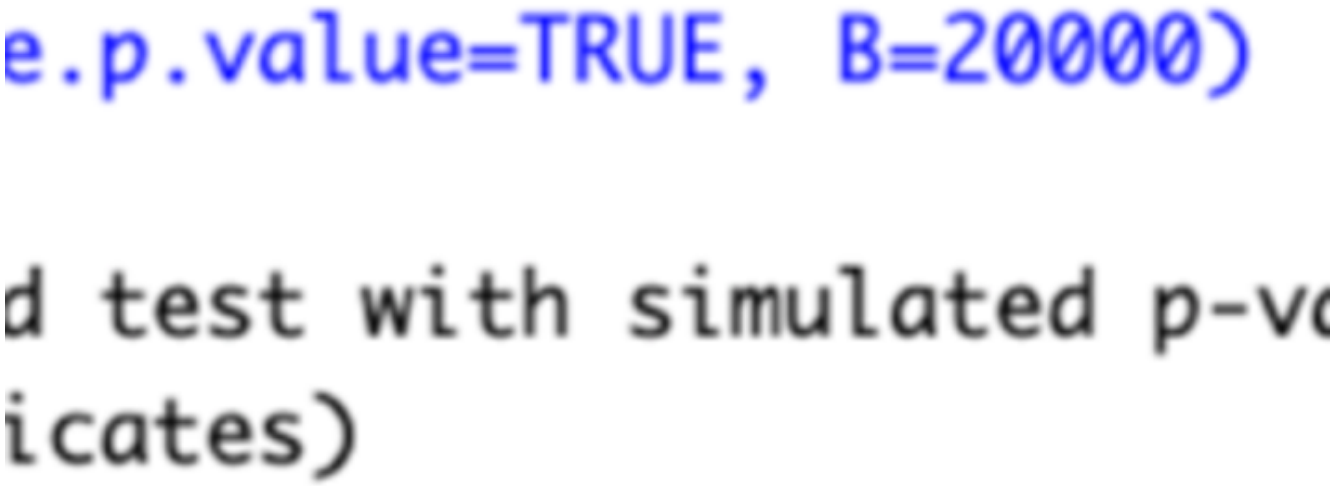


# 『データサイエンスのための統計学入門』と「say」

♡ 6

Yutaka MOTEKI
2023年1月22日 13:06

「カイ二乗検定：リサンプリング方式」の項目を読んでみた。

訳書129ページに、リサンプリングのアルゴリズムの説明(**注1**)の中で、1から3までのステップがあり、4番目に「２から３のステップを1,000回繰り返す」とある。なぜ1,000回なのかと疑問に思った。その回数にどのような意味があるのかと思って英文を確認してみると、**「say」という単語**が省略されて訳されてることがわかった。

「２から３のステップを、**たとえば**、1,000回繰り返す」ということであった。

> Repeat steps 2 and 3, **say**, 1,000 times.

Practical Statistics  for Data Scientists, 2nd edition, p.126.

訳書の次のページでは、1,000回ではなく、2,000回繰り返した例が掲載されている(**注2**)。つまり、

```
chisq.test(clicks, simulate.p.value=TRUE, B=20000)
```

を実行した結果である。（設定しなくても、デフォルトがB=20000であるようだ。）

```
> clicks
           Click No-click
Headline A    14      986
Headline B     8      992
Headline C    12      988
> chisq.test(clicks, simulate.p.value=TRUE, B=20000)

        Pearson's Chi-squared test with simulated p-value
        (based on 20000 replicates)

data:  clicks
X-squared = 1.6659, df = NA, p-value = 0.4856
```

130ページに記載の実行例と同じもの

Rでの出力結果を見ると、カイ２乗値(**注3**)が約1.67、p値が0.4856ということになっている。

[注]

(1) リサンプリングのアルゴリズムの説明がなされているのだが、訳書では、「このリサンプリングのアルゴリズムを次のように**検定**できる」となっている。アルゴリズムが検定の対象なのだろうか。そう考えると、日本語としてここで検定という言葉を使うのはおかしいと思う。また、次に出てくるのはアルゴリズム

の説明であり、「統計的検定」の計算を具体的にやっているわけではない。test という英語は統計学的な文脈の中であっても「検定」という日本語にいつも置き換えられるわけではない。

> We can **tes**t with this resampling algorithm:
>
> 1. Constitute a box with 34 ones (clicks) and 2,966 zeros (no clicks).
> 2. Shuffle, take three separate samples of 1,000, and count the clicks in each.
> 3. Find the squared differences between the shuffled counts and the expected counts and sum them.
> ...

Practical Statistics  for Data Scientists, 2nd edition, p.126.

(2)　「例えば、1,000回」となっていれば、実行例が2,000回になっていても、つじつまが合わないことにはならない。

(3) 訳書111ページに、「偶然（帰無仮説）の結果よりも5%極端」という表現がある。α（アルファ）についての項目の中である。この部分は、「5パーセントの確率で偶然に生じる結果よりも極端」(more extreme than **5% of the chance result**)という意味になっているだろうか。「偶然の結果**よりも５%**極端」という訳はおかしい。

**[補足]**

　他にもおかしいところがあった。「リサンプリング手続きは、クリック率が偶然によるよりは大きいことを検定できる」と訳されている部分があるのだが、これは、後から出てくる「検定は、結果がランダムでも容易に得られることを示す」という部分と矛盾する。実際、p値は0.4853という計算結果が表示されている。

　ここでは、**whetherという単語**を無視して訳してしまっていることがわかった。つまり、「リサンプリング手続きは、クリック率が偶然によるよりは大きい**かどうか**を検定できる」と訳されるべきであった。表現として、「ＡがＢであること」と「ＡがＢであるかどうか」とは、使い分けるべきだろうと思う。そうでないと「検定できる」という表現の意味が不明瞭になる。

> A resampling procedure can test **whether** the click rates differ to an extent greater than chance might cause.

Practical Statistics  for Data Scientists, 2nd edition, p.125.