

1. カテゴリカルデータの分析

仮説検定に関する基本的なことを学んだうえで、今度は心理学でよく使われる検定について見ていきましょう。では、どこから始めればよいのでしょうか。全ての教科書がスタート地点に関する合意を持つわけではないのですが、ここでは“ χ^2 検定”（この章では、“カイ二乗 (にじょう)chi-square” と発音します^{*1}）と“t-検定”（Chapter ??）から始めます。これらの検定は科学的実践において頻繁に使用されており、“回帰”（Chapter ??）や“分散分析”（Chapter ??）ほど強力ではないのですがそれらよりはるかに理解しやすいものとなっています。

“カテゴリカルデータ”という用語は“名義尺度データ”の別名に過ぎません。説明していないのではなく、ただデータ分析の文脈では、“名義尺度データ”よりも“カテゴリカルデータ”という言葉を使う傾向があるのです。なぜかは知りません。なににせよ、**カテゴリカルデータの分析** はあなたのデータが名義尺度の際に適用可能なツールの集合を指示しています。しかし、カテゴリカルデータの分析に使用できるツールには様々なものがあり、本章では一般的なツールの一部のみを取り上げます。

1.1

The χ^2 (カイ二乗) 適合度検定

χ^2 適合度検定は、最も古い仮説検定の一つです。この検定は世紀の変わり目に Karl Pearson 氏が考案したもので（Pearson1900）、Ronald Fisher 氏によっていくつかの修正が加えられました（Fisher1922）。名義尺度変数に関する観測度数分布が期待度数分布と合致するかどうかを調べます。例えば、ある患者グループが実験的処置を受けており、彼・彼女らの状況が改善されたか、変化がないか、悪化したかを確認するために健康状態が評価されたとします。各カテゴリー（改善、変化なし、悪化）の数値が、標準的な処置条件で期待される数値と一致するかどうかを判断するために、適

^{*1}また“カイ二乗 (にじょう)chi-squared”とも呼ばれる

合性検定は適用できます。もう少し、心理学を交えて考えてみましょう。

1.1.1 カードデータ

何年にもわたる多くの研究が、人が完全にランダムにふるまおうとすることの難しさを示しています。ランダムに「行動」しようとしても、我々はパターンや構造に基づいて考えてしまいます。そのため、「ランダムになにかをしてください」と言われたとしても人々が実際に行うことはランダムなものにはなりません。結果として、人のランダム性 (あるいは非ランダム性) に関する研究は、我々が世界をどのように捉えているのかについての深遠な心理学的問いを数多く投げかけます。このことを念頭に置いて、非常に簡単な研究について考えてみましょう。シャッフルされたカードのデッキを想像して、このデッキの中から「ランダムに」一枚のカードを頭の中で選ぶようお願いしたとします。一枚目のカードを選んだ後、二枚目のカードを心の中で選択してもらいます。二つの選択に関して、注目するのは選ばれたカードのマーク (ハート、クラブ、スペード、ダイヤモンド) です。これをたとえば $N = 200$ にやってもらうよう依頼した後、選択されようとしたカードが本当にランダムに選ばれているかどうかをデータを確認して調べてみましょう。データは `randomness.csv` に入っており、JASP で開くと 3 つの変数が表示されるでしょう。変数 `id` は各参加者に対する一意識別子であり、二つの変数 `choice_1` と `choice_2` は参加者が選択したカードのマークを意味しています。

今回は、参加者の選んだ最初の選択肢に注目してみましょう。‘Descriptives’ - ‘Descriptive Statistics’ の下にある [Frequency tables](#) オプションを選択して、選択された各マークの数をカウントしてみましょう。以下が得られたものです:

クラブ	ダイヤモンド	ハート	スペード
35	51	64	50

この小さな度数分布表はとても有益です。この表を見れば、人はクラブよりもハートを選びやすいかもしれないというわずかなヒントを得られますが、それが実際にそうであるのか偶然の賜物であるのかどうかは見るだけでは明らかではありません。なので、それを知るためにはなんらかの統計分析をしなければならないでしょう。それが、次のセクションでお話しすることになります。

よろしい。ここからは、先ほどの表を分析対象のデータとして扱います。しかしながら、このデータについて数学的に語らなければならないために、表記の意味について明確にしておくことは大事でしょう。数学的表記では、人が読める単語である "observed (観測された)" を文字 O に短縮して、観測位置を示すために下付き文字を使用します。なので、この表における二番目の観測変数は数学では O_2 として記述します。日本語表記と数学記号の関係を以下に示します:

ラベル	インデックス, i	数学, シンボル	数値
クラブ, ♣	1	O_1	35
ダイヤモンド, ◇	2	O_2	51
ハート, ♥	3	O_3	64
スペード, ♠	4	O_4	50

これではっきりしたでしょう。また、数学者は特定の事柄よりも一般的な事柄について話したがるので、 O_i という表記が見られるでしょう。これは、 i 番目のカテゴリーに属する観測変数を意味します (i は 1、2、3、4 のいずれか)。最後に、観測された頻度数に言及したい場合、統計家は観測値をベクトル^{*2}に分類します。これは、太字を使用して \mathbf{bmO} とします。

$$\mathbf{O} = (O_1, O_2, O_3, O_4)$$

繰り返しますが、これは新しいものでも興味深いものでもありません。ただの表記です。 $\mathbf{O} = (35, 51, 64, 50)$ ということ、私がしているのは観測された度数表の記述 (i.e., **observed**) ですが、数学表記を用いてそれを参照します。

1.1.2 帰無仮説と対立仮説

先ほどのセクションで指摘したように、我々の研究仮説は「人はカードをランダムに選択しない」です。これから行いたいことはこれを統計的仮説に変換してから、それらの仮説に関する統計検定を構築することです。説明予定のテストは**ピアソンの χ^2 (カイ二乗) 適合度検定**であり、よくあることですが、まずは帰無仮説の注意深い構築から始めなければなりません。今回はかなり簡単です。まず、帰無仮説を言葉にしてみましょう：

H_0 : 4 つ全てのマークは同じ確率で選択される

さて、これは統計学なので、同じことを数学っぽく言えなければなりません。これをするために、 j 番目のマークが選ばれる場合の真の確率を参照するときには表記 P_j を用いましょう。もし帰無仮説が真であれば、4 つのマークがそれぞれ 25% の確率で選択されます。言い換えれば、帰無仮説は $P_1 = .25, P_2 = .25, P_3 = .25$ そして $P_4 = .25$ としたものです。ただし、観測された頻度数をデータ全体の要約ベクトル \mathbf{O} として分類するように、帰無仮説と対応する確率として \mathbf{P} を用います。そのため、帰無仮説を記述する確率の集合を $\mathbf{P} = (P_1, P_2, P_3, P_4)$ とすると、以下のようになります：

H_0 : $\mathbf{P} = (.25, .25, .25, .25)$

この例では、帰無仮説は全ての確率が互いに等しい確率のベクトル \mathbf{P} と対応します。しかし、常に

^{*2}ベクトルは同じ基本型のデータ要素のシーケンスです