福島県立医大 大平論文に対する公開質問状　オンライン（ZOOM）学習会

p 値とは？ 信頼区間とは？ 正しい理解へ

大倉弘之

元京都工芸繊維大学（統計数学）

2020 年 10 月 24 日（土）

はじめに

甲状腺がんが放射能への被ばくにより起こったのか．これが問題の焦点である．そこで，用いられている統計の議論で最もよく耳にするのが p 値という用語である．そして，この p 値が 0.05（5%）より小さければ，被ばくの影響がある（有意）と判断したり，0.05 以上であれば，影響があるとは言えない（非有意）と判断するような，議論が行われる．ところが，この議論や p 値の概念そのものにも多くの誤解や誤用がある．近年，特に医学生物学関係の学術雑誌等で p 値のみに基づいた議論の論文は受け付けないとか，信頼区間の有用性が再認識されるなど，世界的に注意喚起が行われるようになっている．

私の報告では，p 値や信頼区間の正しい理解や，誤用などの概略を解説した後，質問状に出てくる統計処理の意味などを改めて振り返って，解説する予定である．

参考文献

1. 「ロスマンの疫学　科学的思考への誘い」（篠原出版新社，2004年）（Rothman, K. J.:*Epidemiology; An Introduction*, Oxford Univ. Press, 2002 の和訳）
2. 柳川堯:「P 値　その正しい理解と適用」（近代科学社，2018）
3. Wasserstein, RL., Lazar, NA. Editorial:The ASA's statement on *p*-values: Context, process, and purpose. *Am Stat* 2016; 70: 129–133.（米合衆国統計協会声明の日本計量生物学会による和訳）: `http://www.biometrics.gr.jp/news/all/ASA.pdf`
4. Agresti, A.: *Categorical Data Analysis*, Wiley, 2013.
5. 奥村晴彦:「R で楽しむ統計」（共立出版，2016）
6. 奥村晴彦氏の web site の「統計・データ解析」のページ `https://okumuralab.org/~okumura/stat/`
7. 折笠秀樹: P 値論争の歴史, Jpn Pharmacol Thor（薬理と治療）vol. 46, 2018, 1273–1279.

p 値と統計的有意性

- p 値とは，特定の統計モデルの下で，実際に観測されたデータの出現確率とそれより偏った（極端な）データが出現するすべての場合の確率の総和を意味する（創始者は Fisher）.
- 特定の統計モデルとして典型的なのは，いわゆる「帰無仮説」の下で定まる統計モデルで，今回の論文に即して言えば，被ばく量と疾病の発生とが無関係との仮説の下で確率が計算される.
- p 値が小さいということは，実際に観察されたデータと仮定されている統計モデルが両立しにくいことを意味する.
- ネイマン・ピアソン (Neyman-Pearson) 流の統計的仮説検定では，仮説 H_0 から定まる統計モデルに基づいて p 値 p を計算し，
  $p < \alpha \implies \mathrm{H}_0$ を棄却（reject）（「有意である」という）
  $p \geqq \alpha \implies \mathrm{H}_0$ を採択（accept）（「有意でない」）←単に棄却しないという意味だが…
  α（5%など）は有意水準と呼ばれる.

p 値論争の歴史

1922	Fisher p 値導入，分割表の正確検定
1933	Neyman-Pearson 仮説検定の理論 （帰無仮説，対立仮説，α 過誤，β 過誤）
1986 頃	American J. Public Health で論争
1987	C. Poole（Rothman 一派）p 値関数
2010 以降	p 値への非難，Nature など
2016	ASA 声明

統計的有意性と P 値に関する ASA 声明

ASA=American Statistical Association

1. P 値はデータと特定の統計モデル（訳注:仮説も統計モデルの要素の一つ）が矛盾する程度を示す指標の一つである。
2. P 値は，調べている仮説が正しい確率や，データが偶然のみで得られた確率を測るものではない。
3. 科学的な結論や，ビジネス，政策における決定は，P 値がある値（訳注:有意水準）を超えたかどうかにのみ基づくべきではない。
4. 適正な推測のためには，すべてを報告する透明性が必要である。
5. P 値や統計的有意性は，効果の大きさや結果の重要性を意味しない。
6. P 値は，それだけでは統計モデルや仮説に関するエビデンスのよい指標とはならない。

p 値の目安（by Stuart Pocock, 2015）

$p < 0.001$	Overwhelming evidence（極めて強い証拠）
	高度有意
$0.001 \leqq p < 0.01$	Strong evidence（強い証拠）
	1%有意
$0.01 \leqq p < 0.05$	Some evidence（証拠あり）
	5%有意
$0.05 \leqq p < 0.1$	Insufficient evidence（証拠不十分）
	有意傾向
$0.1 \leqq p$	No evidence（証拠なし）
	非有意

背景情報の適切な扱いの下で，統計的な議論を活かすことができる．

p 値とは

簡単な例を考える．観察データが次に分割表に記されているとする．

	疾病有	疾病無	計
暴露有	4	6	10
暴露無	2	18	20
計	6	24	30

分割表の周辺度数（縦横の計）を固定したまま，

4	6
2	18

の部分を

6	4
0	20

5	5
1	19

4	6
2	18

3	7
3	17

2	8
4	16

1	9
5	15

0	10
6	14

と変化させる．

p 値とは（つづき）

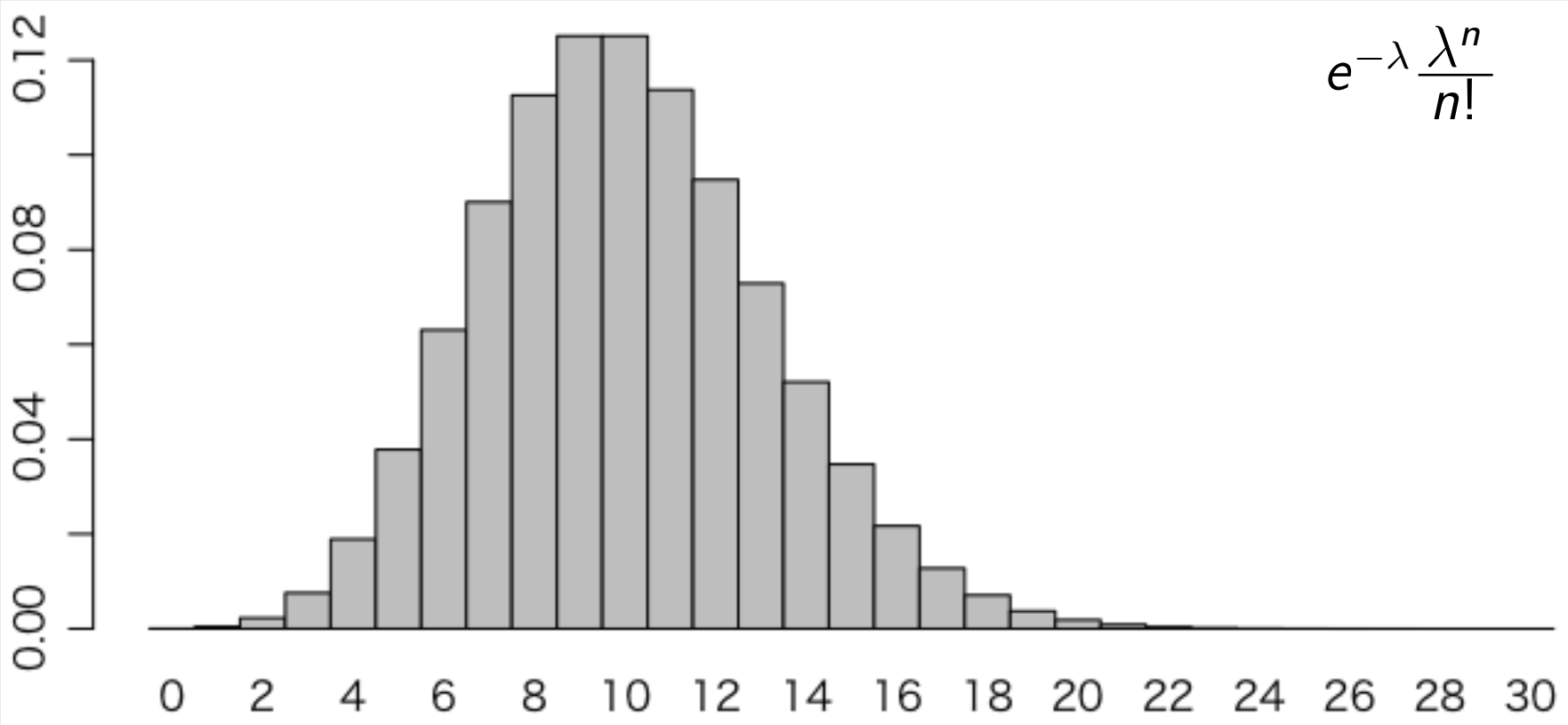
6	4	5	5	4	6	3	7	2	8	1	9	0	10
0	20	1	19	2	18	3	17	4	16	5	15	6	14

暴露と疾病が無関係として，これらの確率は，順に
0.00035，0.00849，$\underline{0.06720}$，0.23039，0.36718，0.26111，0.06528
と計算される．ここで，p 値は
$p = 0.00035 + 0.00849 + \underline{0.06720} + 0.06528 = 0.14132$ すなわち約 14.1%である．
一番右端も確率がより小さいという意味で偏ったデータ（両側 p 値）．
5%有意となるデータは左端の 2 つのみ（$p = 0.035\%$，0.88%）．

下側 p 値：観察データから左端までの確率の総和 $p = 7.604\%$
上側 p 値：観察データから右端までの確率の総和 $p = 99.12\%$
これらを，片側 p 値と総称する．

信頼区間の簡単な例（奥村晴彦氏の web site からの引用）

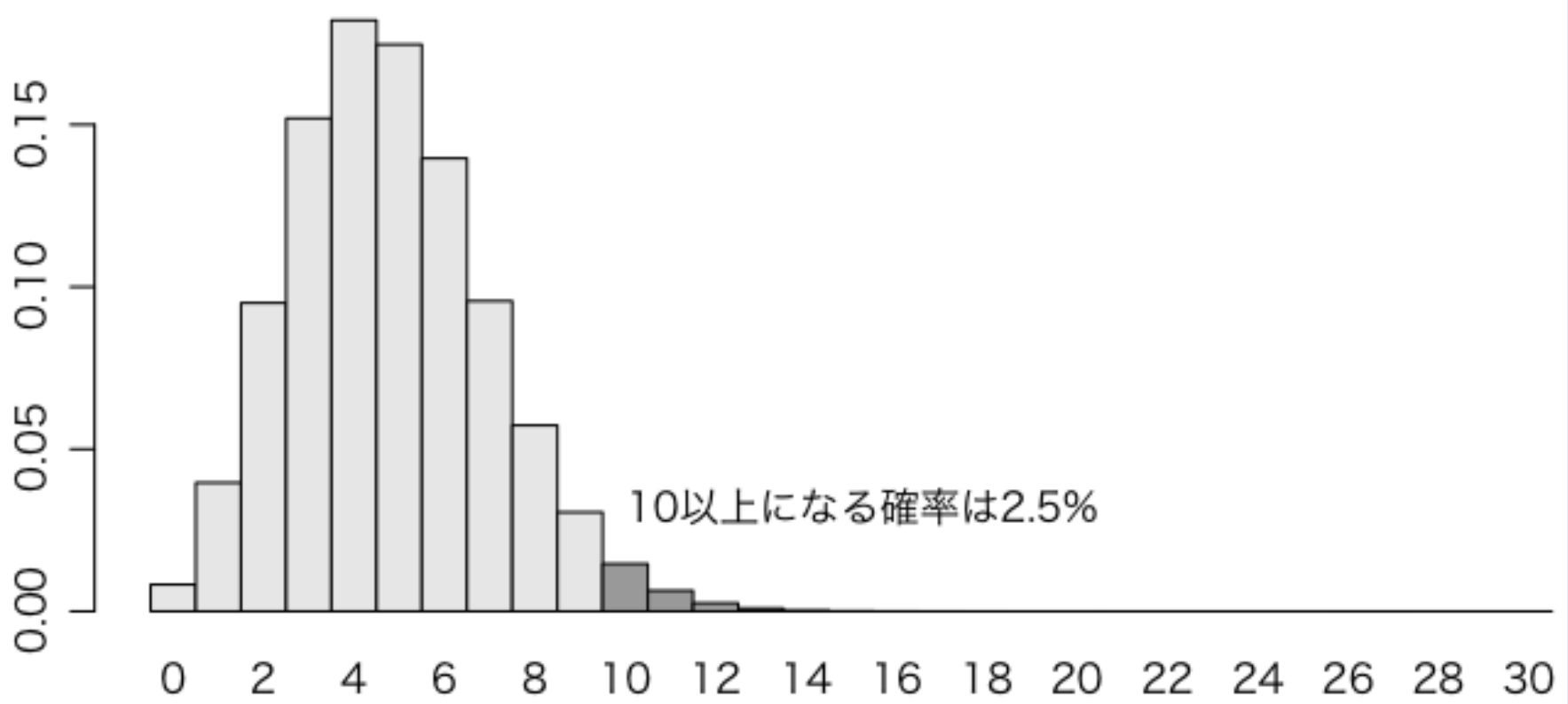
ポアソン分布を例にとる．例えば 1 分間当たり放出する放射線の個数が該当する．1 分間で 10 個の放出を観測したとする．平均が $\lambda = 10$ であるポアソン分布の確率分布は



10 回観測される確率を最大とする λ の値がちょうど 10 である．その意味で，$\lambda = 10$ は最尤推定値と呼ばれる．

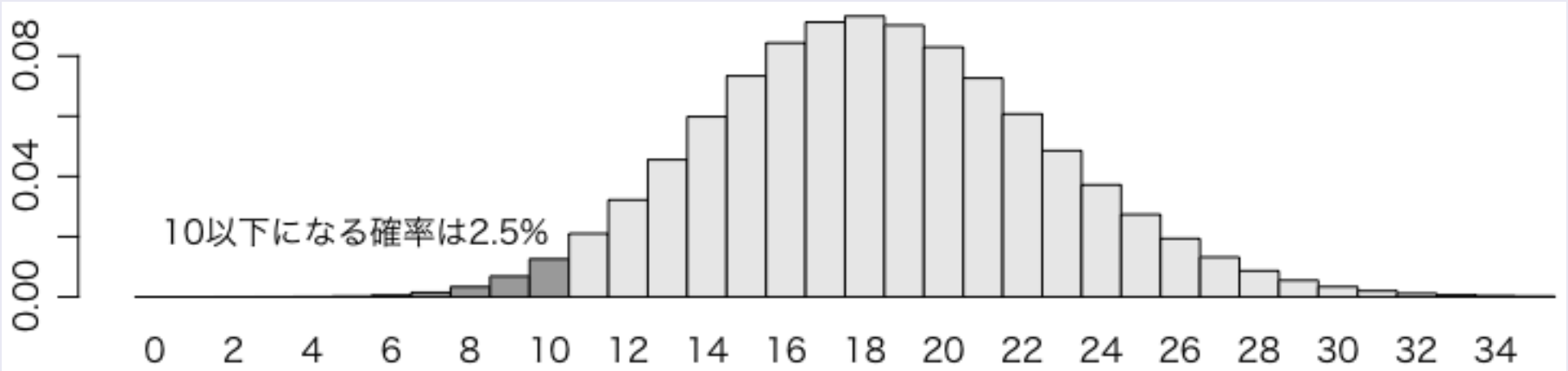
信頼区間の簡単な例（奥村晴彦氏の web site からの引用）

λ の値に対する信頼区間は，次のようにして求める．まず，λ の値を色々動かして，10 個以上を観測する確率（上側 p 値）が 2.5% になるようにすると，そのとき，大体 $\lambda = 4.8$ である



信頼区間の簡単な例（奥村晴彦氏の web site からの引用）

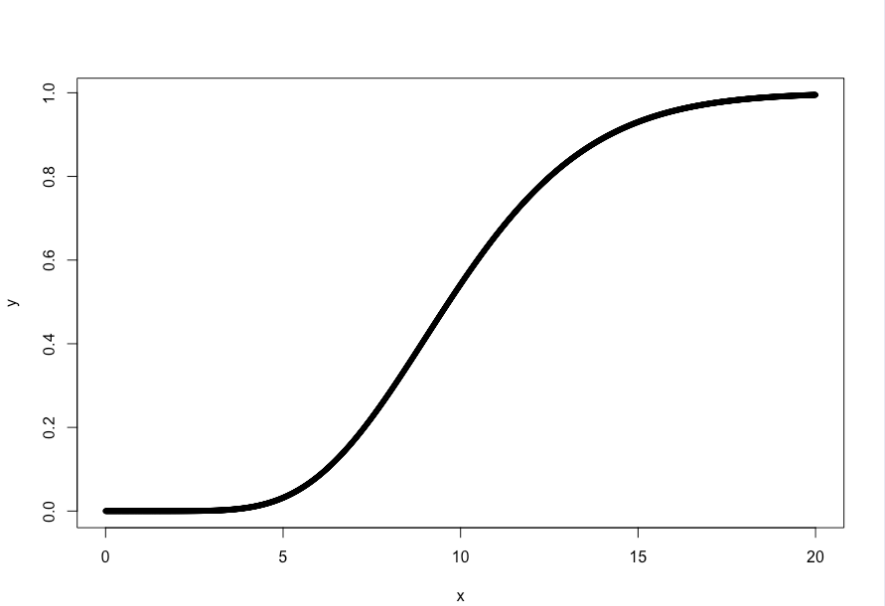
次に，やはり λ の値を色々動かして，10 個以下を観測する確率（下側 p 値）が 2.5% になるようにすると，そのとき，大体 $\lambda = 18.4$ である



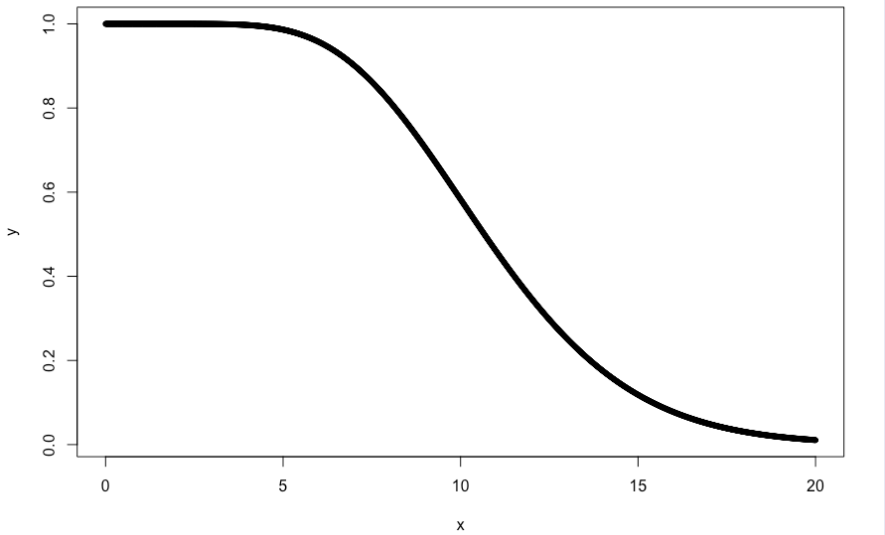
以上の $\lambda = 4.8$ と $\lambda = 18.4$ は各片側 p 値の和である 5% の水準で観測値 10 個と確率分布が両立する，あるいは矛盾するギリギリの境目の値であり，これらを端点とする区間 $4.8 \leqq \lambda \leqq 18.4$ が λ の 95%信頼区間である.

このように，信頼区間は様々な仮説に対する検定結果から逆算して得られる.

p 値関数と信頼区間



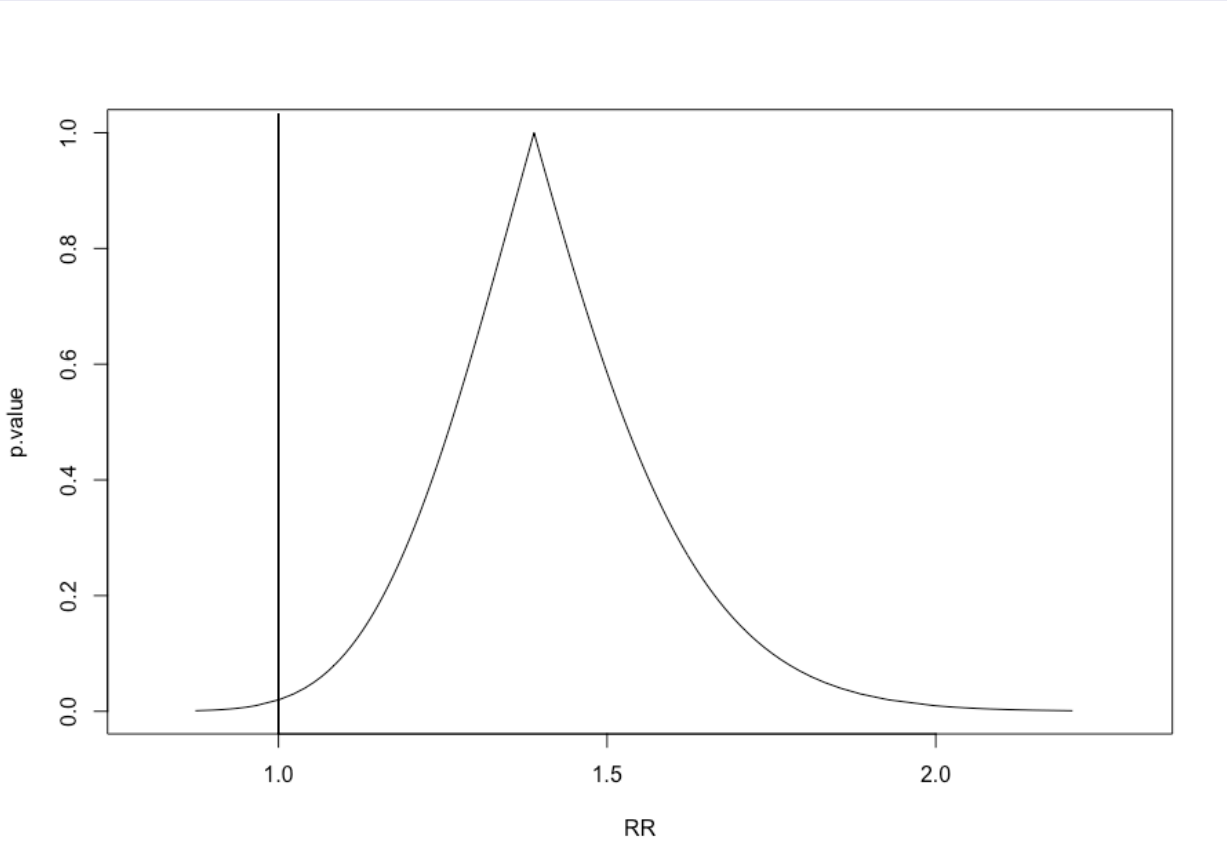
←上側 p 値関数



←下側 p 値関数

p 値関数と信頼区間

質問状 1. (iii) の分割表の相対リスク (RR) に対する p 値関数



有意性検定は $RR = 1$ のところだけの議論. p 値関数は RR の可能な値全てについての情報を持っている. 全体を見た総合的な判断が不可欠.

ご清聴ありがとうございました.