

多変量解析 <初歩の初歩>に入門

連関の分析; 仮説検定の基礎の基礎

なかもとけいこ
kenakamoto@nifty.com

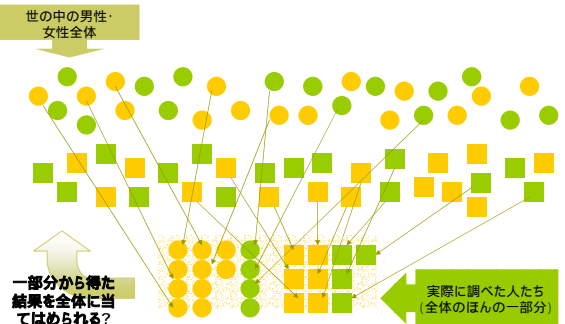
今回の目的

- 今回は多変量解析の勉強ではないです。いわゆる推測統計学の基礎の基礎。
- 共起関係の有無を調べる手法として、連関の検定を勉強する
- 統計的仮説検定について勉強する
- ついでにExcelのピボットテーブルの使い方を覚える

たとえば次のような場面

- ケータイの機種
 - Aさん: 知り合いの男の子10人と女の子14人にどこのケータイを使っているか聞いてみたら, docomoを使っているのは, 男の子のうちでは6人, 女の子のうちでは10人だった。
 - Bさん: …
 - Aさん: パーセントに直すと, 男の子は60%, 女の子は71%. だから, (一般的にいうと,) 女の子の方がdocomoを使っている人は多いと思う。
 - Bさん: …なんか納得できん。たまたまちゃうん?

たまたまじゃないの? の意味



偶然を利用して仮説を検証する

- つまり, 「偶然選んだサンプルにだけ当てはまるのではなく, 全体(母集団)にも結果を一般化できる」という風にいえばよい。
- その前提として, 母集団から標本が偏りなく(ランダムに)抽出されていることが必要。
 - 自分の仮説に都合のいい例ばかりを引っ張ってきて, 「全体」については何も言えない。
- その上で, 標本から得られた結果が, (仮説が成り立たない場合には) 非常に低い確率でしか得られないことを示せばよい。
 - この確率を求めるために利用されるのが, 様々な確率分布。
 - 基本的に偉い先生達が求めてくれているので, それに当てはめて計算するだけでよい。
- 仮説検定というのは要するにこういうことです。

仮説検定の考え方

- 確かめたい仮説(H1)を立てる
- 反対の仮説(帰無仮説; H0)を立てる
- 母集団(知りたい対象全体)を想定する。
- 母集団から標本を抽出する。
 - 母集団全体からまんべんなく標本を採れるようにする。
 - 基本は無作為抽出
- 標本について関連する特徴を測定する。
- もしも帰無仮説(H0)が正しかったという前提を置いたときに, 測定した結果がどれくらいの確率で得られるかを計算する。
- その確率がうんと小さければ(例えば, 5%以下であれば), 「偶然で済ませるには珍しすぎる」と考えてみる。
- 帰無仮説(H0)を棄却する(偶然でないならば, 前提が間違ってるんだらう)。
- H1(対立仮説)を採択する。

仮説検定の考え方:例

- (H1) 男の子よりも女の子の方がdocomoを持っている割合が多いに違いない
- (H0) 男の子と女の子ではdocomoを持っている割合は変わらない
- 母集団を[ケータイを所有している日本人の成人男女全体]と想定する.
- 母集団から標本を抽出する.
 - たとえば、電話帳と乱数表を利用して、無作為に調査対象者を選ぶ、とか.
 - でも、実際は無作為抽出は難しく、特定の標本に偏ることになりがち → 後述.
- 標本について関連する特徴を測定 (性別と所有している携帯の機種)
- H0が正しいという前提のもとで、標本について得た値(男性10人中6人, 女性14人中10人)がどれくらいの確率で得られるかを計算する.
 - この場合、 χ^2 (カイ二乗と読む)値というものを求めると、0.34.
 - χ^2 分布を調べると、男女でdocomoの割合に差がなかったとしても、55%くらいの確率で標本にこれくらいの差が現れることが分かる.
- H0は棄却できない.
- なので、H1を採択できない. → 有意(味)significant な差がない.

連関を(うまく)調べる必要性

- 例えば次のような場合...
 - 女性の発話を100個集めてきて、終助詞の有無を調べたところ、100個中80個で終助詞が用いられていることが分かった.
 - ここから、女性は終助詞をよく用いるということを結論.
 - (必ずしも論理的に正しくはないが)この結論が含意すること
女性は終助詞をよく使うけれども、男性はあまり使わない(あるいは、男性はよくは使わない).
 - つまり、性別と終助詞の使用には連関性がある.
- はたして、このデータからそう結論できるか?

それはちょっとムリ...

- なぜなら、右のような場合もあろうから

- 男性も女性も発話の約8割に終助詞を使用している.
- 性別に無関連に終助詞はよく使われると結論するのが妥当.
対立カテゴリーについても調べないと連関には言及できない.

- でも、男性の方が(ちょっとだけ)、終助詞を使ってる発話が多いよ?

- たまたま、かどうか確認する必要がある.
- 検定の出番!

	終助詞あり	終助詞なし	合計
男性	80 (80%)	20 (20%)	100 (100%)
女性	45 (75%)	15 (25%)	60 (100%)
合計	125	35	160

こういう表をクロス集計表と呼ぶ

クロス集計表上での連関の検討(1)

- 完全に連関がない=独立のとき

- クロス集計表の周辺度数から、中身のセルを計算することができる.
- つまり、どの行、どの列をとっても比率が一定になっている状態.
- この状態を、連関がない場合(H0が成り立つ場合)の期待値と見なせる
- 実際に観測された値がこの期待値から有意にずれているか調べれば、二つの変数が独立かどうか検定することができる.

	終助詞あり	終助詞なし	合計
男性	125X(100/160) = 78.125	35X(100/160) = 21.875	100 (100/160=62.5%)
女性	125X(60/160) = 46.875	35X(60/160) = 13.125	60 (60/160=37.5%)
合計	125	35	160

クロス集計表上での連関の検討(2)

- カイ2乗検定 (χ^2 検定)

- 期待値と実測値の乖離の程度が有意なものかどうかを検討するための検定.

$$\chi^2 = \sum_i \sum_j \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

O_{ij} : 観測値

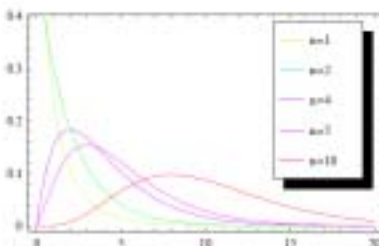
E_{ij} : 期待値

- 帰無仮説 (H0: 二つの変数は独立である)が正しい場合には、 χ^2 値は、自由度 (n-1) × (m-1) の χ^2 分布に従う.

クロス集計表上での連関の検討(3)

- χ^2 分布.

χ^2 分布の確率密度関数



$$\chi^2_m = \sum_{i=1}^m Z_i^2 = \sum_{i=1}^m \frac{(X_i - \mu)^2}{\sigma^2}$$

$$f(x) = \frac{1}{2^{\frac{m}{2}} \Gamma(\frac{m}{2})} x^{\frac{m}{2}-1} e^{-\frac{x}{2}}$$

ただし、

$$\Gamma(m) = \int_0^{\infty} e^{-x} x^{m-1} dx \quad (\text{ガンマ関数})$$

自由度が大きくなると、正規分布に近づく

クロス集計表上での連関の検討(4)

自由度 degree of freedom (df)

- 自由に変動できる測定値の個数
- 実験や調査のデザイン, 検定する仮説で決まる数値.
 - 得られるデータの内容的特徴を表現するのではなく, 形式的特徴を表す.

	終助詞あり	終助詞なし	合計
男性	80	20	100
女性	45	15	60
合計	125	35	160

たとえば、
先ほどの例だと...

クロス集計表上での連関の検討(5)

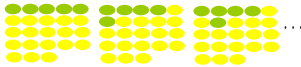
χ²検定が適用できない場合

- セルの中に0に近い値がある
- 周辺度数に10以下程度の小さな値がある
 - χ²分布への当てはまりが悪くなるため, χ²検定を適用することは望ましくない.
- 周辺度数を固定した上で, 各セルの度数にある特定のパターンが得られる場合の数を数え上げ, それぞれの確率を求めれば, 実際に観測されたパターンが生じる確率を求められるはず.
 - ようするに, 理論的な分布とかを考えずに, クロス集計表に現れているパターンの出現確率を直接求める.
 - これが, Fisherの直接(確率)法: Fisher's exact test

クロス集計表上での連関の検討(6)

Fisherの直接法

23人中5人が終助詞を使い, 18人が使わないという組み合わせ.



$${}_{23}C_5 = 33649$$

この内1人が男性, 4人が女性である組み合わせ

$${}_{13}C_1 \times {}_{10}C_4 = 13 \times 210 = 2730$$

したがって表のようなデータが得られる確率は,

$$P = \frac{{}_{13}C_1 \times {}_{10}C_4}{{}_{23}C_5} = \frac{2730}{33649} = 0.0811$$

この値が設定した有意水準以下なら帰無仮説を棄却

クロス集計表上での連関の検討(7)

Fisherの直接法

- 一般化して書くと...

$$P = \frac{n_1! n_2! n_3! n_4!}{N! n_{11}! n_{12}! n_{21}! n_{22}!}$$

	カテゴリー1	カテゴリー2	合計
条件1	n_{11}	n_{12}	$n_{1\cdot}$
条件2	n_{21}	n_{22}	$n_{2\cdot}$
合計	$n_{\cdot 1}$	$n_{\cdot 2}$	N

- 観測されたパターンが生じる確率と観測されたパターン以上に対立仮説を支持するパターンの確率を求め, それらの確率の総和(すなわち, P)が, 設定した有意水準以下なら帰無仮説を棄却する.

言語学研究での適用例

Stefanowitch (To appear?)より無断転載

Table 1: The distribution of *ryu* in the (conter-bases) of NP (HSC-1)

	<i>ryu</i>	no	Other items	Row total
in the center of NP	27	108	3,879	4,014
in the base of NP	45	108	468	621
Column totals	72	216	4,347	

If we submit these figures to a distributional statistic such as the Fisher exact test or the Chi-square test, we find that the *p*-value is exceptionally small (Fisher exact: $p=0.14E-08$, Chi-square: $\chi^2=32.84$ $df=1$, $p=1.69E-06$). This tells us that the noun *ryu* is indeed highly distinctive for the use of the two patterns, but it does not tell us for which one, since any distributional statistic can determine whether the observed frequencies in one or more cells of a table deviate significantly from the expected frequencies, but cannot determine the direction of deviation. Thus, in order to determine which of the two

Excelを利用したクロス集計表の作成

代表的なExcelでのコーディング形式からピボットテーブルを利用して作成

→ やってきましょう。

連関の大きさの検討

- ここまで話してきた χ^2 検定やFisherの直接法は、期待値から観測値が有意にずれているかを検定する方法。
 - つまり、独立性からの有意なずれが認められるかで連関の有無を検討している。
 - 連関の大きさ(あるいは強さ)を考えるには向いていない。
- 連続量での相関係数のように連関の強さを数値化するには、各種の連関係数を用いるとよい。
 - 四分点相関係数() -1 1
 - Yuleの連関係数(Q) -1 Q 1
 - クラメールの連関係数(V) 0 V 1 など

ランダム・サンプリングについて

- 心理学でも言語学でも実際に手にできるデータは「まず母集団を想定」→「標本をランダムにサンプリング(無作為抽出)」という手続きをとることは非常に困難。
 - 統計的推測のための前提条件が満たされていない。
 - 一つの方法は得られたサンプルから結論を導くのに「実際のサンプルからその結果を一般化しても無理がないと思われる母集団」を新たに想定すること(南風原, 2002)。
 - 誰もが納得する客観的な基準はなく、研究者それぞれが主観的に行うしかない。
 - その上で、類似の研究をさまざまなサンプルに対して行うことで、どこまで適用可能で、どこから不可能かを見極めていく作業が必要になる。

連関の分析のまとめ

- 今回は次のことを学びました。
 - 統計的仮説検定の基本的な考え方
 - 慣れない内はちょっとひねくれたやり方に思えるかも。
 - 連関の分析をするために必要なこと
 - クロス集計表の作成
 - 2乗値の求め方
 - Fisherの直接法
 - ピボットテーブルを使ったクロス集計表の作り方
- 仮説検定のための統計量は検定対象となる数値の性質(2つの平均の差、3つ以上の平均の差、相関係数の大きさなど)によって異なります。必要が生じたら、自分で本を読んだりして勉強しましょう。