0.1

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。

0.1.1 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に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、ダイアモンド) 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	ダイアモンド	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 it 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
ダイアモンド, ♢	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[*1]に分類します。これは、太字を使用して bmO とします。

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

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。$\boldsymbol{O} = (35, 51, 64, 50)$ ということで、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。

0.1.2 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

[*1] ベクトルは同じ基本型のデータ要素のシーケンスです

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 $\boldsymbol{O}$ として分類するように、帰無仮説と対応する確率として $\boldsymbol{P}$ を用います。そのため、帰無仮説を記述する確率の集合を $\boldsymbol{P} = (P_1, P_2, P_3, P_4)$ とすると、以下のようになります:

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

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 $\boldsymbol{P}$ と対応します。しかし、常にそうである必要はありません。例えば、もし実験課題で他のマークの 2 倍クラブが含まれているデッキを想像してもらう場合には、帰無仮説は $\boldsymbol{P} = (.4, .2, .2, .2)$ となるでしょう．確率がすべて正の値であり、その合計が 1 である限りは、それは帰無仮説として正当な選択です。ですが、適合度検定では一般的に全てのカテゴリーが同様の確率である帰無仮説を用います。そのため、ここではそれに固執します。

対立仮説 H_1 はどうでしょうか? 我々の関心は、関係する確率が全て同じではないこと (つまり、人々の選択が完全にランダムではなかったこと) を実証することです。その結果、「人にやさしい (負担の小さい)」バージョンの仮説はこんな感じです：

H_0: 4 つのマークが同じ確率で選択される
H_1: 少なくとも 1 つのマークの選択確率が 0.25 *ではない*

そして「数学者にやさしい」バージョンはこうです:

H_0: $\boldsymbol{P} = (.25, .25, .25, .25)$
H_1: $\boldsymbol{P} \neq (.25, .25, .25, .25)$

0.1.3 “適合度” 検定の統計量

この段階で、観測された頻度数 $\boldsymbol{O}$ と検定予定の帰無仮説と対応する確率の集合である $\boldsymbol{P}$ を我々は手にしています。ここでしたいことは、帰無仮説検定の構築です。いつものように、H_1 に対して H_0 を検定したい場合には、検定統計量が必要です。適合度検定の基本的なトリックは、データが帰無仮説にどれだけ「近いのか」を測定する検定統計量を組み立てることです。もし帰無仮説が真であるときの「期待値」がデータと似ていなければ、帰無仮説は真ではないでしょう。オーケイ、帰無仮説が真であるならどうなるだろう？ 正しい言い方をすれば、「期待度数」とは何かということです。$N = 200$ の観測データがあり、(もし帰無仮説が真であれば) ハートの選択確率は $P_3 = .25$ で、ハートの期待値は $200 \times .25 = 50$ ですよね? より具体的には、もし E_i が、「帰無仮説が真であるときにカテゴリー i の選択数の期待値」とすると次のようになります。

$$E_i = N \times P_i$$

この計算はとても簡単です。もし 4 つのカテゴリーに分類されうる 200 個の観測データがあって、カテゴリー全ての選択確率が同じだとすれば、各カテゴリーの観測データは 50 であると期待されますよね？

さて、これをどのように検定統計量に変換するのでしょうか？ 明らかに、我々のしたいことは各カテゴリーの*期待*値 (E_i) と*観測*値 (O_i) の比較です。そしてこの比較に基づいて、我々は良い検定統計量を導き出すことができるはずです。まずは、帰無仮説が期待した結果と我々が実際に得られた結果との差を計算しましょう。つまり、「観測値から期待値を引いた」差得点である $O_i - E_i$ を計算します。これを図示すると次の表のようになります。

		♣	♢	♡	♠
期待度数	E_i	50	50	50	50
観測度数	O_i	35	51	64	50
差得点	$O_i - E_i$	-15	1	14	0

つまり我々の計算によって、帰無仮説の予測よりも人はハートを多く、クラブを少なく選択していることがわかりました。しかし、ちょっと考えてみると、この素朴な違いは、私たちが求めているものとはちょっと違うようです。直観的に、帰無仮説の予測が少なすぎた場合 (ハート) も予測が多すぎた場合 (クラブ) も同程度によくないことに感じられます。なので、クラブではマイナスでハートではプラスだというのはちょっと変な感じです。これを解決する一つの簡単な方法は全てを二乗することで、ここでは二乗された差を計算します $(O_i - E_i)^2$。前回同様、これは手作業でできます：

		♣	♢	♡	♠
期待度数	E_i	50	50	50	50
観測度数	O_i	35	51	64	50
差得点	$O_i - E_i$	-15	1	14	0
二乗差	$(O_i - E_i)^2$	225	1	196	0

さぁ、これで一歩前進です。今手にしているものは、帰無仮説が悪い予測をしたときには大きく (クラブとハート)、良い予測をしたときには小さくなる数の集合です。次は、後述予定の技術的理由により、それらの数を期待度数 E_i で割って、**調整された** 二乗値 $\frac{(E_i - O_i)^2}{E_i}$ を計算しています。今回の例では全てのカテゴリーで $E_i = 50$ となるので、あまり面白い計算ではないですが、とりあえずやってみましょう：

		♣	♢	♡	♠
期待度数	E_i	50	50	50	50
観測度数	O_i	35	51	64	50
差得点	$O_i - E_i$	-15	1	14	0
二乗差	$(O_i - E_i)^2$	225	1	196	0
調整済み二乗差	$(O_i - E_i)^2/E_i$	4.50	0.02	3.92	0.00

要するに、ここで得たのは 4 つの「エラー」得点で、それぞれが観測度数に対する帰無仮説の予測から生じた「間違い」の大きさを示しています。そして、これを一つの有用な検定統計量に変換するためには、それらの数を単に足し合わせることが一つのやり方です。その結果を **適合度:goodness-of-fit** とよび、慣習的には χ^2 (カイ二乗) または (頭文字をとって)GOF ともよばれています。$4.50 + 0.02 + 3.92 + 0.00 = 8.44$ として計算可能です。

k をカテゴリーの総数だとすれば (i.e., カードデータの例だと $k = 4$)、χ^2 統計量は以下のように与えられます：

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

直観的に、もし χ^2 が小さければ観測データ O_i が帰無仮説の予測する E_i に近づき、帰無仮説を棄却するためには大きな χ^2 が必要となるはずです。

計算の結果、カードデータセットでは $\chi^2 = 8.44$ という値が得られました。次の疑問はこれが帰無仮説を棄却するのに十分な値なのか、ということです。

0.1.4 適合度の標本分布

χ^2 の値が帰無仮説を棄却するほどに大きいかどうかを決めるために、帰無仮説が真である場合の χ^2 に関する標本分布はどうなるかを理解する必要があるでしょう。ということで、今回のセクションではそれをやっていきます。この標本分布がどのように構成されるかの詳細をお見せして、次のセクションではそれを仮説検定の構築に用います。さて本題。標本分布が $k - 1$ の自由度を持つ **χ^2 (カイ二乗) 分布**であると喜んで受け入れる人は、このセクションの残りをスキップできます。しかし、もし適合度検定の *仕組み* を理解したいひとは、ぜひこの先をお読みください。

よし、帰無仮説が実際に真であると仮定しましょう。もしそうであれば、ある観測変数が i 番目のカテゴリーに属する真の確率は P_i となります。まぁ結局、それはほぼほぼ帰無仮説の定義です。これが意味するところについて考えます。これは、「自然」が、重み付きのコイン (i.e., 表が出る確率は P_j) を裏返すことで観測値がカテゴリー i に含まれるかどうかを決定するというようなものです。

したがって、自然がこれらのコインの N を反転させること（データセット内の観測ごとに 1 つ）を想像して、正確に O_i を頭に浮かべることで、観測された頻度 O_i を考えることができます。明らかに、これは実験を考える上ではかなり奇妙なやり方です。ですが、このシナリオには見覚えがありますね（と私は期待してますよ)。セクション **??**、二項分布のケースとまったく同じ設定です。言い換えれば、帰無仮説が真であれば、観測された度数は二項分布のサンプリングによって生成されたことになります。

$$O_i \sim \text{Binomial}(P_i, N)$$

中心極限定理 (Section **??**) の説明を思い出すと、特に N が大きく P_i が 0 または 1 に近 *すぎ* ない時に、二項分布は正規分布と近似して見えるようになります。いいかえれば、$N \times P_i$ が十分に大きければいいのです。また、別の言い方をすれば、期待度数 E_i が十分に大きい場合 O_i の理論的な分布は近似的に正規分布となります。さらにいえば、O_i が正規分布していれば、そのとき $(O_i - E_i)/\sqrt{E_i}$ も正規分布します。E_i は固定の値なので、E_i を引いて $\sqrt{E_i}$ で割ることで正規分布の平均と標準偏差が変化しますが、それだけです。では早速、適合度統計量*とはなにか*について見ていきましょう。今しているのは正規分布するものをたくさん集めて、二乗して、それからそれらを足し合わせているのです。おっと。これも見たことありますね！ セクション **??**でお話ししたように、標準正規分布 (i.e., 平均 0, 標準偏差 1) を持つものをたくさん集めて二乗してから足し合わせると、その結果はカイ二乗分布となります。これで適合度統計量の標本分布がカイ二乗分布であることを帰無仮説が予測している、ということがわかりました。イイね。

最後にもう一つ、いわゆる自由度についてお話しときましょう。セクション **??**を思い返せば、足し合わせるものの数は k、結果として生じるカイ二乗分布の自由度は k になると言いましたね. しかし、このセクションの冒頭で述べたのはカイ二乗適合度検定の自由度は $k-1$ であるということです。どうしたというのでしょう？ ここでの答えは、私たちが注目しているのは、純粋に *独立した* ものが同時に足し合わされている数だということです。また、次のセクションでお話ししますが、たとえ k 個分追加したとしても真に独立しているのは $k-1$ 個のみであり、自由度は $k-1$ だけです. それが次のセクションでの話題です。[*2]

0.1.5 自由度

セクション **??** でカイ二乗分布を紹介したときに、「**自由度**」が実際に *意味* するところは少し曖昧でした. 明らかに、そこは重要な点です。Figure **??**を見ると、自由度を変化させるとカイ二乗分布の形がかなり大きく変わっています。しかし、それはなんなのでしょうか？ 分布を紹介して正規分布との関係性を説明したときに、ある答えを提供しました：私が二乗して足し合わせた「正規分布する」数です。ですが、多くの人にとって、それは抽象的でちっとも参考になりません。ここで本当に

[*2] もし適合度統計量の式を $k-1$ の独立したものの数の和として書き直すと、$k-1$ の自由度を持つカイ二乗の「適切な」標本分布を得られます。その数学の詳細を示すことは入門書の範囲を逸脱しています。ここでしたいのは、なぜ適合度統計量がカイ二乗分布と関連しているのかについて理解してもらうことです。

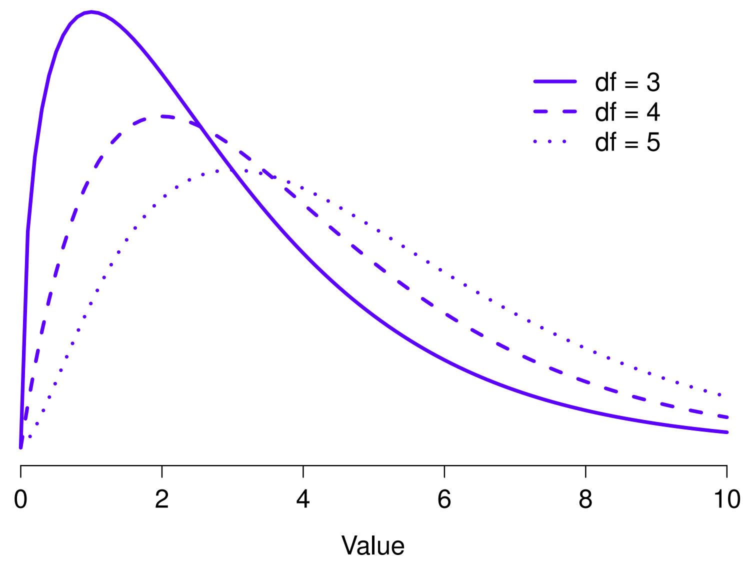

Figure1 「自由度」の異なる χ^2 (カイ二乗) 分布。

目指すべきなのは、我々の持つデータを用いて自由度を理解することです。ではいってみましょう。

自由度の基本的な考え方はとてもシンプルです。データの記述に使用する明確な「量」を数え上げることで計算して、それらのデータが満たさなければならない「制約」をすべて引き算します。[*3] これでは少し曖昧なので、具体的な例としてのカードデータを使いましょう。4 カテゴリー (ハート、クラブ、ダイアモンド、スペード) の観測度数に対応する O_1、O_2、O_3、O_4 の 4 つを用いてデータを記述します。それら 4 つの数が今回の実験での *ランダムな結果* です。しかし、実験には固定の制約が組み込まれています：サンプルサイズが N。[*4]つまり、ハートを選んだ人の数、ダイアモンドを選んだ人の数、クラブを選んだ人の数がわかれば、スペードを選んだ数は正確に把握できます。言い換えれば、4 つの数を用いてデータが記述されますが、実際には $4-1=3$ の自由度にしか対応していません。ちょっと違う考え方は、関心のある 4 つの*確率*があること (ここでも 4 つのカテゴリーに

[*3] これは単純すぎると指摘せざるを得ないとは思っています。大体はその説明でうまくいくんですが、整数ではない自由度の値に出くわすこともあります。気にしすぎることはありません；そんなときは「自由度」が少し厄介な概念であり、私がここでしているような単純なストーリーがすべてではないことを思い返してください。入門編では単純なストーリーに固執するのがいいんですが、このストーリーが崩壊することは事前に警告しておくのがベストだと思います。もし警告がなければ $df=3.4$ みたいなのを目にしたときに混乱してしまい、(正確に) 私が教えなかったことに気付くのではなく (不正確に) 私が教えたことを誤解したんじゃないかと考えてしまいます。

[*4] 実際問題として、サンプルサイズは常に固定されていません。例えば、一定期間にわたって実験をする際に参加者数は参加する人数に依存しているかもしれません。まぁそれは今の目的には関係ないです。

対応します) に注意することですが、それらの確率の合計値は 1 でなければならないという制約が課されます。そうなると自由度は $4 - 1 = 3$ になります。観測度数の観点から考えたいのか、確率の観点から考えたいのかに関係なく、答えは同じです。一般的に、k グループを含む実験で χ^2 (カイ二乗) 適合度検定を実行すると、自由度は $k - 1$ になります。

0.1.6　帰無仮説検定

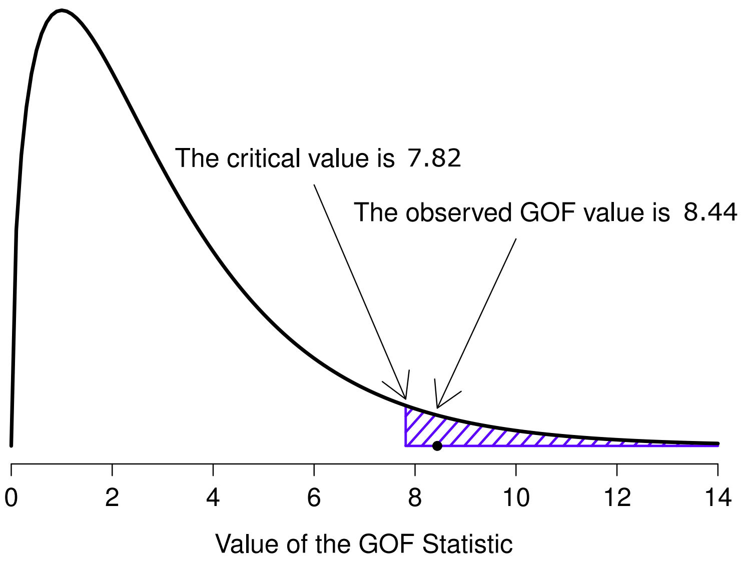

Figure2　χ^2 (カイ二乗) 適合度検定で仮説検定がどのように機能するのかを示す図。

仮説検定を構築するプロセスの最終段階は、棄却域とは何なのかを理解することです。いわば、χ^2 の値によって帰無仮説が棄却されます。以前のように、χ^2 の値が大きいということは帰無仮説が実験データの予測を行うのに不十分であったことを意味します。一方で χ^2 の値が小さいということは帰無仮説が支持されていることを意味します。したがって、χ^2 が棄却値よりも大きいければ帰無仮説を棄却し、χ^2 が棄却値よりも小さければ帰無仮説を保持する、というのは非常に賢明なやり方です。チャプター**??**で紹介した言語を使えば、カイ二乗適合度検定は常に **片側検定** です。そのとおり。あとはこの重要な値が何であるかを考えればいいわけです。そしてそれはとても簡単です。もし有意水準を $\alpha = .05$ と設定して (すなわち、Type1 エラーを 5% で許容する) 検定を行いたい場合、帰無仮説が真である際に χ^2 がその値を超える可能性が 5% になるように、棄却値を選択する必要があります。これを示したのが図 **??**です。

Degrees of Freedom	Probability								
	0.95	0.90	0.70	0.50	0.30	0.10	0.05	0.01	0.001
1	0.004	0.016	0.148	0.455	1.074	2.706	3.841	6.635	10.828
2	0.103	0.211	0.713	1.386	2.408	4.605	5.991	9.210	13.816
3	0.352	0.584	1.424	2.366	3.665	6.251	7.815	11.345	16.266
4	0.711	1.064	2.195	3.357	4.878	7.779	9.488	13.277	18.467
5	1.145	1.610	3.000	4.351	6.064	9.236	11.070	15.086	20.515
6	1.635	2.204	3.828	5.348	7.231	10.645	12.592	16.812	22.458
7	2.167	2.833	4.671	6.346	8.383	12.017	14.067	18.475	24.322
8	2.733	3.490	5.527	7.344	9.524	13.362	15.507	20.090	26.124
9	3.325	4.168	6.393	8.343	10.656	14.684	16.919	21.666	27.877
10	3.940	4.865	7.267	9.342	11.781	15.987	18.307	23.209	29.588
	Non-significant						Significant		

Figure3　カイ二乗分布の棄却値表

あぁ、ですが、あなたの質問が聞こえてくるようです、$k-1$ の自由度を持つカイ二乗分布の棄却値はどのように見つけましょうか？ 何年も前に私が最初に心理統計の講義を受講した際に、図 **??**のような棄却値表でそれらの棄却値を調べていました。この図を見ると、p=0.05 で自由度 3 の χ^2 分布の棄却値は 7.815 であることがわかります。

なので、もし計算された χ^2 統計量が棄却値 7.815 よりも大きければ、帰無仮説を棄却できます(帰無仮説 H_0 は 4 つのマークが同じ確率で選択される、ということを思い出してくださいね)。以前実際に計算したので (i.e., $\chi^2 = 8.44$)、帰無仮説は棄却できます。基本的にはこれで終わりです。いまや「適合度に関するピアソンの χ^2 検定」がわかりましたね。ラッキーですね。

0.1.7　JASP でのやり方

当然ですが、JASP はこれらの計算を行う分析を提供します。メインの ‘Analyses’ ツールバーから ‘Frequencies’ - ‘Multinomial Test’ を選択しましょう。次に、表示される分析ウィンドウで、分析したい変数 (`choice_1`) を ‘Factor’ ボックスに移動します。また、‘Descriptives’ のチェックボックスをクリックして、結果の表に期待度数を表示しましょう。これら全てを実行すると、図 **??**のように JASP 上で分析結果が表示されます。JASP では上記で手計算したのと同じ期待度数と統計量が得られ、自由度 3 の chi^2 値はもちろん 8.44 となります。そんで p=0.038 です。JASP が 自由度 3 の

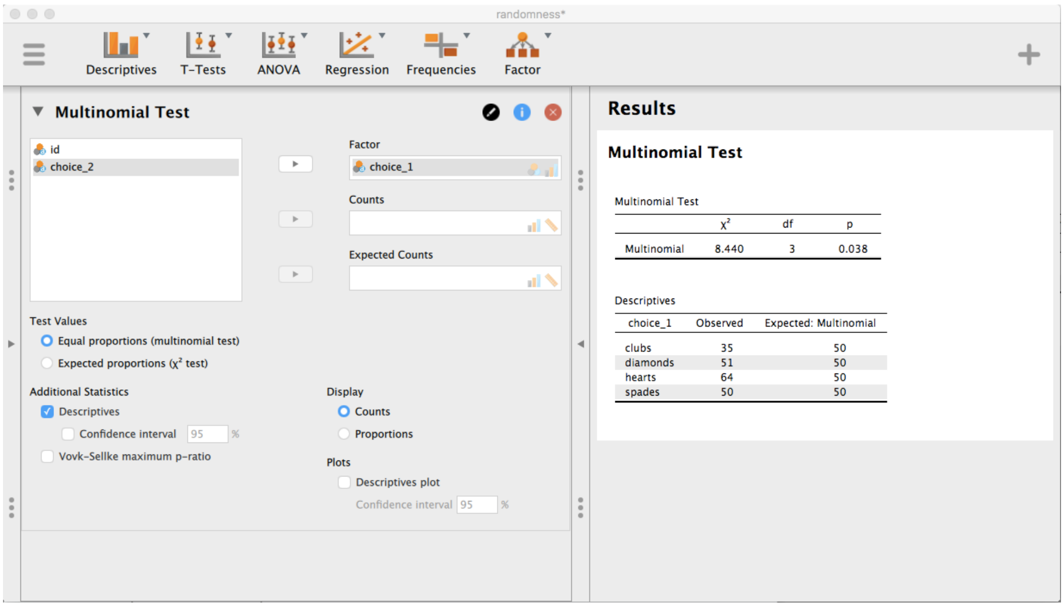

Figure4 JASP の χ^2 適合度検定で観測・期待度数を示しています。

χ^2 値による p 値を出してくれるので、棄却する p 値のしきい値を見る必要がなくなりました。

0.1.8 異なる帰無仮説の指定

適合度検定をしたいけれど、全てのカテゴリーが同じように選択されないという帰無仮説を持ってる場合にはどうしたらいいか、現段階では疑問に思うかもしれません。例えば、赤のカードを 60% 、黒のカードを 40% で選ぶはずだ、という理論的予測をした人がいて (なぜそんな予測をしたのかはわかりませんが)、他の好みがなかったとしましょう。もしそうであれば、ハート、ダイアモンドが 30% 、スペード、クラブが 20% で選択される確率を持った帰無仮説になります。いいかえればハートとダイアモンドはそれぞれ 60 回 (200 回中の 30% なので 60 回です) 選択され、スペードとクラブはそれぞれ 40 回 (200 回中の 20% なので 40 回です) 選択されるでしょう。ばかばかしい理論ではありますが、それでも、この明示的に指定された帰無仮説を JASP のデータでテストするのは非常に簡単です。分析ウィンドウ (図 **??**を見てください) で、'Expected Proportions (χ^2 test)' のラジオボタンをクリックできます。これをすれば、選択した変数に関する期待度数を入力するための選択肢が存在していて、我々の場合だとこれは `choice_1` です。図 **??**にあるように新しい帰無仮説を反映した数に変化して、結果の変化を確認しましょう。

期待された度数はこの通りです：

	♣	♢	♡	♠
期待度数 E_i	40	60	60	40

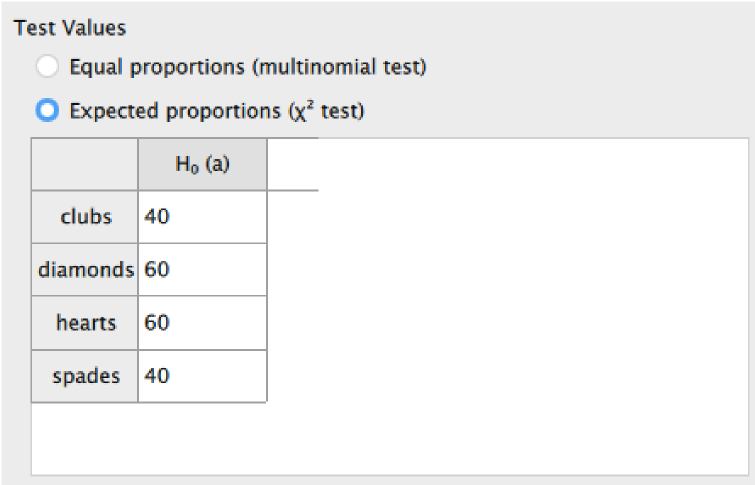

Figure5　JASP での χ^2 適合度検定における期待割合の変更

そして自由度 3 の χ^2 統計量は 4.742 で $p = 0.192$ です。では、更新された仮説と期待度数の結果は前回のものとは異なっています。結果として χ^2 検定統計量と p 値は異なるものになっています。残念ながら p 値は.192 なので、帰無仮説を棄却できません (セクション **??** を振り返って、理由を思い出してください)。帰無仮説はばかばかしいものであるにもかかわらず、データはそれに対する十分なエビデンスを提供していません。

0.1.9　検定結果の報告方法

これで検定がどのように機能するのか、そして素晴らしい JASP 風味な魔法の計算箱を用いて検定を行うやり方がわかったでしょう。次に知る必要があることは結果をどのように報告するのかです。せっかく実験を設計・実行してデータを分析したのにそれをだれにも伝えなければ意味がありませんよ！ 分析を報告する際に必要なことについてお話しましょう。これまでどおりカードマークの例で説明します。この結果をある論文なりなんなりに記述したいのであれば、通常の報告方法では次のように書きます：

> 実験に参加した 200 人のうち、最初にハートを選択したのが 64 人、ダイアモンドを選択したのが 51 人、スペードを選択したのが 50 人、クラブを選択したのが 35 人でした。全マークの選択確率が同一であるかどうかを検定するために、カイ二乗適合度検定を行った。結果は有意であり ($\chi^2(3) = 8.44, p < .05$)、人は完全にランダムなマークの選択をしなかったことがわかります。

これはかなりわかりやすいですし、うまくいけば非常に無難な感じになります。とはいえ、この記述にはいくつか注意すべき点があります：

- *記述統計は統計的検定に先行する*。つまり、検定を行う前にデータがどのようなものかを読み手に伝えたのです。一般的には、これは良い実践です。読み手があなたのデータをあなたのように理解しているわけではないことを常に念頭に置きましょう。あなたがデータを適切に記述しなければ、読み手は統計的検定の意味が理解できず泣き寝入りすることでしょう。
- *検定に用いた帰無仮説について記述する*。正直に言うと、書き手は常にこれをする必要はありませんが、曖昧さが存在するような状況や読み手が使用されている統計ツールを熟知しているとは限らないときにはしばしばそれはいい考えなのです。ほとんどの場合、読み手はあなたの用いる検定の詳細について承知していない（あるいはおぼえていない）ので、それらを「思い出させる」のは一種の礼儀ですよ！ 適合度検定である限りは、科学者なら持っているであろう検定の知識に頼れるはずです（統計学の入門でほとんどカバーされてますので）。しかし、帰無仮説を（簡潔に！）明言化しておくのはそれでもいい考えです。なぜなら帰無仮説はあなたが検定に利用するものによっては異なる可能性がありますからね。これまでのカードの例でいえば、帰無仮説は 4 つのマークを選択する確率が同一 (i.e., $P_1 = P_2 = P_3 = P_4 = 0.25$) というものでしたが、その仮説が特別なものというわけではありません。適合度検定に $P_1 = 0.7$ で $P_2 = P_3 = P_4 = 0.1$ という帰無仮説を適用することも簡単にできました。なので帰無仮説を説明しておくのは読み手にとって助かります。また、帰無仮説を数式でなく言葉に記述しましたが、それは全く問題ありません。もしお望みであれば、数式で記述することはできます。ですが、多くの読み手は記号よりも単語のほうが読みやすいので、書き手は帰無仮説をできる限り言葉で表現しがちです。
- *「統計ブロック」が含まれている*。検定結果を報告するときには、結果がただ有意であったというだけでなく、「重要な」統計情報全てを報告する「統計ブロック」(i.e., 括弧の中にある数学っぽい部分) を含めました。カイ二乗適合度検定に関しては、報告された情報は検定統計量 (適合度統計量が 8.44)、検定に用いた分布 (自由度 3 の χ^2 で $\chi^2(3)$ と短く表現される)、結果が有意であったかどうか (今回は $p < .05$) の情報です。統計ブロックに含める必要のある情報は検定によって異なり、新しい検定を紹介するたびに統計ブロックがどんな感じになるべきかを紹介します。[*5] しかし、常に読み手が望めばその検定結果を確認できるように十分な情報を常に提供しておくというのは一般的な原則です。
- *結果が解釈されている*。結果が有意であったことを示すのに加えて、結果の解釈を提供しました (i.e., 人はランダムなマークの選択をしなかった)。これもまた読み手に対する優しさです。なぜなら、それはデータに何が起きたのか、何を信じればいいかについての情報を読み手

[*5] まぁまぁ。統計をどのように報告するべきかという慣習は学問分野によって多少異なりがちです。私は心理学者なので、心理学分野での報告方法にこだわる傾向があります。ですが、結果を確認できるように読み手に対して十分な情報を提供するという一般的な原則は極めて普遍的だと、私は思います。

に伝えるからです。もしこういうのを含めなければ、何が起きているのかを読み手が理解するのは非常に困難でしょう。[*6]

なんでもそうなんですが、読み手に対して説明することを第一に考えてください。結果を報告するということは他の人間とコミュニケーションすることであることを常に覚えておくようにしましょう。私はこれまでに、書き手が全ての数字が含まれているかだけに専念して読み手とコミュニケーションすることを失念してしまったために、何度もレポート・論文・科学的文献でさえも結果セクションが難解になってしまったものを目にしてきました。

0.1.10 統計的表記についてのコメント

サタンは統計も聖句を引用することも同じように楽しむ
– H.G. Wells

もしあなたがよく読んでいて、私と同じくらい数学的衒学者であれば、前のセクションで書いたカイ二乗検定に関して少しだけ気になっているかもしれないことが一つあります。“$\chi^2(3) = 8.44$” と書くのはなにか違和感がある、と思われるかもしれません。結局のところ、8.44 になるのは適合度統計量なので、$X^2 = 8.44$ あるいは GOF$= 8.44$ と書くべきだったのではないでしょうか？ これは*標本分布* (i.e., $df = 3$ の χ^2) と *検定統計量* (i.e., X^2) を混同しているようです。χ と X はとても似ているので、タイプミスだと思った人もいるでしょう。奇しくも、そうではありません。$\chi^2(3) = 8.44$ という記述は本質的に、「検定統計量の標本分布は $\chi^2(3)$ であり、検定統計量の値が 8.44 です」という記述を非常に凝縮した方法です。

ある意味で、これはばかげたことです。カイ二乗の標本分布を持つ検定統計量なんて*ごまん*とあるのです。適合度検定で用いた X^2 統計量はその中の一つにすぎません (一番エンカウント率の高いものではありますが)。賢明で完全に組織化された世界の中では、*常に*検定統計量と標本分布には別々の名前がつけられます。そうすれば、統計ブロック自体が研究者が計算したものを正確に伝えてくれます。時々こういうことが起こります。例えば、ピアソンの適合度検定で用いられた検定統計量は X^2 ですが、G-検定として知られる密接に関連した検定があります[*7]が (**Sokal1994**)、そこでは検定

[*6]一部の人にとっては、このアドバイスは奇妙に聞こえるか、少なくともテクニカルレポートの書き方における「一般的な」アドバイスとは矛盾するものであるかもしれません。よくあることとして、学生は「結果」セクションはデータの記述と統計分析の報告用であり、「考察」セクションは解釈を提供するためのものであると言われます。確かにその通りなのですが、あまりにも文字通りに解釈してしまう人が多いのではないでしょうか。わたしがよくやっているのは結果セクションにデータの迅速かつ単純な解釈を提供することです。それにより、読み手はデータが示していることを理解できます。そして考察では、自分の結果がどのようにこれまでの科学的文献と整合するのかについてより大きなストーリーを語るようにしています。要するに、「解釈は考察の中で行う」というアドバイスで結果セクションを解釈できないゴミにさせてはいけません。読み手による理解こそが*より*重要なのです。

[*7]複雑なことに、G 検定とは*尤度比検定*として知られる一連の検定の特殊なケースです。この本では尤度比検定はカバーしていませんが、知っておくとかなり便利なものです。

統計量が G です。偶然にも、ピアソンの適合度検定と G 検定はともに同じ帰無仮説を検定し、標本分布も全く同じものです (i.e., $k-1$ の自由度を持つカイ二乗分布). もしカードデータに対して適合度検定でなく G 検定を行った場合、最終的に検定統計量は $G = 8.65$ となり、以前に獲得した $X^2 = 8.44$ とは少し異なり、p 値も少し小さくなります $p = .034$。検定統計量、標本分布、p 値の順に報告するのが慣例と仮定しましょう。もしそうであれば、二つの状況で異なる統計ブロックができます：オリジナルの結果は $X^2 = 8.44, \chi^2(3), p = .038$、一方で G 検定の新しいバージョンは $G = 8.65, \chi^2(3), p = .034$ と記述されます。しかし凝縮報告基準を用いると、オリジナルの結果だと $\chi^2(3) = 8.44, p = .038$、新しい方だと $\chi^2(3) = 8.65, p = .034$ と書かれるので、実際にどちらの検定を行ったのかは不明瞭です。

では、統計ブロックの中身が行った検定を一意に特定する世界に住んでみたくないですか？ 人生はごちゃごちゃしてますもの。我々は (統計ツールのユーザーとして) キレイで整理整頓されている状態を望みます。プロダクトのようにその状態を*デザイン*されたものが欲しいのですが、人生はそうはいきません。統計学は他と同じように知的学問であり、そのため、誰も完全に理解していない、大規模に分散され、部分的に協調的だったり競争的だったりするプロジェクトです。私とあなたがデータ分析ツールとして用いるものは統計学の神様による所業から作られたものではなかったのです。それらは多くの人たちによって発明され、学術雑誌に論文として出版され、他の人たちによって実装・訂正・修正され、他のだれかが教科書を通して学生に説明しました。結果として名前さえない検定統計量が *たくさん* 存在し、対応する標本分布と同じ名前がつけられてます。のちに見るように、χ^2 分布に従う検定統計量は「カイ二乗統計量」と呼ばれ、t 分布の場合は「t 統計量」などと呼ばれます。ですが、χ^2 と G の例で示したように、同じ標本分布を持つ二つの異なるものは、やはり、異なるものになります。

最終的に、実際に行った検定がなんであるかを明確にすることはしばしば良い考えです。特に一般的でないものをやっているときには。「カイ二乗検定」というだけでは、あなたが話している検定がどういうものかは不明瞭です。二つの有名なカイ二乗検定が適合度検定と独立性検定 (Section **??**) であるために、多くの統計訓練を受けている読み手は推測できるでしょう。とはいえ、気を付けなければならないことなんです。

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。説明していないことではなく、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。なんにせよ、**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 (じじょう)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に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、ダイアモンド) 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	ダイアモンド	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 it 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
ダイアモンド, ♢	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 bmO とします。

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

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。$\boldsymbol{O} = (35, 51, 64, 50)$ ということで、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 $\boldsymbol{O}$ として分類するように、帰無仮説と対応する確率として $\boldsymbol{P}$ を用います。そのため、帰無仮説を記述する確率の集合を $\boldsymbol{P} = (P_1, P_2, P_3, P_4)$ とすると、以下のようになります:

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

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 $\boldsymbol{P}$ 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

そうである必要はありません。例えば、もし実験課題で他のマークの 2 倍クラブが含まれているデッキを想像してもらう場合には、帰無仮説は $\boldsymbol{P} = (.4, .2, .2, .2)$ となるでしょう．確率がすべて正の値であり、その合計が 1 である限りは、それは帰無仮説として正当な選択です。ですが、適合度検定では一般的に全てのカテゴリーが同様の確率である帰無仮説を用います。そのため、ここではそれに固執します。

対立仮説 H_1 はどうでしょうか? 我々の関心は、関係する確率が全て同じではないこと (つまり、人々の選択が完全にランダムではなかったこと) を実証することです。その結果、「人にやさしい (負担の小さい)」バージョンの仮説はこんな感じです：

H_0: 4 つのマークが同じ確率で選択される
H_1: 少なくとも 1 つのマークの選択確率が 0.25 *ではない*

そして「数学者にやさしい」バージョンはこうです:

H_0: $\boldsymbol{P} = (.25, .25, .25, .25)$
H_1: $\boldsymbol{P} \neq (.25, .25, .25, .25)$

1.1.3 “適合度” 検定の統計量

この段階で、観測された頻度数 $\boldsymbol{O}$ と検定予定の帰無仮説と対応する確率の集合である $\boldsymbol{P}$ を我々は手にしています。ここでしたいことは、帰無仮説検定の構築です。いつものように、H_1 に対して H_0 を検定したい場合には、検定統計量が必要です。適合度検定の基本的なトリックは、データが帰無仮説にどれだけ「近いのか」を測定する検定統計量を組み立てることです。もし帰無仮説が真であるときの「期待値」がデータと似ていなければ、帰無仮説は真ではないでしょう。オーケイ、帰無仮説が真であるならどうなるだろう？ 正しい言い方をすれば、「期待度数」とは何かということです。$N = 200$ の観測データがあり、(もし帰無仮説が真であれば) ハートの選択確率は $P_3 = .25$ で、ハートの期待値は $200 \times .25 = 50$ ですよね? より具体的には、もし E_i が、「帰無仮説が真であるときにカテゴリー i の選択数の期待値」とすると次のようになります。

$$E_i = N \times P_i$$

この計算はとても簡単です。もし 4 つのカテゴリーに分類されうる 200 個の観測データがあって、カテゴリー全ての選択確率が同じだとすれば、各カテゴリーの観測データは 50 であると期待されますよね？

さて、これをどのように検定統計量に変換するのでしょうか？ 明らかに、我々のしたいことは各カテゴリーの*期待*値 (E_i) と*観測*値 (O_i) の比較です。そしてこの比較に基づいて、我々は良い検定統計量を導き出すことができるはずです。まずは、帰無仮説が期待した結果と我々が実際に得られた結果との差を計算しましょう。つまり、「観測値から期待値を引いた」差得点である $O_i - E_i$ を計算

します。これを図示すると次の表のようになります。

		♣	♢	♡	♠
期待度数	E_i	50	50	50	50
観測度数	O_i	35	51	64	50
差得点	$O_i - E_i$	-15	1	14	0

つまり我々の計算によって、帰無仮説の予測よりも人はハートを多く、クラブを少なく選択していることがわかりました。しかし、ちょっと考えてみると、この素朴な違いは、私たちが求めているものとはちょっと違うようです。直観的に、帰無仮説の予測が少なすぎた場合 (ハート) も予測が多すぎた場合 (クラブ) も同程度によくないことに感じられます。なので、クラブではマイナスでハートではプラスだというのはちょっと変な感じです。これを解決する一つの簡単な方法は全てを二乗することで、ここでは二乗された差を計算します $(O_i - E_i)^2$。前回同様、これは手作業でできます：

		♣	♢	♡	♠
期待度数	E_i	50	50	50	50
観測度数	O_i	35	51	64	50
差得点	$O_i - E_i$	-15	1	14	0
二乗差	$(O_i - E_i)^2$	225	1	196	0

さぁ、これで一歩前進です。今手にしているものは、帰無仮説が悪い予測をしたときには大きく (クラブとハート)、良い予測をしたときには小さくなる数の集合です。次は、後述予定の技術的理由により、それらの数を期待度数 E_i で割って、**調整された** 二乗値 $\frac{(E_i - O_i)^2}{E_i}$ を計算しています。今回の例では全てのカテゴリーで $E_i = 50$ となるので、あまり面白い計算ではないですが、とりあえずやってみましょう：

		♣	♢	♡	♠
期待度数	E_i	50	50	50	50
観測度数	O_i	35	51	64	50
差得点	$O_i - E_i$	-15	1	14	0
二乗差	$(O_i - E_i)^2$	225	1	196	0
調整済み二乗差	$(O_i - E_i)^2/E_i$	4.50	0.02	3.92	0.00

要するに、ここで得たのは 4 つの「エラー」得点で、それぞれが観測度数に対する帰無仮説の予測から生じた「間違い」の大きさを示しています。そして、これを一つの有用な検定統計量に変換するためには、それらの数を単に足し合わせることが一つのやり方です。その結果を **適合度:goodness-of-fit** とよび、慣習的には χ^2 (カイ二乗) または (頭文字をとって)GOF ともよばれています。$4.50 + 0.02 + 3.92 + 0.00 = 8.44$ として計算可能です。

k をカテゴリーの総数だとすれば (i.e., カードデータの例だと $k = 4$)、χ^2 統計量は以下のように与えられます：

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

直観的に、もし χ^2 が小さければ観測データ O_i が帰無仮説の予測する E_i に近づき、帰無仮説を棄却するためには大きな χ^2 が必要となるはずです。

計算の結果、カードデータセットでは $\chi^2 = 8.44$ という値が得られました。次の疑問はこれが帰無仮説を棄却するのに十分な値なのか、ということです。

1.1.4 適合度の標本分布

χ^2 の値が帰無仮説を棄却するほどに大きいかどうかを決めるために、帰無仮説が真である場合の χ^2 に関する標本分布はどうなるかを理解する必要があるでしょう。ということで、今回のセクションではそれをやっていきます。この標本分布がどのように構成されるかの詳細をお見せして、次のセクションではそれを仮説検定の構築に用います。さて本題。標本分布が $k-1$ の自由度を持つ **χ^2 (カイ二乗) 分布**であると喜んで受け入れる人は、このセクションの残りをスキップできます。しかし、もし適合度検定の *仕組み* を理解したいひとは、ぜひこの先をお読みください。

よし、帰無仮説が実際に真であると仮定しましょう。もしそうであれば、ある観測変数が i 番目のカテゴリーに属する真の確率は P_i となります。まぁ結局、それはほぼほぼ帰無仮説の定義です。これが意味するところについて考えます。これは、「自然」が、重み付きのコイン (i.e., 表が出る確率は P_j) を裏返すことで観測値がカテゴリー i に含まれるかどうかを決定するというようなものです。したがって、自然がこれらのコインの N を反転させること（データセット内の観測ごとに 1 つ）を想像して、正確に O_i を頭に浮かべることで、観測された頻度 O_i を考えることができます。明らかに、これは実験を考える上ではかなり奇妙なやり方です。ですが、このシナリオには見覚えがありますね（と私は期待してますよ)。セクション **??**、二項分布のケースとまったく同じ設定です。言い換えれば、帰無仮説が真であれば、観測された度数は二項分布のサンプリングによって生成されたことになります。

$$O_i \sim \text{Binomial}(P_i, N)$$

中心極限定理 (Section **??**) の説明を思い出すと、特に N が大きく P_i が 0 または 1 に近 *すぎ* ない時に、二項分布は正規分布と近似して見えるようになります。いいかえれば、$N \times P_i$ が十分に大きければいいのです。また、別の言い方をすれば、期待度数 E_i が十分に大きい場合 O_i の理論的な分布は近似的に正規分布となります。さらにいえば、O_i が正規分布していれば、そのとき $(O_i - E_i)/\sqrt{E_i}$ も正規分布します。E_i は固定の値なので、E_i を引いて $\sqrt{E_i}$ で割ることで正規分布の平均と標準偏差が変化しますが、それだけです。では早速、適合度統計量 *とはなにか*について見ていきましょう。

今しているのは正規分布するものをたくさん集めて、二乗して、それからそれらを足し合わせているのです。おっと。これも見たことありますね！ セクション **??**でお話ししたように、標準正規分布 (i.e., 平均 0, 標準偏差 1) を持つものをたくさん集めて二乗してから足し合わせると、その結果はカイ二乗分布となります。これで適合度統計量の標本分布がカイ二乗分布であることを帰無仮説が予測している、ということがわかりました。イイね。

最後にもう一つ、いわゆる自由度についてお話しときましょう。セクション **??**を思い返せば、足し合わせるものの数は k、結果として生じるカイ二乗分布の自由度は k になると言いましたね. しかし、このセクションの冒頭で述べたのはカイ二乗適合度検定の自由度は $k-1$ であるということです。どうしたというのでしょう？ ここでの答えは、私たちが注目しているのは、純粋に *独立した* ものが同時に足し合わされている数だということです。また、次のセクションでお話ししますが、たとえ k 個分追加したとしても真に独立しているのは $k-1$ 個のみであり、自由度は $k-1$ だけです. それが次のセクションでの話題です。[*3]

1.1.5 自由度

セクション **??** でカイ二乗分布を紹介したときに、「**自由度**」が実際に *意味* するところは少し曖昧でした. 明らかに、そこは重要な点です。Figure **??**を見ると、自由度を変化させるとカイ二乗分布の形がかなり大きく変わっています。しかし、それはなんなのでしょうか？ 分布を紹介して正規分布との関係性を説明したときに、ある答えを提供しました：私が二乗して足し合わせた「正規分布する」数です。ですが、多くの人にとって、それは抽象的でちっとも参考になりません。ここで本当に目指すべきなのは、我々の持つデータを用いて自由度を理解することです。ではいってみましょう。

自由度の基本的な考え方はとてもシンプルです。データの記述に使用する明確な「量」を数え上げることで計算して、それらのデータが満たさなければならない「制約」をすべて引き算します。[*4] これでは少し曖昧なので、具体的な例としてのカードデータを使いましょう。4 カテゴリー (ハート、クラブ、ダイアモンド、スペード) の観測度数に対応する O_1、O_2、O_3、O_4 の 4 つを用いてデータを記述します。それら 4 つの数が今回の実験での *ランダムな結果* です。しかし、実験には固定の制約が組み込まれています：サンプルサイズが N。[*5]つまり、ハートを選んだ人の数、ダイアモンドを

[*3] もし適合度統計量の式を $k-1$ の独立したものの数の和として書き直すと、$k-1$ の自由度を持つカイ二乗の「適切な」標本分布を得られます。その数学の詳細を示すことは入門書の範囲を逸脱しています。ここでしたいのは、なぜ適合度統計量がカイ二乗分布と関連しているのかについて理解してもらうことです。

[*4] これは単純すぎると指摘せざるを得ないとは思っています。大体はその説明でうまくいくんですが、整数ではない自由度の値に出くわすこともあります。気にしすぎることはありません；そんなときは「自由度」が少し厄介な概念であり、私がここでしているような単純なストーリーがすべてではないことを思い返してください。入門編では単純なストーリーに固執するのがいいんですが、このストーリーが崩壊することは事前に警告しておくのがベストだと思います。もし警告がなければ $df = 3.4$ みたいなのを目にしたときに混乱してしまい、(正確に) 私が教えなかったことに気付くのではなく (不正確に) 私が教えたことを誤解したんじゃないかと考えてしまいます。

[*5] 実際問題として、サンプルサイズは常に固定されていません。例えば、一定期間にわたって実験をする際に参加者数は参加する人数に依存しているかもしれません。まぁそれは今の目的には関係ないです。

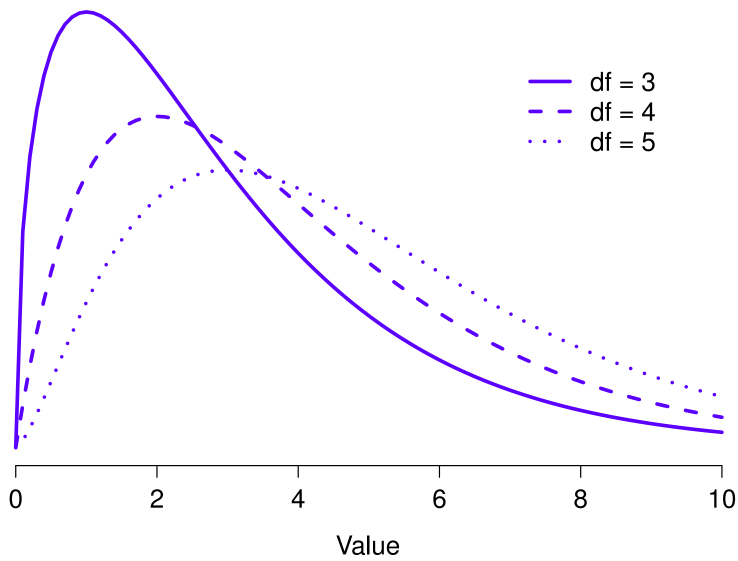

Figure1.1 「自由度」の異なる χ^2 (カイ二乗) 分布。

選んだ人の数、クラブを選んだ人の数がわかれば、スペードを選んだ数は正確に把握できます。言い換えれば、4 つの数を用いてデータが記述されますが、実際には $4-1=3$ の自由度にしか対応していません。ちょっと違う考え方は、関心のある 4 つの*確率*があること (ここでも 4 つのカテゴリーに対応します) に注意することですが、それらの確率の合計値は 1 でなければならないという制約が課されます。そうなると自由度は $4-1=3$ になります。観測度数の観点から考えたいのか、確率の観点から考えたいのかに関係なく、答えは同じです。一般的に、k グループを含む実験で χ^2 (カイ二乗) 適合度検定を実行すると、自由度は $k-1$ になります。

1.1.6 帰無仮説検定

仮説検定を構築するプロセスの最終段階は、棄却域とは何なのかを理解することです。いわば、χ^2 の値によって帰無仮説が棄却されます。以前のように、χ^2 の値が大きいということは帰無仮説が実験データの予測を行うのに不十分であったことを意味します。一方で χ^2 の値が小さいということは帰無仮説が支持されていることを意味します。したがって、χ^2 が棄却値よりも大きいければ帰無仮説を棄却し、χ^2 が棄却値よりも小さければ帰無仮説を保持する、というのは非常に賢明なやり方です。チャプター**??**で紹介した言語を使えば、カイ二乗適合度検定は常に **片側検定** です。そのとおり。あとはこの重要な値が何であるかを考えればいいわけです。そしてそれはとても簡単です。もし

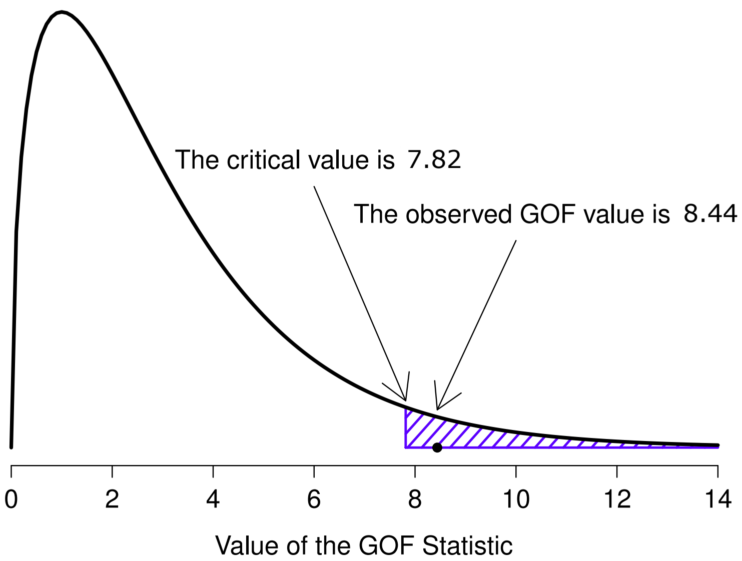

Figure1.2　χ^2 (カイ二乗) 適合度検定で仮説検定がどのように機能するのかを示す図。

有意水準を $\alpha = .05$ と設定して (すなわち、Type1 エラーを 5% で許容する) 検定を行いたい場合、帰無仮説が真である際に χ^2 がその値を超える可能性が 5% になるように、棄却値を選択する必要があります。これを示したのが図 **??**です。

あぁ、ですが、あなたの質問が聞こえてくるようです、$k-1$ の自由度を持つカイ二乗分布の棄却値はどのように見つけましょうか？ 何年も前に私が最初に心理統計の講義を受講した際に、図 **??**のような棄却値表でそれらの棄却値を調べていました。この図を見ると、p=0.05 で自由度 3 の χ^2 分布の棄却値は 7.815 であることがわかります。

なので、もし計算された χ^2 統計量が棄却値 7.815 よりも大きければ、帰無仮説を棄却できます (帰無仮説 H_0 は 4 つのマークが同じ確率で選択される、ということを思い出してくださいね)。以前実際に計算したので (i.e., $\chi^2 = 8.44$)、帰無仮説は棄却できます。基本的にはこれで終わりです。いまや「適合度に関するピアソンの χ^2 検定」がわかりましたね。ラッキーですね。

1.1.7　JASP でのやり方

当然ですが、JASP はこれらの計算を行う分析を提供します。メインの ‘Analyses’ ツールバーから ‘Frequencies’ - ‘Multinomial Test’ を選択しましょう。次に、表示される分析ウィンドウで、分析し

Degrees of Freedom	Probability								
	0.95	0.90	0.70	0.50	0.30	0.10	0.05	0.01	0.001
1	0.004	0.016	0.148	0.455	1.074	2.706	3.841	6.635	10.828
2	0.103	0.211	0.713	1.386	2.408	4.605	5.991	9.210	13.816
3	0.352	0.584	1.424	2.366	3.665	6.251	7.815	11.345	16.266
4	0.711	1.064	2.195	3.357	4.878	7.779	9.488	13.277	18.467
5	1.145	1.610	3.000	4.351	6.064	9.236	11.070	15.086	20.515
6	1.635	2.204	3.828	5.348	7.231	10.645	12.592	16.812	22.458
7	2.167	2.833	4.671	6.346	8.383	12.017	14.067	18.475	24.322
8	2.733	3.490	5.527	7.344	9.524	13.362	15.507	20.090	26.124
9	3.325	4.168	6.393	8.343	10.656	14.684	16.919	21.666	27.877
10	3.940	4.865	7.267	9.342	11.781	15.987	18.307	23.209	29.588
	Non-significant						Significant		

Figure1.3　カイ二乗分布の棄却値表

たい変数 (`choice_1`) を 'Factor' ボックスに移動します。また、'Descriptives' のチェックボックスをクリックして、結果の表に期待度数を表示しましょう。これら全てを実行すると、図 **??**のように JASP 上で分析結果が表示されます。JASP では上記で手計算したのと同じ期待度数と統計量が得られ、自由度 3 の chi^2 値はもちろん 8.44 となります。そんで p=0.038 です。JASP が 自由度 3 の χ^2 値による p 値を出してくれるので、棄却する p 値のしきい値を見る必要がなくなりました。

1.1.8　異なる帰無仮説の指定

適合度検定をしたいけれど、全てのカテゴリーが同じように選択されないという帰無仮説を持ってる場合にはどうしたらいいか、現段階では疑問に思うかもしれません。例えば、赤のカードを 60% 、黒のカードを 40% で選ぶはずだ、という理論的予測をした人がいて (なぜそんな予測をしたのかはわかりませんが)、他の好みがなかったとしましょう。もしそうであれば、ハート、ダイアモンドが 30% 、スペード、クラブが 20% で選択される確率を持った帰無仮説になります。いいかえればハートとダイアモンドはそれぞれ 60 回 (200 回中の 30% なので 60 回です) 選択され、スペードとクラブはそれぞれ 40 回 (200 回中の 20% なので 40 回です) 選択されるでしょう。ばかばかしい理論ではありますが、それでも、この明示的に指定された帰無仮説を JASP のデータでテストするのは非常に簡単です。分析ウィンドウ (図 **??**を見てください) で、'Expected Proportions (χ^2 test)' のラジオボタンをクリックできます。これをすれば、選択した変数に関する期待度数を入力するための選択肢が存在していて、我々の場合だとこれは `choice_1` です。図 **??**にあるように新しい帰無仮説

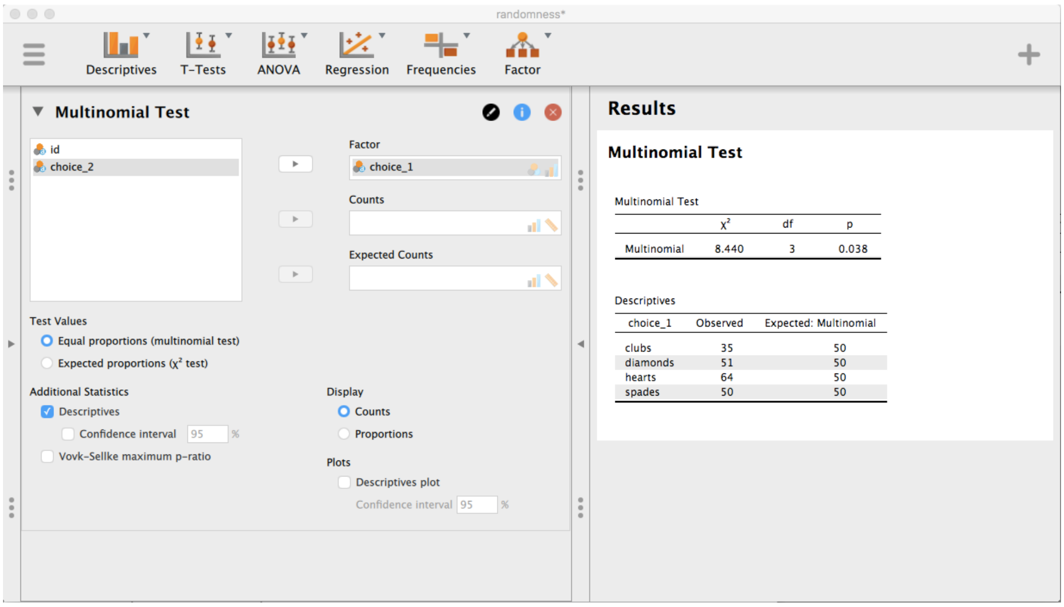

Figure1.4　JASP の χ^2 適合度検定で観測・期待度数を示しています。

を反映した数に変化して、結果の変化を確認しましょう。

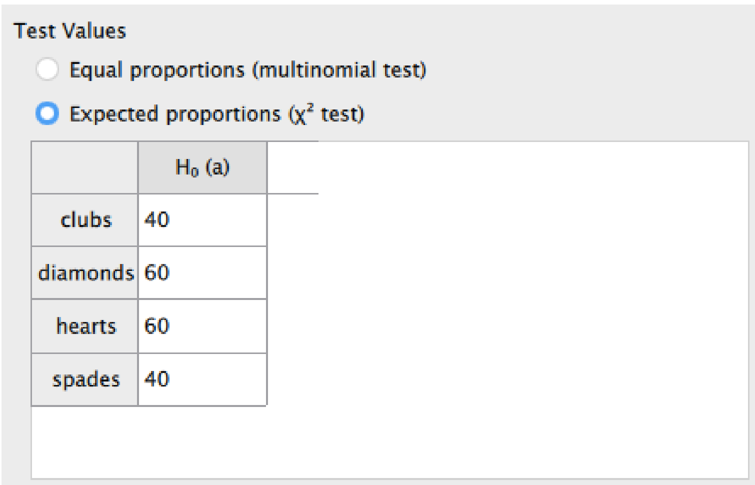

Figure1.5　JASP での χ^2 適合度検定における期待割合の変更

期待された度数はこの通りです：

		♣	♢	♡	♠
期待度数	E_i	40	60	60	40

そして自由度 3 の χ^2 統計量は 4.742 で $p = 0.192$ です。では、更新された仮説と期待度数の結果は前回のものとは異なっています。結果として χ^2 検定統計量と p 値は異なるものになっています。残念ながら p 値は.192 なので、帰無仮説を棄却できません (セクション **??** を振り返って、理由を思い出してください)。帰無仮説はばかばかしいものであるにもかかわらず、データはそれに対する十分なエビデンスを提供していません。

1.1.9 検定結果の報告方法

これで検定がどのように機能するのか、そして素晴らしい JASP 風味な魔法の計算箱を用いて検定を行うやり方がわかったでしょう。次に知る必要があることは結果をどのように報告するのかです。せっかく実験を設計・実行してデータを分析したのにそれをだれにも伝えなければ意味がありませんよ！ 分析を報告する際に必要なことについてお話しましょう。これまでどおりカードマークの例で説明します。この結果をある論文なりなんなりに記述したいのであれば、通常の報告方法では次のように書きます：

> 実験に参加した 200 人のうち、最初にハートを選択したのが 64 人、ダイアモンドを選択したのが 51 人、スペードを選択したのが 50 人、クラブを選択したのが 35 人でした。全マークの選択確率が同一であるかどうかを検定するために、カイ二乗適合度検定を行った。結果は有意であり ($\chi^2(3) = 8.44, p < .05$)、人は完全にランダムなマークの選択をしなかったことがわかります。

これはかなりわかりやすいですし、うまくいけば非常に無難な感じになります。とはいえ、この記述にはいくつか注意すべき点があります：

- *記述統計は統計的検定に先行する*。つまり、検定を行う前にデータがどのようなものかを読み手に伝えたのです。一般的には、これは良い実践です。読み手があなたのデータをあなたのように理解しているわけではないことを常に念頭に置きましょう。あなたがデータを適切に記述しなければ、読み手は統計的検定の意味が理解できず泣き寝入りすることでしょう。
- *検定に用いた帰無仮説について記述する*。正直に言うと、書き手は常にこれをする必要はありませんが、曖昧さが存在するような状況や読み手が使用されている統計ツールを熟知しているとは限らないときにはしばしばそれはいい考えなのです。ほとんどの場合、読み手はあなたの用いる検定の詳細について承知していない（あるいはおぼえていない）ので、それらを「思い出させる」のは一種の礼儀ですよ！ 適合度検定である限りは、科学者なら持っているであろう検定の知識に頼れるはずです（統計学の入門でほとんどカバーされてますので)。しかし、帰

無仮説を（簡潔に！）明言化しておくのはそれでもいい考えです。なぜなら帰無仮説はあなたが検定に利用するものによっては異なる可能性がありますからね。これまでのカードの例でいえば、帰無仮説は 4 つのマークを選択する確率が同一 (i.e., $P_1 = P_2 = P_3 = P_4 = 0.25$) というものでしたが、その仮説が特別なものというわけではありません。適合度検定に $P_1 = 0.7$ で $P_2 = P_3 = P_4 = 0.1$ という帰無仮説を適用することも簡単にできました。なので帰無仮説を説明しておくのは読み手にとって助かります。また、帰無仮説を数式でなく言葉に記述しましたが、それは全く問題ありません。もしお望みであれば、数式で記述することはできます。ですが、多くの読み手は記号よりも単語のほうが読みやすいので、書き手は帰無仮説をできる限り言葉で表現しがちです。

- *「統計ブロック」が含まれている*。検定結果を報告するときには、結果がただ有意であったというだけでなく、「重要な」統計情報全てを報告する「統計ブロック」(i.e., 括弧の中にある数学っぽい部分) を含めました。カイ二乗適合度検定に関しては、報告された情報は検定統計量 (適合度統計量が 8.44)、検定に用いた分布 (自由度 3 の χ^2 で $\chi^2(3)$ と短く表現される)、結果が有意であったかどうか (今回は $p < .05$) の情報です。統計ブロックに含める必要のある情報は検定によって異なり、新しい検定を紹介するたびに統計ブロックがどんな感じになるべきかを紹介します。[*6] しかし、常に読み手が望めばその検定結果を確認できるように十分な情報を常に提供しておくというのは一般的な原則です。
- *結果が解釈されている*。結果が有意であったことを示すのに加えて、結果の解釈を提供しました (i.e., 人はランダムなマークの選択をしなかった)。これもまた読み手に対する優しさです。なぜなら、それはデータに何が起きたのか、何を信じればいいかについての情報を読み手に伝えるからです。もしこういうのを含めなければ、何が起きているのかを読み手が理解するのは非常に困難でしょう。[*7]

なんでもそうなんですが、読み手に対して説明することを第一に考えてください。結果を報告するということは他の人間とコミュニケーションすることであることを常に覚えておくようにしましょう。私はこれまでに、書き手が全ての数字が含まれているかだけに専念して読み手とコミュニケーションすることを失念してしまったために、何度もレポート・論文・科学的文献でさえも結果セクション

[*6] まぁまぁ。統計をどのように報告するべきかという慣習は学問分野によって多少異なりがちです。私は心理学者なので、心理学分野での報告方法にこだわる傾向があります。ですが、結果を確認できるように読み手に対して十分な情報を提供するという一般的な原則は極めて普遍的だと、私は思います。

[*7] 一部の人にとっては、このアドバイスは奇妙に聞こえるか、少なくともテクニカルレポートの書き方における「一般的な」アドバイスとは矛盾するものであるかもしれません。よくあることとして、学生は「結果」セクションはデータの記述と統計分析の報告用であり、「考察」セクションは解釈を提供するためのものであると言われます。確かにその通りなのですが、あまりにも文字通りに解釈してしまう人が多いのではないでしょうか。わたしがよくやっているのは結果セクションにデータの迅速かつ単純な解釈を提供することです。それにより、読み手はデータが示していることを理解できます。そして考察では、自分の結果がどのようにこれまでの科学的文献と整合するのかについてより大きなストーリーを語るようにしています。要するに、「解釈は考察の中で行う」というアドバイスで結果セクションを解釈できないゴミにさせてはいけません。読み手による理解こそが*より*重要なのです。

が難解になってしまったものを目にしてきました。

1.1.10 統計的表記についてのコメント

> *サタンは統計も聖句を引用することも同じように楽しむ*
>
> – H.G. Wells

もしあなたがよく読んでいて、私と同じくらい数学的衒学者であれば、前のセクションで書いたカイ二乗検定に関して少しだけ気になっているかもしれないことが一つあります。“$\chi^2(3) = 8.44$” と書くのはなにか違和感がある、と思われるかもしれません。結局のところ、8.44 になるのは適合度統計量なので、$X^2 = 8.44$ あるいは GOF= 8.44 と書くべきだったのではないでしょうか？ これは*標本分布* (i.e., $df = 3$ の χ^2) と *検定統計量* (i.e., X^2) を混同しているようです。χ と X はとても似ているので、タイプミスだと思った人もいるでしょう。奇しくも、そうではありません。$\chi^2(3) = 8.44$ という記述は本質的に、「検定統計量の標本分布は $\chi^2(3)$ であり、検定統計量の値が 8.44 です」という記述を非常に凝縮した方法です。

ある意味で、これはばかげたことです。カイ二乗の標本分布を持つ検定統計量なんて*ごまんと*あるのです。適合度検定で用いた X^2 統計量はその中の一つにすぎません (一番エンカウント率の高いものではありますが)。賢明で完全に組織化された世界の中では、*常に*検定統計量と標本分布には別々の名前がつけられます。そうすれば、統計ブロック自体が研究者が計算したものを正確に伝えてくれます。時々こういうことが起こります。例えば、ピアソンの適合度検定で用いられた検定統計量は X^2 ですが、G-検定として知られる密接に関連した検定があります[*8]が (**Sokal1994**)、そこでは検定統計量が G です。偶然にも、ピアソンの適合度検定と G 検定はともに同じ帰無仮説を検定し、標本分布も全く同じものです (i.e., $k-1$ の自由度を持つカイ二乗分布). もしカードデータに対して適合度検定でなく G 検定を行った場合、最終的に検定統計量は $G = 8.65$ となり、以前に獲得した $X^2 = 8.44$ とは少し異なり、p 値も少し小さくなります $p = .034$。検定統計量、標本分布、p 値の順に報告するのが慣例と仮定しましょう。もしそうであれば、二つの状況で異なる統計ブロックができます：オリジナルの結果は $X^2 = 8.44, \chi^2(3), p = .038$、一方で G 検定の新しいバージョンは $G = 8.65, \chi^2(3), p = .034$ と記述されます。しかし凝縮報告基準を用いると、オリジナルの結果だと $\chi^2(3) = 8.44, p = .038$、新しい方だと $\chi^2(3) = 8.65, p = .034$ と書かれるので、実際にどちらの検定を行ったのかは不明瞭です。

では、統計ブロックの中身が行った検定を一意に特定する世界に住んでみたくないですか？ 人生はごちゃごちゃしてますもの。我々は (統計ツールのユーザーとして) キレイで整理整頓されている状態を望みます。プロダクトのようにその状態を*デザイン*されたものが欲しいのですが、人生はそ

[*8] 複雑なことに、G 検定とは*尤度比検定*として知られる一連の検定の特殊なケースです。この本では尤度比検定はカバーしていませんが、知っておくとかなり便利なものです。

うはいきません。統計学は他と同じように知的学問であり、そのため、誰も完全に理解していない、大規模に分散され、部分的に協調的だったり競争的だったりするプロジェクトです。私とあなたがデータ分析ツールとして用いるものは統計学の神様による所業から作られたものではなかったのです。それらは多くの人たちによって発明され、学術雑誌に論文として出版され、他の人たちによって実装・訂正・修正され、他のだれかが教科書を通して学生に説明しました。結果として名前さえない検定統計量が *たくさん* 存在し、対応する標本分布と同じ名前がつけられてます。のちに見るように、χ^2 分布に従う検定統計量は「カイ二乗統計量」と呼ばれ、t 分布の場合は「t 統計量」などと呼ばれます。ですが、χ^2 と G の例で示したように、同じ標本分布を持つ二つの異なるものは、やはり、異なるものになります。

最終的に、実際に行った検定がなんであるかを明確にすることはしばしば良い考えです。特に一般的でないものをやっているときには。「カイ二乗検定」というだけでは、あなたが話している検定がどういうものかは不明瞭です。二つの有名なカイ二乗検定が適合度検定と独立性検定 (Section **??**) であるために、多くの統計訓練を受けている読み手は推測できるでしょう。とはいえ、気を付けなければならないことなんです。

1.2

独立性 (もしくは連関) に関する χ^2 検定

GUARDBOT 1:	*動くな!*
GUARDBOT 2:	*あなたはロボットですか、それとも人間ですか?*
LEELA:	*ロボット…私たちはね。*
FRY:	*ええと、うん! たった 2 台のロボットがそれをロボット化しています! え?*
GUARDBOT 1:	*テストを実施しましょう。*
GUARDBOT 2:	*次のうち、どれが一番いいですか?* *A: 子犬、B: いとしいあの娘からの可愛い花、* *もしくは C: 適切にフォーマットされた大きなデータファイル。*
GUARDBOT 1:	*選びなさい!*

– フューチュラマ、“Fear of a Bot Planet”

先日、私は*チャペック 9* という惑星に住む人々の風変わりな習慣を紹介するドキュメンタリーアニメを見てたんです。どうやら、首都にアクセスするためには訪問者は自身が人間でなくロボットであることを証明しなければならないようです。訪問者が人間であるかどうかを判断するために、ネイティブは子犬、花、大きく適切にフォーマットされたデータファイルのうちどれが好ましいかを尋

ねます。「こいつぁ賢い」と思いました。「しかし人間とロボットが同じ好みを持っていたら？ そうなると、あまりいいテストにはならないのではないだろうか？」たまたま、チャペック *9* の権限を持つ市民がこれをチェックするために使用したテストデータを手にすることができました。彼らがしたことはとてもシンプルであったことがわかりました。彼らは多くのロボットと人間を見つけて、彼らに何が好きかを尋ねました。彼らのデータを `chapek9.csv` というファイルに保存しておいたので、JASP で読み込むことができるようになってます。個人を特定する `ID` 変数と同じように、二つの名義変数が存在します、`species` と `choice`。このデータセットには、選択を求められた人 (ロボットと人の両方を「人」として数えます) ごとで計 180 のデータがあります。具体的には、人間のデータが 93 人、ロボットのデータが 87 人であり、圧倒的にデータファイルのが好まれています。これを自分で確かめたいときには、'Descriptives' - 'Descriptive Statistics' ボタンをクリックして、JASP に頻度表を求めることができます。しかし、この要約は我々の興味がある疑問にこたえるものではありません。その疑問を明らかにするために、データのより細かい記述が必要でしょう。ここでは `species` /it によって分類された `choices` を見ていきます。すなわち、データをクロス集計する必要があるのです。JASP では、'Frequencies' - 'Contingency Tables' ボタンを使って、'Columns' ボックス内に `species` を、'Rows' ボックス内に `choice` を移動させることでこれができます。この手続きは以下のような表が作成されるはずです：

	ロボット	人間	合計
子犬	13	15	28
お花	30	13	43
データ	44	65	109
合計	87	93	180

ここから、人間の多くがデータファイルを選択している一方で、ロボットは好みが大きく分かれる傾向にあることがわかります。*なぜ*人間がデータファイルを選びがちなのかという疑問はいまは脇に置いておくとして (確かに、これはかなり奇妙に思えますが)、まず最初にデータセット内の人間とロボットによる選択の違いが統計的に有意であるかどうかを判断します。

1.2.1 仮説検定の構築

このデータをどう分析しましょうか？ 具体的には、私の*研究*仮説は「人間とロボットでは答え方が違う」というものなので、「人間とロボットが答え方が同じ」という*帰無*仮説を検証するにはどうすればよいでしょうか？ 前回と同様に、データを記述するための表記法を確立することから始めます。

	ロボット	人間	合計
子犬	O_{11}	O_{12}	R_1
お花	O_{21}	O_{22}	R_2
データ	O_{31}	O_{32}	R_3
合計	C_1	C_2	N

この表記では、O_{ij} は選択を求められたときには i (子犬、お花、データ) という回答を行った j (ロボット、人間) という種族の回答者の数 (観測度数) であるとしています。観測の総数はいつものように N と書きます。最後に、行の合計を R_i (例：R_1 はお花を選択した人の総数)、列の合計を C_j (例：C_1 はロボットの総数) で記述します。[*9]

では次に、帰無仮説がなにをいっているのか考えてみましょう。ロボットと人間が同じような答え方をするのであれば、「ロボットが子犬という」確率は「人間が子犬という」確率と同じであり、他の二つの可能性についても同じであるということです。つまり、P_{ij} を「種 j が i の反応をする確率」とする、帰無仮説は次のようになります：

H_0: 以下のすべてが真である：
$P_{11} = P_{12}$ ("子犬" という確率が同じ),
$P_{21} = P_{22}$ ("お花" という確率が同じ), and
$P_{31} = P_{32}$ ("データ" という確率が同じ).

そして実は、帰無仮説では真の選択確率が選択する人の種に依存していないということを主張しているので、P_i がその確率を指すものとします (例：P_1 は子犬を選択する真の確率)。

次に、適合度検定のときと大部分は同じように、期待度数を計算する必要があります。つまり、各観測度数 O_{ij} に関して、帰無仮説が何を期待しているのかを把握する必要があります。この期待度数は E_{ij} と表現しましょう。今回は、ちょっしたコツが必要です。種 j に属す人数が C_j で選択肢 i を選択する真の確率が (種族に関係なく) P_i であれば、期待度数はちょうど：

$$E_{ij} = C_j \times P_i$$

さて、これですべてといえばそりゃそうなのですが、ある問題があります。適合度検定のときとは違い、帰無仮説は P_i に関する値を特定していません。それはデータから推定 (Chapter **??**) しなければならないのです！ 幸いなことに、これはとても簡単にできます。もし 180 人中 28 人が花を選択したのであれば、花を選ぶ確率は 28/180 であり、おおよそ .16 になります。もしこれを数学的に表現すると、「選択肢 i を選ぶ確率の推定値は行の合計を全サンプルサイズで割ったものである」と

[*9] 技術的メモです。検定を説明した方法では、列の合計は固定されていて (i.e., 研究者はロボット 87 人と人間 93 人を調査しようとしていた)、行の合計は変動する (i.e., 28 人が子犬をえらんでいることがわかった) というふりをしてます。私が持っている数理統計学の教科書 (**Hogg2005**) の用語を使うと、私は技術的には同質性のカイ二乗検定という用語、行列の合計値が実験の結果によって変動するものである場合には独立性のカイ二乗検定という用語を用いるべきです。この本の初稿では、まさにそのようにしていました。しかし、この 2 つのテストは同一のものであることが判明したので、まとめています。

いうことです：

$$\hat{P}_i = \frac{R_i}{N}$$

したがって、期待度数は行の合計と列の合計の積 (ようは掛け算) を観測値の合計で割ったものとして書くことができます:*10

$$E_{ij} = \frac{R_i \times C_j}{N}$$

期待度数の計算方法が分かったので、適合度検定で用いたのと同じ戦略に従って、検定統計量を定義するのは簡単です。実際、それはほとんど *同じ* 統計量です。

r 行と c 行を持つ分割表の場合、X^2 統計量を定義する式は

$$X^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(E_{ij} - O_{ij})^2}{E_{ij}}$$

唯一の違いは、行と列の両方を合計していることを示すために 2 つの総和記号 (i.e., $\sum$) を含めなければいけないところです。

前回と同様に、X^2 の値が大きいことは帰無仮説がデータの記述に不十分であることを示しており、X^2 の値が小さいことは帰無仮説がデータを十分に説明できていることを示します。したがって、もし X^2 が十分に大きければ帰無仮説を棄却したいと思います。

当然、この統計量は χ^2 分布しています。ここで知る必要があるのは関与している自由度の数ですが、それはそれほど難しいことではありません。以前にも言った通り、(大抵) 自由度は分析しているデータポイントの数から制約の数を引いたものに等しいと考えることができます。r 行と c 列の分割表には $r \times c$ の観測度数が含まれるので、これが観測数の合計になります。制約はどうでしょう？ ここでは、少しトリッキーです。答えは常に同じで

$$df = (r-1)(c-1)$$

なんですが、自由度がこの値を *なぜ* とるのかについての説明は実験計画によって異なります。議論の便宜上、ロボット 87 人と人間 93 人を対象に調査するとし (列の合計は実験者が固定)、行の合計は自由に変更できるようにしたとしましょう (行の合計は確率変数)。ここで適用する制約について考えてみましょう。さて、実験者によって列の合計を意図的に固定したので、c のほうに制約があります。しかし、実はそれだけではありません。帰無仮説にいくつかのフリーパラメータがあったことを覚えていますか (すなわち、P_i を推定しなければならなかった)？ これらも同様に重要です。この本ではその理由を説明はしませんが、帰無仮説の全フリーパラメータは追加の制約みたいなもので

*10 技術的には、ここでの E_{ij} は推定値なので、$\hat{E}_{ij}$ と書くべきでしょう。でも、誰もやらないから、私もやりません。ふふ。

す。ではそれらはいくつあるのでしょうか？ ふむ、それらの確率の合計は 1 でなければならないので、それらは $r-1$ になります。なので、自由度はこうなります：

$$\begin{aligned} df &= (\text{観測数}) - (\text{制約の数}) \\ &= (rc) - (c + (r-1)) \\ &= rc - c - r + 1 \\ &= (r-1)(c-1) \end{aligned}$$

代わりに、合計サンプルサイズ N のみが実験者によって固定されていると仮定しましょう。つまり、最初に見た 180 人に質問したところ、87 人がロボットで、93 人が人間であることがわかりました。今回の論法は少し異なるものですが、それでも同じ答えを導き出します。帰無仮説には選択確率と対応する自由パラメータは $r-1$ ですが、種の確率に対応する自由パラメータも*また* $c-1$ になります。ランダムにサンプリングした人がロボットである確率も推定しなければなりませんからね。[*11] 最後に、観測数の合計は N に固定したので、それがもう一つの制約になります。なので、今は rc の観測データと $(c-1)+(r-1)+1$ の制約があります。それは何をもたらすのでしょうか？

$$\begin{aligned} df &= (\text{観測数}) - (\text{制約の数}) \\ &= rc - ((c-1) + (r-1) + 1) \\ &= rc - c - r + 1 \\ &= (r-1)(c-1) \end{aligned}$$

すばらしい。

1.2.2 JASP での検定

オーケイ、検定の仕組みが分かったところで、JASP での実装方法を見ていきましょう。面倒な計算をして長い時間をかけて学んでもらいたいのはやまやまですが、それは意味がないと思ってます。前回のセクションでは長い時間をかけて適合度検定のやり方を紹介しました。独立性の検定は概念的には違いがないので、わざわざ新しいことを学ぶことはありません。その代わり、簡単なやり方をご紹介します。JASP で検定を実行した後（‘Frequencies’ - ‘Contingency Tables’）、JASP の結果ウィンドウで分割表の下を見れば χ^2 統計量が表示されます。

2 d.f. の χ^2 統計量は 10.722 で p-値 = 0.005 です。

簡単でしょう？ JASP に期待度数を見せるようお願いだってできますよ - ‘Cells’ オプションの ‘Counts’ - ‘Expected’ を押せば、分割表に期待度数が出てきます。その際には、効果量があると便利です。‘Statistics’ オプションのチェックボックスから Cramer’s V を選択すれば、Cramer’s V の値である 0.244 が得られます。これについては、もう少し詳しく説明します。

この出力は、結果を書き出すのに十分な情報を与えてくれます：

[*11] 私たちの多くが実生活で心配している問題

> Pearson's χ^2 が種と選択の間の有意な連関を明らかにしました ($\chi^2(2) = 10.7$, $p < .01$)。ロボットは「花が好き」と答える人が多いようですが、人は「データが好き」と答える人が多いようです。

ここでも、データに何が起きているのかを理解するのに役立つようなほんの少しの解釈を提供します。後ほどディスカッションパートにて、背景をもうちょっと説明します。違いを説明するため、後ほど私が言うであろうことを紹介しておきます：

> 人はロボットよりもデータファイルを好むようであるという事実は、いくらか直観に反します。しかし、チャペック 9 の権限を持つ市民は、人間を見つけると殺して解剖するという残念な傾向があるので、文脈上その回答傾向にはある程度意味があります。そのため、潜在的に望ましくない結果を回避するために、人間の参加者が質問に正直に答えなかったというのは十分ありうることです。これは方法論的に、相当な弱点であるといえるでしょう。

これは反応効果の極端な例として分類されるんじゃないかと思います。明らかにこの事例は問題が深刻なものなので、人とロボットの間の好みの違いを理解するための道具としては、この研究にはほとんど価値がありません。しかし、この例こそが統計的に有意な結果を得る (帰無仮説を棄却して対立仮説を支持する) ことと科学的価値のあるものを見つけること (データには大きな方法論的問題があるため、研究仮説にとって興味深い情報は提供しない) の違いを示していることでしょう。

1.2.3 追伸

後になってそのデータは作り物で、私は仕事をせずにただアニメを見ていただけだったということが判明しました。

1.3 連続性の補正

さて、ちょっとした余談です。これまでほんの少しウソをついていました。自由度が 1 のみである場合には、計算方法を少し変える必要があります。以前指摘していたことを思い出しましょう：χ^2 検定は近似に基づいており、特に大きな N で二項分布が正規分布のように見えるようになるという仮定に基づいています。これの問題点は、特に自由度が 1 しかない場合には上手くいかないことが多いです (例：2×2 の分割表で独立性の検定を行うとき)。主な理由は X^2 統計量の真の標本分布は実際には離散的ですが (カテゴリカルデータを扱ってますからね)、χ^2 分布は連続的だからです。

これにより、系統的な問題が生じる可能性があります。特に N が小さく $df = 1$ のとき、適合度統計量は “大きくなりすぎる” 傾向があります。つまり、予想よりも大きい α の値になります (あるいは、p 値が小さくなりすぎます)。

Yates1934 は、適合度統計量を次のように再定義する簡単な補正方法を提案しました：

$$\chi^2 = \sum_i \frac{(|E_i - O_i| - 0.5)^2}{E_i}$$

基本的には、どこでも 0.5 を引いていくだけです。

Yates の論文を読んだ限りでは、補正は基本的にハックです。それは何か原理的な理論に基づいているわけではありません。むしろ、テストの動作を検証し、補正版がうまく機能することを確認しています。JASP では、‘Statistics’ オプションの ‘χ^2 continuity correction’ というチェックボックスで、この補正を指定できます。

1.4 効果量

先に述べたように (Section **??**)、研究者は効果量の大きさを報告するのが一般的になってきています。なので、カイ二乗検定を実施し、それが有意であることがわかったとします。これで、変数間の連関 (独立性の検定) あるいは指定された確率からの逸脱 (適合度検定) が存在することがわかりました。ここでは、効果量を報告しましょう。つまり、連関や逸脱があるとして、それはどの程度の強さなのでしょうか？

報告するために選択可能な指標と、それらを計算可能な道具が存在します。それらのすべてを説明するつもりはないですが、最も一般的に報告されている効果量に焦点を当てます。

デフォルトでは、頻繁に報告されがちな二つの指標は ϕ 統計量と Cramér's V として知られるやや優れたバージョンです。

数学的に、それはとても単純です。ϕ 統計量を計算するために、X^2 の値をサンプルサイズで割って、それの平方根をとるのです：

$$\phi = \sqrt{\frac{X^2}{N}}$$

ここでの考え方は ϕ 統計量は 0 (まったく関連がない) から 1 (完全な関連) の間の範囲をとるというものですが、分割表が 2×2 よりも大きい場合にはうまくいかず、これはとても面倒です。より大きな表の場合、実際には $\phi > 1$ を得ることができますが、これはかなりよくないです。

そこで、これを補正するために、通常は **Cramer1946** によって提案された V 統計量を報告することが好まれます。これは ϕ に対するかなり簡単な調整です。r 行と c 列の分割表がある場合、二つの数のうちでより小さい値となる $k = \min(r, c)$ を定義します。そうすれば、**Cramér's** V 統計量は以下のようになります。

$$V = \sqrt{\frac{X^2}{N(k-1)}}$$

これで完了です。これはとても人気な指標のようです。計算が簡単で、完全にばかげていない答えが得られるからでしょうね。Cramer’s V では、その値は 0 (まったく関連がない) から 1 (完全な関連) までの範囲をとります。

1.5 検定の前提条件

どのような統計検定にも前提条件があり、その条件が満たされているかを確認することは通常良い考え方です。このチャプターでこれまで取り上げてきたカイ二乗検定の場合、前提条件は以下の通りです：

- *期待度数が充分に大きい。*前のセクションで、二項分布は正規分布とよく似ているために生じた χ^2 の標本分布がどのように見えたか覚えていますか？ さて、チャプター **??** で説明したように、これは観測数が十分に大きい場合にのみ成立します。これが実際に意味していることは、期待度数のすべてを適度に大きくする必要があるということです。適度な大きさとはどのくらいなのでしょうか？ 意見が分かれますが、標準的な前提は期待度数のすべてを約 5 以上であることが望ましいとされているようです。ただし、より大きな表の場合は期待度数の少なくとも 8 割が 5 以上で、1 以下のものがなければ問題ないでしょう。しかし、私が調べた限りではそれらは大まかなガイドラインとして提案されている (**Cochran1954**) だけで、厳格なルールではなく、やや保守的なものなようです (**Larntz1978**)。

- *データは互いに独立している。*カイ二乗検定のやや隠れた前提条件の一つは、観測値が独立していると純粋に信じなければならないことです。これが私の言いたいことです。特定の病院で生まれた男の子の割合に関心があるとします。産科病棟を歩き回り、20 人の女の子と 10 人の男の子を観察します。かなり説得力のある違いじゃないですか？ しかし後になって、実は私が同じ病棟を 10 回分ただ歩き回っていたことがわかり、実際には女の子 2 人と男の子 1 人しか見ていなかったみたいです。そうなると説得力に欠けますよね？ 私が最初に行った 30 回分の*観測*はほとんどが独立していないものであり、実際には 3 回分の独立した観測を行っただけのようです。明らかにこれは極端な (そして非常にばかばかしい) 例ですが、基本的な問題点を示しています。非独立性は “物事を難しくする” のです。愚かな病院の例で示したように、帰無仮説を誤って棄却を引き起こす場合もあればその逆の場合もあります。もう少し馬鹿げた例を挙げると、もし私がカード実験をちょっと違うやり方で行ったときにどうなるかを考えてみましょう。200 人がランダムにカードを 1 枚選ぶのを想像する代わりに、50 人に 4 枚のカードを選ぶようお願いしたとします。一つの可能性は、全員がハート、クラブ、ダイヤ、スペードを一つずつ選択することです (“代表制ヒューリスティック”; Tversky & Kahneman 1974)。これはきわめてランダムでない行動ですが、この場合 4 つのマークに関する観測頻度は 50 になるはずです。この例では、観測が独立してないという事実は (4 枚のカード選択だとそれらが互いに関連するため) 逆効果をもたらし、誤って帰無仮説を保持してしまうことになります。

もし独立性の仮定が満たされない場合に、McNemar 検定や Cochran 検定のようなノンパラメトリック検定を用いることができるかもしれません。同じように、もしセル内の期待度数が小さすぎるのであれば、Fisher 正確性検定を確認してください。現時点では、JASP はこれらのテストを実装してませんが、後で確認してみてください！ 今のところ、これらのテストが存在するという言及のみにとどめますが、その説明は本書の範囲を超えています。

1.6 Summary

この章で説明する重要なアイディアは：

- さまざまなカテゴリーの観測度数に関する表があり、帰無仮説がそれらを比較するための “既知” の確率を提供している場合に、χ^2 (カイ二乗) 適合度検定 (Section **??**) が使用されます。
- χ^2 (カイ二乗) 独立性の検定 (Section **??**) は二つのカテゴリカル変数を分割表 (クロス集計) があるときに使用されます。帰無仮説は、変数の間に関係性や連関がないというものになります。

- 分割表に関する効果量は様々な方法で測定できる (Section **??**)。特に注目したのは、Cramér's V 統計量です。
- 二種類の Pearson の検定は二つの前提に依拠している：期待度数が十分に大きく、観測が独立していること (Section **??**)。ある種の独立性や計数性の違反に対して、様々なノンパラメトリック検定が使用できます。

もしあなたがカテゴリカルデータ分析についてもっと勉強したいのであれば、タイトルが示すように *Introduction to Categorical Data Analysis* **Agresti1996** が最初の選択肢として良いでしょう。もし入門書では物足りない (あるいは取り組んでいる問題を解決できそうもない) 場合には、**Agresti2002** の *Categorical Data Analysis* はいかがでしょうか。後者はより高度なテキストなので、本書からいきなりそっちにいってしまうのはちょっときついかもしれません。