x+1)$, $2(n-x)$ の F 分布の上側 $100(\alpha/2)$ パーセント点。

例 : いま、 $x=0, n=10$ とします。例えば、10 人の患者さんを観察してある疾病が起こる確率を推定しようとしたところ、実際には観察期間中に一人も病気を発症しなかった場合を想定します。このとき信頼区間は以下のように得られます。

$$\text{漸近信頼区間 : } \left(\hat{p} - z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right) = (0, 0)$$

$$\text{正確信頼区間 : } \left(\frac{x}{(n-x+1)F_1+x}, \frac{(x+1)F_2}{(x+1)F_2+n-x} \right) = (0, 0.308)$$

観察対象の患者さんが 10 人しかいなかったとすると、たとえ発症が観察できなかったとしても、それは発症確率が 0% であるからと考えるよりは、むしろ発症確率は正の値であるにもかかわらず、たまたま発症が観察されなかったと解釈するほうが自然です。その意味で、正確信頼区間のほうが納得できる結果です。

3. 2×2 分割表とリスクの解釈

第1回の講義でみたとおり、離散の数値をとるカテゴリ変数は度数分布によって要約されます。第1回で取り上げた定義を、再掲します。

定義 カテゴリの種類を**水準**と呼びます。ABO式血液型というカテゴリデータであれば、A型 B型 AB型 O型といった種類が水準に当たります。それぞれの水準に含まれるデータの個数を**度数**と言います。また、データ全体に占める度数の割合（度数/サンプル数）を**割合**もしくは**相対度数**と呼びます。この水準、度数、割合をまとめたものを**度数分布表**と言います。

分割表 (Contingency table) : 2種類のカテゴリデータの水準の組み合わせごとに、度数を求めた表。クロス集計表ともいう。

2×2分割表

	疾患有り	疾患なし
要因陽性	a	c
要因陰性	b	d

さて、疾患のなんらかの検査法があつて、その検査法の性能を評価したいと思います。このとき、どのような指標が望ましいでしょうか。

正答率 (accuracy rate) : 上の例でいえば、検査結果が陽性で実際に疾患があつた a 人と、検査結果が陰性で疾患がなかった d 人は予測が当たった人たちです。したがって正答率は以下のように求められます。

$$\text{accuracy rate} = \frac{(a+d)}{(a+b+c+d)}$$

正答率はわかりやすい概念ですが、欠点もあります。例えば非常にまれな疾患で、人口の0.1%しか罹患しない病気があつたとします。このとき、「患者さんの情報を一切無視して、すべての人に疾患なし、と判断する」という検査方法を考えたとします。(検査とは呼べませんが) この方法で患者さんを診断した場合、正答率は99.9%になります(0.1%しか罹患しないわけですから)。しかしこの方法では、もし仮に本当に罹患した患者さんがいた場合は、100%確実に診断を誤ってしまいます。それでは良い診断方法とは言えませんし、正答率が高いことのみを理由として検査方法を評価できないことを示しています。

そこで、以下の概念を導入します。

【定義】

疾患に罹患していたにもかかわらず、検査結果が陰性であった人の割合を**偽陰性率 (false negative rate)**という。疾患に罹患していた人の中で、検査結果が陽性であった人の割合 = $1 - \text{偽陽性率}$ を、この検査方法の**感度 (sensitivity)**と呼ぶ。 $a/(a+c)$

逆に疾患に罹患していないにもかかわらず、検査結果が陽性だった人の割合を**偽陽性率 (false positive rate)**、疾病に罹患していなかった人のうち検査結果が陰性だった人の割合 (= $1 - \text{偽陽性}$)を**特異度 (specificity)**と呼ぶ。 $d/(b+d)$

前に挙げた、まれな疾患に対して常に疾患なしと診断する方法は、特異度は100%ですが、感度は0%で、肝心の疾患に対してきわめて感度の悪い診断方法であるということになります。

一方予測の確からしさを測るために、以下の概念があります。

【定義】

検査結果が陽性だった人のうち、実際に疾患に罹患していた人の割合を**陽性的中率 (positive predictive value)** $a/(a+b)$ 、逆に陰性だった人のうち疾患に罹患していなかった人の割合を**陰性的中率 (negative predictive value)** $d/(c+d)$ と呼ぶ。

上で考えた検査方法は、要請、陰性の2つの値しか持ちませんでした。検査の方法によっては3段階以上の結果があったり、検査数値が実数地で得られたりする場合があります。その場合、検査数値のどこにカットポイントを置いて陽性、陰性を定義すればよいかという問題が起こります。検査数値が、3段階以上もしくは実数値をとる場合、カットポイントのとり場所により異なった複数の2×2分割表にデータをまとめることが出来ます。検査数値の順序に従って分割表を作り、感度、特異度が求めることができます。この一連の分割表の感度、特異度を元に、縦軸に感度、横軸に(1-特異度)をとったグラフを**ROC 曲線 (Receiver Operating Characteristic curve)**と呼び、ROC 曲線の下を面積を**AUC (Area Under Curve)**と呼びます。(次ページに例を示します) 感度、特異度ともに1に近いほうが望ましいわけですから、感度は1に近く(1-特異度)は0に近いグラフが望ましいことになります。

例：疾患の有無と検査数値について、以下のようなデータが得られたとします。
 (柳川 堯, 木 由布子「バイオ統計の基礎—医薬統計入門」近代科学社 より)

検査数値

	0	1	2	3	4	計
疾患有り	1	0	6	11	12	30
疾患なし	9	2	11	8	0	30

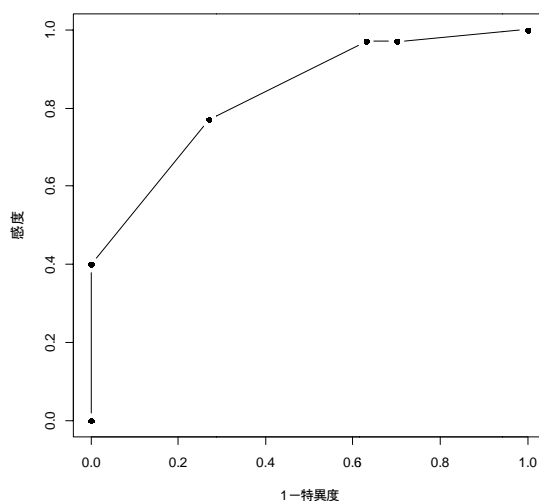
このとき、検査数値が 0 vs. 1~4 で表を分割すると以下の2×2 分割表が得られます。

	0	1~4
疾患有り	1	29
疾患なし	9	21

この分割表の感度、特異度はそれぞれ感度=0.97, 特異度=0.3 となります。同様に、カットポイントの位置を 0, 1 vs. 2~4, 0~2 vs. 3, 4, 0~3 vs. 4 のようにずらしていくと、感度と特異度の関係として以下が得られます。(柳川、荒木 (2012))

感度	1	1	0.73	0.37	0.3	0
特異度	0	0.40	0.77	0.97	0.97	1

ここから、ROC 曲線は以下ようになります。



4. 分割表の検定

第1回で検討した通り、分割表を用いて2つの要因の関連の強さを検定するには、一方の要因の水準によってデータを群に分け、群ごとに母比率が一定であるかどうかを検定します。

$$H_0: p_1 = p_2 \text{ 母比率が一定}$$

$$H_1: p_1 \neq p_2 \text{ 母比率が異なる}$$

もし、2つの要因の間に関連がなければ、群によって母比率は一定のはずですから、帰無仮説が受け入れられます。逆に帰無仮説が棄却されれば2つの要因の間に何らかの関連が存在したことになります。分割表の検定には **χ^2 検定 (chi-squared test)** と **Fisher の直接法 (Fisher's exact test)** が可能ですが、第1回で触れたように正確な p 値を計算できる **Fisher の直接法が第一選択** になります。分割表が大きすぎる、などの理由で Fisher の直接法を計算できないときのみ、**Yates の連続補正** を行った χ^2 検定を行ってください。

分割表とオッズ比

今、改めて 2×2 分割表を考えてみます。

2×2分割表

	疾患有り	疾患なし
要因陽性	a	b
要因陰性	c	d

このとき陽性の時の母比率（疾患ありの確率）を p_1 、陰性の時の母比率（疾患ありの確率）を p_2 とすれば、陽性、陰性それぞれのオッズは

$$\text{陽性オッズ} = \frac{p_1}{1-p_1}, \quad \text{陰性オッズ} = \frac{p_2}{1-p_2}$$

となり、オッズ比は

$$\text{オッズ比} = \frac{\frac{p_1}{1-p_1}}{\frac{p_2}{1-p_2}} = \frac{p_1(1-p_2)}{p_2(1-p_1)}$$

となります。帰無仮説 $H_0: p_1 = p_2$ は、オッズ比=1 を検定することになります。

5. シンプソンのパラドックス

分割表で考える2つの要因の双方に影響を与える因子を、**交絡因子 (confounding factor)** と呼びます。交絡因子を無視して検定を行うと、本来存在した関連が見えなくなる、あるいはその逆が起こることがあります。

いま、A, B 2つの要因の間の分割表を考えているとします。第3の要因 Z の値によって、標本が二つの層に分けられたとします。それぞれの分割表は以下の通りとします。

	Z1		Z2	
	B	Not B	B	Not B
A	80	16	50	452
Not A	160	160	10	452

いずれの場合も、オッズ比は 5.0 に等しく、要因 A があるほうがリスクが大きいことがわかります。これに対して、もし交絡因子 Z を無視して、2つの分割表を統合してしまうとすると次のようになります。

	B	Not B
A	130	468
Not A	170	612

この分割表からオッズ比を計算するとオッズ比は 1 になることがわかります。つまり、交絡因子 Z を無視したために、本来存在した要因 A のリスクが見えなくなってしまった、ということです。

交絡要因が存在するという事は、本来データの中に性質の異なるサブグループが存在するという事ですから、これを無視しては本質を見失います。

このような場合は、層ごとに分割表の検定を行い、それを統合する形で全体の検定を行います。このような検定を **Mantel-Haenszel 検定** と呼びます。

Take Home Message

1. 比と割合と率
2. 母比率の推定と検定
 - Clopper-Pearson の正確信頼区間
3. 分割表の推定
 - 感度、特異度、その他
 - ROC 曲線
4. 分割表の検定
5. シンプソンのパラドックス
 - 交絡因子
6. Mantel-Haenszel 検定

参考文献：

柳川 堯 (著), 荒木 由布子 (著)

「バイオ統計の基礎—医薬統計入門」近代科学社 (2010/02) ISBN-10:
4764903873

柳川 堯 (著) 「離散多変量データの解析」共立出版 (1986/12/1) ISBN-10:
4320013883

以上