外国語教育メディア学会（LET）関西支部 メソドロジー研究部会 2014 年度 第 6 号報告論集
小林 雄一郎（pp. 85–95）

コーパス言語学研究における頻度差の検定と効果量

小林 雄一郎
日本学術振興会

概要

コーパスを用いた言語研究では，複数の頻度データを比較することが多い。そして，一般的には，複数の頻度データに統計的に意味のある差（有意差）が存在するかどうかを検証するために，有意性検定と呼ばれる統計処理が行われる。しかしながら，検定には，サンプル・サイズが大きくなれば，結果として得られる *p* 値が小さくなる傾向があることが知られている。そして，*p* 値が小さいと，実質的な差がない場合にも，「有意差あり」という誤った解釈が導かれる危険性がある。そのようなときには，検定の結果だけでなく，何らかの効果量を提示する必要がある。以下，本稿では，頻度差の検定を行う際の注意点を述べ，効果量（オッズ比，φ係数，クラメールの *V*）についての解説を行う。

Keywords: コーパス，検定，効果量，サンプル・サイズ

1. はじめに

コーパスを用いた言語研究では，複数の頻度データを比較することが多い。そして，一般的には，複数の頻度データに統計的に意味のある差（有意差）が存在するかどうかを検証するために，有意性検定 (significance testing) と呼ばれる統計処理が行われる。例えば，Hommerberg and Tottie (2007) は，カイ二乗検定 (chi-square test) と呼ばれる有意性検定を用いて，イギリス英語とアメリカ英語における *try to* と *try and* の頻度分布には，有意差が存在することを示している。また，Lorenz (1998) は，同様にカイ二乗検定を用いて，ドイツ人英語学習者と英語母語話者による形容詞強調の頻度に有意差があることを示している。

しかしながら，検定には，サンプル・サイズが大きくなれば，結果として得られる *p* 値が小さくなる傾向があることが知られている。そして，*p* 値が小さいと，実質的な差がない場合にも，「有意差あり」という誤った解釈が導かれる危険性がある。そのようなときには，検定の結果だけでなく，何らかの効果量 (effect size) を確認する必要がある。以下，本稿では，コーパス言語学研究において頻度差の検定を行う際の注意点を述べ，効果量についての解説を行う。[1]

2. 頻度差の検定

2.1 カイ二乗検定

例えば，2 つの異なるコーパスを検索し，類義語 X と Y の頻度を以下のような分割表 (contingency table) の形式で集計したとする（表 1)。

表 1

頻度差の検定のためのサンプル・データ (1)

	Corpus A	Corpus B
Word X	12	9
Word Y	8	11

そして，Corpus A と Corpus B では，Word X を使う割合と Word Y を使う割合に有意差があるのかどうかを知りたいとする。そのような場合に，コーパス言語学の分野で最もよく用いられる検定手法は，カイ二乗検定である (McEnery, Xiao, & Tono, 2006, p. 55)。[2]

カイ二乗検定では，両方のコーパス間で頻度の差がないと仮定した場合の理論値（期待値）と実際に得られた頻度（観測値）のずれの度合いを調べて，その度合いが大きいほど，統計的に意味のある差が存在するとみなされる (Oakes, 1998, pp. 24–25)。ここでは，フリーの統計処理ツールである R を使って，表 1 のデータにカイ二乗検定を実行し，Corpus A と Corpus B において，Word X を使う割合と Word Y を使う割合に有意差があるのかどうかを調べる。[3]

R でカイ二乗検定を行うには，`chisq.test` 関数を用いる。[4] 以下の R スクリプトのうち，行頭の `>` は 1 つのコマンドの開始位置を示すものであり，分析者が入力する必要はない。また，`#` で始まる部分はコメントであり，実際の処理では省略可能である。つまり，実際の処理では，ボールド体の部分のみを入力すればよい。

```
> # データの入力
> dat1 <- matrix(c(12, 8, 9, 11), nrow = 2)    # nrow で行数を指定
> colnames(dat1) <- c("Corpus A", "Corpus B")    # 行ラベルを指定
> rownames(dat1) <- c("Word X", "Word Y")    # 列ラベルを指定
> # 入力したデータの確認
> dat1
      Corpus A Corpus B
Word X       12        9
Word Y        8       11
> # カイ二乗検定
```

```
> chisq.test(dat1, correct = F)    # イェーツの連続補正なし

      Pearson's Chi-squared test

data:  dat1
X-squared = 0.9023, df = 1, p-value = 0.3422
```

表 1 のデータにカイ二乗検定を行った結果は，上記ボックスの一番下の行に表示されている (X-squared = 0.9023, df = 1, p-value = 0.3422)。ここでの p 値は 0.3422 であり，Corpus A と Corpus B における Word X と Word Y の頻度の割合には，有意差が見られなかった。

2.2 検定とサンプル・サイズ

検定は，数多くの分野で活用されている統計手法である。しかしながら，コーパスを使った言語研究では，分割表の各セルに非常に大きな数値が入ることも稀ではなく，そのような場合，検定結果として得られる p 値が非常に小さなものとなる。以下に，サンプル・サイズが検定結果に影響を与える具体例を示す。これは，表 1 における 4 つの頻度を全て 10 倍にしたデータに対して，カイ二乗検定を行った結果である。

```
> # データの入力
> dat2 <- matrix(c(120, 80, 90, 110), nrow = 2)
> colnames(dat2) <- c("Corpus A", "Corpus B")
> rownames(dat2) <- c("Word X", "Word Y")
> # 入力したデータの確認
> dat2
      Corpus A Corpus B
Word X      120       90
Word Y       80      110
> # カイ二乗検定
> chisq.test(dat2, correct = F)

      Pearson's Chi-squared test

data:  dat2
X-squared = 9.0226, df = 1, p-value = 0.002667
```

上記のカイ二乗検定の結果を見ると，p 値は 0.002667 であり，Corpus A と Corpus B における Word X と Word Y の頻度の割合には，1%水準での有意差が見られた。つまり，4 つのセルに入っている数値の比率がまったく同じであったとしても，数値そのもの（サンプル・サイズ）が大きくなればなるほど，有意差が出やすくなるのである。[5] 例えば，大規模コーパスにおける機能語の頻度を集計すると，個々のセルに入る数値が数千や数万にのぼることもあり得る。従って，コーパスを用いた言語研究では，p 値だけでなく，効果量と呼ばれるサンプル・サイズによらない指標を確認することが不可欠である。

3. 効果量

3.1 オッズ比

カイ二乗検定の効果量として，オッズ比 (odds ratio)，リスク比 (risk ratio)，リスク差 (risk difference) などの様々な指標が存在するが，その中でオッズ比が最も一般的かつ有用であるとされている (Bourenstein, Hedges, Higgins, and Rothstein, 2009, p. 36; Field, Miles, and Field, 2012, p. 826)。オッズ比は，ある事象の起こりやすさを 2 つのデータで比較するために用いられる。前掲の表 1 の例で言えば，「Corpus A における Word X の頻度と Word Y の頻度の割合（比率)」を「Corpus B における Word X の頻度と Word Y の頻度の割合（比率)」で割ったものがオッズ比である。[6] 以下の R スクリプトと表中の数値との関係では，`dat1[1, 1]`（1 行目 1 列目）が 12，`dat1[2, 1]`（2 行目 1 列目）が 8，`dat1[1, 2]`（1 行目 2 列目）が 9，`dat1[2, 2]`（2 行目 2 列目）が 11 にそれぞれ対応している。

```
> # オッズ比の計算
> (dat1[1, 1] / dat1[2, 1]) / (dat1[1, 2] / dat1[2, 2])
[1] 1.833333
```

上記の R スクリプトからも分かるように，オッズ比の計算は非常に簡単である。そして，その結果は，「Corpus A で Word X が使われる割合は，Corpus B で Word X が使われる割合よりも約 1.83 倍である」ということを意味しており，解釈も容易である。[7] なお，オッズ比は 0 が下限であり，もし 1 を下回っている場合は，Corpus B で Word X が使われる割合よりも Corpus A で Word X が使われる割合の方が大きいことを意味する。

また，前述のように，効果量はサンプル・サイズの影響を受けないため，表 1 における 4 つの頻度を全て 10 倍にしたデータからオッズ比を計算しても，その結果は，10 倍する前のデータの結果と同じである。

```
> # オッズ比の計算
> (dat2[1, 1] / dat2[2, 1]) / (dat2[1, 2] / dat2[2, 2])
[1] 1.833333
```

オッズ比の計算には，R の vcd パッケージの oddsratio 関数を使うのが便利である。[8] この関数を使えば，オッズ比の信頼区間や *p* 値も簡単に計算することができる。

```
> # パッケージのインストール
> install.packages("vcd", dependencies = T)     # 初回のみ
> # パッケージの読み込み
> library(vcd)
> # オッズ比の計算
> oddsratio(dat1, log = F)
[1] 1.833333
> # 信頼区間（下限値，上限値）の計算
> confint(oddsratio(dat1, log = F))
         lwr      upr
[1,] 0.522362 6.434448
> # p 値の計算
> summary(oddsratio(dat1))
     Log Odds Ratio Std. Error z value Pr(>|z|)
[1,]        0.60614    0.64059  0.9462    0.344
```

前述のように，検定結果として得られる *p* 値の大きさは，必ずしも効果量の大きさと一致しない。つまり，有意差があっても実質的な差が小さい場合もあれば，有意差がなくても実質的な差が大きい場合も考えられる。よって，有意差の有無にかかわらず，効果量と信頼区間を提示するべきである (Kline, 2004)。この点に関して，Field et al. (2012) は，研究論文でカイ二乗検定を用いる場合に，カイ二乗統計量 (X-squared)，自由度 (df)，*p* 値 (p-value) に加えて，オッズ比の値と信頼区間（下限値および上限値）を報告することを推奨している (p. 827)。

3.2 φ 係数

オッズ比と同様，カイ二乗検定の効果量としてよく用いられる指標として，φ 係数 (phi coefficient) とクラメールの *V* (Cramér's *V*) が挙げられる。そして，2×2 の分割表に対しては φ 係数，それよりも大きい分割表に対してはクラメールの *V* がそれぞれ用いら

れる (Field, 2009, p. 698)。

φ係数は，本質的にはピアソンの積率相関係数 (Pearson's product-moment correlation coefficient) の絶対値に等しく，0 から 1 までの値をとる。表 1 のデータからφ係数を求める場合は，以下のような計算を行う。

```
> # φ係数の計算
> (dat1[1, 1] * dat1[2, 2] - dat1[1, 2] * dat1[2, 1]) /
sqrt((dat1[1, 1] + dat1[1, 2]) * (dat1[2, 1] + dat1[2, 2]) *
(dat1[1, 1] + dat1[2, 1]) * (dat1[1, 2] + dat1[2, 2]))
[1] 0.1501879
```

上記の計算結果を見ると，φ係数が 0.1501879 であることが分かる。また，R の psych パッケージの phi 関数を使えば，φ係数を簡単に求めることができる。

```
> # パッケージのインストール
> install.packages("psych", dependencies = T)     # 初回のみ
> # パッケージの読み込み
> library(psych)
> # φ係数の計算
> phi(dat1, digit = 8)     # 引数 digits で有効桁数を指定
[1] 0.1501879
```

Cohen (1988) などでは，φ係数が 0.1 以上で「効果量小」，0.3 以上で「効果量中」，0.5 以上で「効果量大」であるとされている。ただし，効果量の大きさは，あくまで目安であり，その基準は研究分野によっても変わる。[9]

3.3 クラメールの *V*

クラメールの *V* は，基本的にはφ係数と同じものだが，2×2 よりも大きい分割表を扱えるように正規化した指標である。ここでは，3 つの異なる習熟度 (Level) を持つ学習者のデータから集計した正用 (Correct) と誤用 (Error) の頻度を分析対象とする（表 2)。

表 2

頻度差の検定のためのサンプル・データ (2)

	Level 1	Level 2	Level 3
Correct	805	414	226
Error	99	38	12

表 2 のデータを R に読み込んで，カイ二乗検定を実行するには，以下のような処理を行う。

```
> # データの入力
> dat3 <- matrix(c(805, 99, 414, 38, 226, 12), nrow = 2)
> colnames(dat3) <- c("Level 1", "Level 2", "Level 3")
> rownames(dat3) <- c("Correct", "Error")
> # 入力したデータの確認
> dat3
       Level 1 Level 2 Level 3
Correct     805     414     226
Error        99      38      12
> # カイ二乗検定
> chisq.test(dat3)

      Pearson's Chi-squared test

data:  dat3
X-squared = 8.4224, df = 2, p-value = 0.01483
```

上記のカイ二乗検定の結果を見ると，p 値は 0.01483 であり，3 段階の習熟度における正用と誤用の頻度の割合には，5%水準での有意差が見られた。

次に，表 2 のデータからクラメールの V を計算してみる。クラメールの V は，「カイ二乗統計量」÷（「行数と列数の少ない方から 1 を引いた数」×「サンプル・サイズ」）の平方根をとった数である。以下のスクリプトでは，`chisq.test(dat3)$statistic` が「カイ二乗統計量」，`min(dim(dat3)) - 1` が「行数と列数の少ない方から 1 を引いた数」，`sum(dat3)` が「サンプル・サイズ」にそれぞれ対応している。

```
> # クラメールのVの計算
> sqrt(as.numeric(chisq.test(dat3)$statistic) / ((min(dim(dat3)) -
1) * sum(dat3)))
[1] 0.07268964
```

また，vcd パッケージの assocstats 関数を使えば，クラメールの V を簡単に求めることができる。

```
> # クラメールのVの計算
> (V <- assocstats(dat3))     # vcd パッケージは前段で読み込み済み
                     X^2 df  P(> X^2)
Likelihood Ratio 9.2592  2 0.0097588
Pearson          8.4224  2 0.0148289

Phi-Coefficient   : 0.073
Contingency Coeff.: 0.072
Cramer's V        : 0.073
```

上記の assocstats 関数の実行結果を見ると，クラメールの V (Cramer's V) が 0.073 であり，カイ二乗統計量 (Pearson X^2) が 8.4224 であることが分かる。有効桁数の違いこそあれ，これらの数値は前段で求めた結果と等しい。因みに，Cohen (1988) などでは，ϕ 係数と同様，クラメールの V が 0.1 以上で「効果量小」，0.3 以上で「効果量中」，0.5 以上で「効果量大」であるとされている。そして，Gries (2013) では，研究論文でカイ二乗検定を用いる場合は，分析対象である頻度表，カイ二乗値，自由度，p 値に加えて，クラメールの V などの効果量を報告すべきであるとされている (p. 5821)。

最後に，クラメールの V の信頼区間を求めるには，以下のような計算を行う。[10]

```
> # パッケージのインストール
> install.packages("MBESS", dependencies = T)     # 初回のみ
> # パッケージの読み込み
> library(MBESS)
> # 非心度の95%信頼区間の計算
> (nc.chisq <- conf.limits.nc.chisq(Chi.Square = V$chisq_tests[2,
1], df = V$chisq_tests[2, 2], conf.level = .95))
$Lower.Limit
```

```
[1] 0.3010138

$Prob.Less.Lower
[1] 0.025

$Upper.Limit
[1] 22.34111

$Prob.Greater.Upper
[1] 0.025
> # クラメールの V の信頼区間（下限値）の計算
>   sqrt((V$chisq_tests[2,   2]   +   as.numeric(nc.chisq[1]))   /
((V$chisq_tests[2, 2] - 1) * sum(dat3)))
[1] 0.03799404
> # クラメールの V の信頼区間（上限値）の計算
>   sqrt((V$chisq_tests[2,   2]   +   as.numeric(nc.chisq[3]))   /
((V$chisq_tests[2, 2] - 1) * sum(dat3)))
[1] 0.1235737
```

上記の計算結果を見ると，クラメールの V の信頼区間の下限値が 0.03799404 であり，上限値が 0.1235737 であることが分かる。

4. おわりに

本稿では，頻度差の検定の効果量として，オッズ比，ϕ 係数，クラメールの V を紹介した。これらの指標を用いることで，サンプル・サイズなどの問題を回避できるだけでなく，実質的な頻度差の解釈が可能になる。従って，検定を行う場合は，単に p 値を確認するだけでなく，効果量を考慮に入れた解釈を行うべきである。また，研究論文に効果量を記載し，その計算に使用した分割表を提示することで，近年大きな注目を集めているメタ分析 (meta-analysis) に役立つ (Bourenstein et al., 2009)。[11] そして，人文科学の分野においても，心理学や応用言語学の論文では，何らかの効果量を報告する義務や慣例が急速に出来上がりつつある (e.g., American Psychological Association, 2009; Plonsky & Oswald, 2014)。このような流れを受けて，今後は，コーパス言語学の研究論文においても，効果量の報告が求められることが予想される。個々の研究論文で効果量を報告することにより，分析結果の適切な記述と解釈がなされることが望ましい。

謝辞

本稿を執筆するにあたって，水本篤先生（関西大学），浦野研先生（北海学園大学），田中省作先生（立命館大学）より，非常に多くの助言を頂きました。記して，ここに感謝の意を申し上げます。

注

1. 紙面の都合上，本稿では，「統計量」，「自由度」，「*p* 値」，「信頼区間」，「相関係数」といった基本的な用語の解説は行わない。これらの用語に関しては，南風原 (2002) などを参照。
2. カイ二乗検定ではなく，対数尤度比検定 (log-likelihood ratio test) やフィッシャーの正確検定 (Fisher's exact test) などが用いられることもある (Baker, Hardie, & McEnery, 2006, p. 31)。
3. R に関する詳細は，Kabacoff (2011) などを参照。同書は，R のインストール方法から丁寧に解説し，データを視覚化する方法，検定や相関といった基本的な統計手法から，多変量解析などの比較的高度な統計手法までを幅広く扱っている。
4. `chisq.test` 関数の引数 `correct` を `T` とすると，イェーツの連続補正 (Oakes, 1998, p. 25) を行ったカイ二乗検定が実行される。しかし，ここでは補正を行わない。
5. R スクリプトは省略するが，表 1 における 4 つの頻度を全て 100 倍にしたデータに対して，カイ二乗検定を行うと，0.1%水準での有意差が見られる。
6. 正確には，「Corpus A における Word Y の頻度に対する Word X の頻度の割合（比率）」を「Corpus B における Word Y の頻度に対する Word X の頻度の割合（比率）」で割ったものである。
7. オッズ比の計算方法には様々なものがあり，それらを計算するための R パッケージも複数存在する。詳しくは，奥村 (2015) などを参照。
8. Gries (2014) は，「オッズ比 0.5 とオッズ比 1.5 は，ともに 1 から 0.5 ずつ離れているために，等しい効果量の大きさである」と誤って解釈される可能性を指摘し，対数オッズ比 (logged odds ratio) を提示するという選択肢に言及している (p. 372)。なお，対数オッズ比とは，オッズ比の対数をとった値であり，0.5 と 1.5 の対数は，それぞれ約-0.7 と約 0.4 となる。
9. ただし，効果量に恣意的な閾値を設けることは，検定における有意水準の恣意性と本質的に同じものであるという考え方もできる。
10. クラメールの *V* の信頼区間を求めるには，非心度の 95%信頼区間を先に計算する必要がある。詳しくは，南風原 (2014) を参照。
11. 効果量の解釈に関して，「どれくらいの効果量であれば，どの程度の実質的な効果があるのか」という特定分野での基準は，メタ分析によって策定され得るものである (Plonsky & Oswald, 2014)。コーパスを用いた言語研究の場合，先行研究の論文中に頻度表（分割表）が記載されていることが多いため，それらのデータから効果量を計算することも可能である。

参考文献

American Psychological Association (2009). *Publication manual of the American Psychological Association* (6th ed.). Washington, DC: American Psychological Association.

Baker, P., Hardie, A., & McEnery, T. (2006). *A glossary of corpus linguistics.* Edinburgh: Edinburgh University Press.

Bourenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis.* Chichester: Wiley.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale: Lawrence Erlbaum.

Field, A. (2009). *Discovering statistics using SPSS* (3rd ed.). London: Sage.

Field, A., Miles, J., & Field, Z. (2012). *Discovering statistics using R.* London: Sage.

Gries, S. Th. (2013). Testing independent relationships. In Chapelle, C. A. (Ed.), *The encyclopedia of applied linguistics* (pp. 5817–5822). Oxford: Wiley-Blackwell.

Gries, S. Th. (2014). Tests, effect sizes, and explorations. In Glynn, D., & Robinson, J. A. (Eds.), *Corpus methods for semantics: Quantitative studies in polysemy and synonymy* (pp. 365–389). Amsterdam: John Benjamins.

南風原朝和 (2002). 『心理統計学の基礎—統合的理解のために』 有斐閣.

南風原朝和 (2014). 『続・心理統計学の基礎—総合的理解を広げ深める』 有斐閣.

Hommerberg, C., & Tottie, G. (2007). *Try to* or *try and*? Verb complementation in British and American English. *ICAME Journal, 31*, 45–64.

Kabacoff, R. I. (2011). *R in action: Data analysis and graphics with R.* New York: Manning.

Kline, R. B. (2004). *Beyond significance testing: Reforming data analysis methods in behavioral research.* Washington, DC: American Psychological Association.

Lorenz, G. (1998). Overstatement in advanced learners' writing: Stylistic aspects of adjective intensification. In Granger, S. (Eds.), *Learner English on computer* (pp. 53–79). London: Longman.

McEnery T., Xiao, R., & Tono, Y. (2006). *Corpus-based language studies: An advanced resource book.* London: Routledge.

Oakes, M. (1998). *Statistics for corpus linguistics.* Edinburgh: Edinburgh University Press.

奥村晴彦 (2015). 「2×2 の表，オッズ比，相対危険度」 http://oku.edu.mie-u.ac.jp/~okumura/stat/2by2.html [Last access: 15th February of 2015]

Plonsky, L., & Oswald, F. L. (2014). How big is "big"? Interpreting effect sizes in L2 research. *Language Learning, 64,* 878–912.