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。

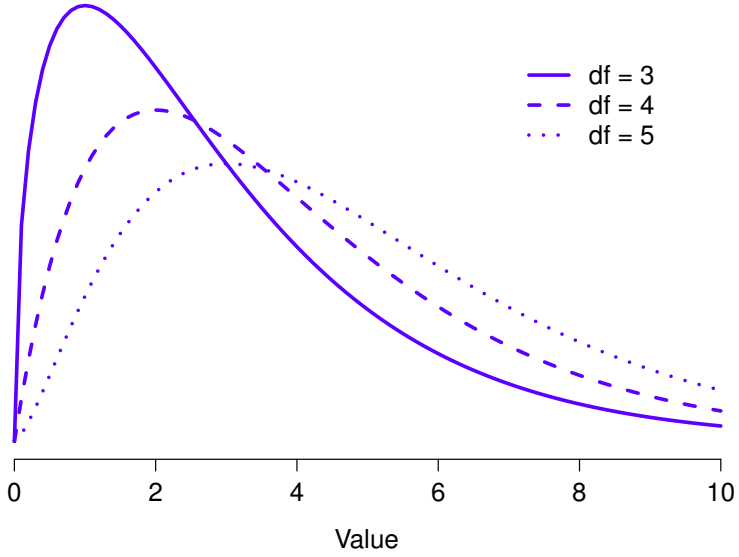

Figure1.1　「自由度」の異なる χ^2 (カイ二乗) 分布。

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

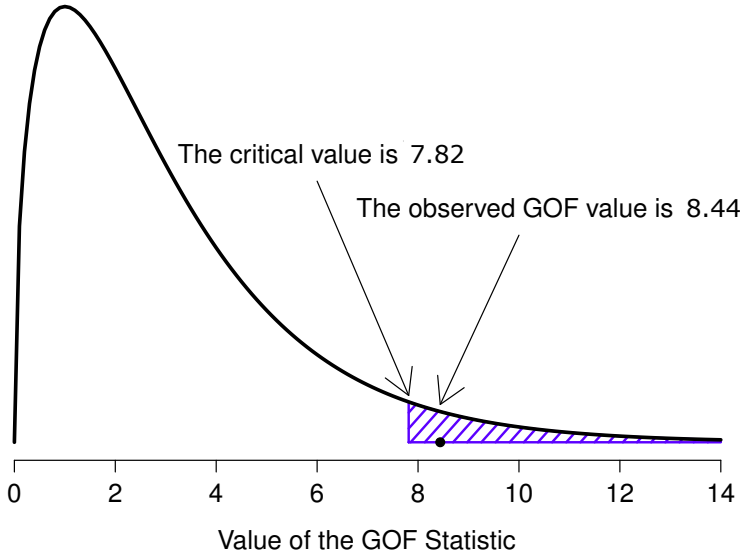

Figure1.2　χ^2 (カイ二乗) 適合度検定で仮説検定がどのように機能するのかを示す図。

. . . . . .

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

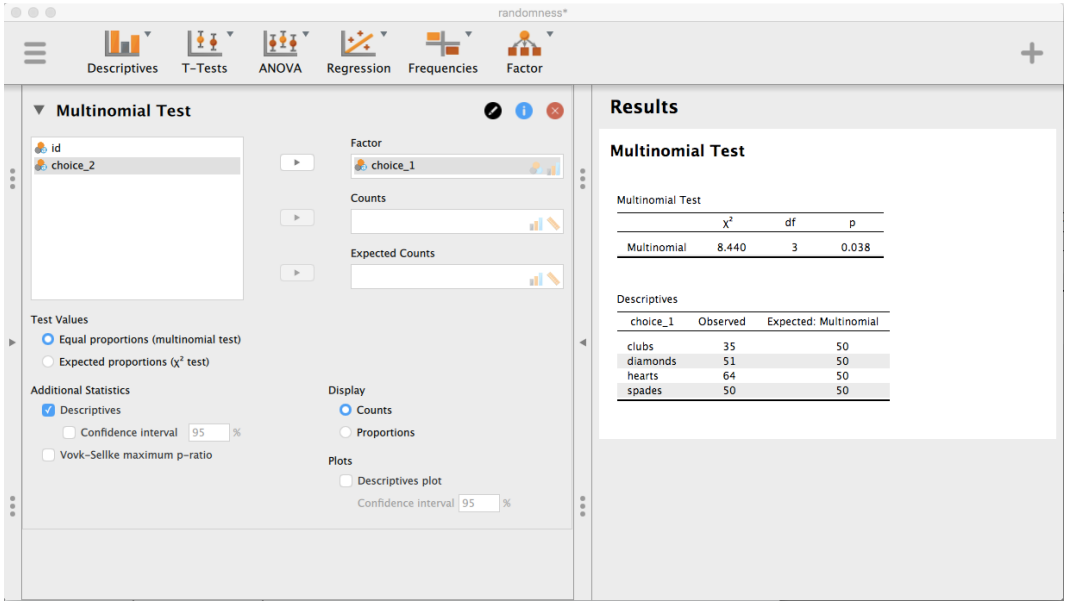

Figure1.4　JASP の χ^2 適合度検定で観測・期待度数を示しています。

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。

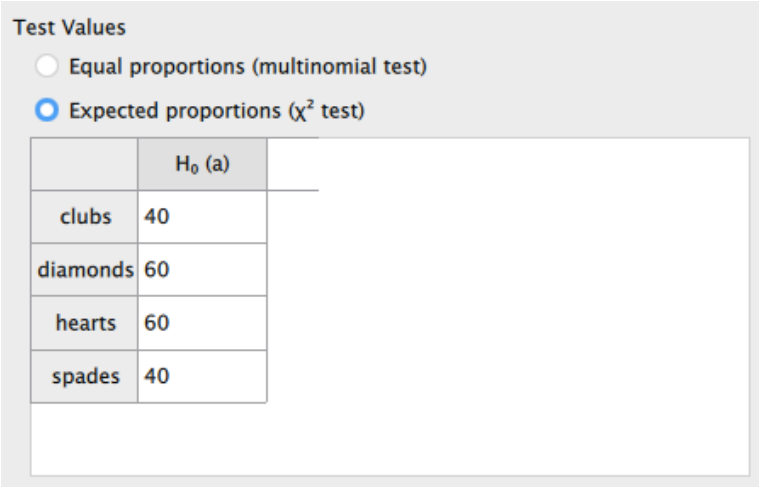

Figure1.5　JASP での χ^2 適合度検定における期待割合の変更

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が

[*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。

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、人生はそう

[*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。

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 The χ^2 test of independence (or association)

GUARDBOT 1: *Halt!*
GUARDBOT 2: *Be you robot or human?*
LEELA: *Robot...we be.*
FRY: *Uh, yup! Just two robots out roboting it up! Eh?*
GUARDBOT 1: *Administer the test.*
GUARDBOT 2: *Which of the following would you most prefer?*
A: A puppy, B: A pretty flower from your sweetie,
or C: A large properly-formatted data file?
GUARDBOT 1: *Choose!*
– Futurama, "Fear of a Bot Planet"

The other day I was watching an animated documentary examining the quaint customs of the natives of the planet *Chapek 9*. Apparently, in order to gain access to their capital city a visitor must prove that they're a robot, not a human. In order to determine whether or not a visitor is human, the natives ask whether the visitor prefers puppies, flowers, or large, properly formatted data files. "Pretty clever," I thought to myself "but what if humans and robots have the same

preferences? That probably wouldn't be a very good test then, would it?" As it happens, I got my hands on the testing data that the civil authorities of *Chapek 9* used to check this. It turns out that what they did was very simple. They found a bunch of robots and a bunch of humans and asked them what they preferred. I saved their data in a file called `chapek9.csv`, which we can now load into JASP. As well as the `ID` variable that identifies individual people, there are two nominal text variables, `species` and `choice`. In total there are 180 entries in the data set, one for each person (counting both robots and humans as "people") who was asked to make a choice. Specifically, there are 93 humans and 87 robots, and overwhelmingly the preferred choice is the data file. You can check this yourself by asking JASP for Frequency Tables, under the 'Descriptives' - 'Descriptive Statistics' button. However, this summary does not address the question we're interested in. To do that, we need a more detailed description of the data. What we want to do is look at the `choices` broken down *by* `species`. That is, we need to cross-tabulate the data. In JASP we do this using the 'Frequencies' - 'Contingency Tables' button, moving `species` into the 'Columns' box and `choice` into the 'Rows' box. This procedure should produce a table similar to this:

	Robot	Human	Total
Puppy	13	15	28
Flower	30	13	43
Data	44	65	109
Total	87	93	180

From this, it's quite clear that the vast majority of the humans chose the data file, whereas the robots tended to be a lot more even in their preferences. Leaving aside the question of *why* the humans might be more likely to choose the data file for the moment (which does seem quite odd, admittedly), our first order of business is to determine if the discrepancy between human choices and robot choices in the data set is statistically significant.

1.2.1 Constructing our hypothesis test

How do we analyse this data? Specifically, since my *research* hypothesis is that "humans and robots answer the question in different ways", how can I construct a test of the *null* hypothesis that "humans and robots answer the question the same way"? As before, we begin by establishing some notation to describe the data:

	Robot	Human	Total
Puppy	O_{11}	O_{12}	R_1
Flower	O_{21}	O_{22}	R_2
Data	O_{31}	O_{32}	R_3
Total	C_1	C_2	N

In this notation we say that O_{ij} is a count (observed frequency) of the number of respondents that are of species j (robots or human) who gave answer i (puppy, flower or data) when asked to make a choice. The total number of observations is written N, as usual. Finally, I've used R_i to denote the row totals (e.g., R_1 is the total number of people who chose the flower), and C_j to denote the column totals (e.g., C_1 is the total number of robots).[*9]

So now let's think about what the null hypothesis says. If robots and humans are responding in the same way to the question, it means that the probability that "a robot says puppy" is the same as the probability that "a human says puppy", and so on for the other two possibilities. So, if we use P_{ij} to denote "the probability that a member of species j gives response i" then our null hypothesis is that:

H_0: All of the following are true:
$P_{11} = P_{12}$ (same probability of saying "puppy"),
$P_{21} = P_{22}$ (same probability of saying "flower"), and
$P_{31} = P_{32}$ (same probability of saying "data").

And actually, since the null hypothesis is claiming that the true choice probabilities don't depend on the species of the person making the choice, we can let P_i refer to this probability, e.g., P_1 is the true probability of choosing the puppy.

Next, in much the same way that we did with the goodness-of-fit test, what we need to do is calculate the expected frequencies. That is, for each of the observed counts O_{ij}, we need to figure out what the null hypothesis would tell us to expect. Let's denote this expected frequency by E_{ij}. This time, it's a little bit trickier. If there are a total of C_j people that belong to species j, and the true probability of anyone (regardless of species) choosing option i is P_i, then the expected frequency is just:

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

[*9] A technical note. The way I've described the test pretends that the column totals are fixed (i.e., the researcher intended to survey 87 robots and 93 humans) and the row totals are random (i.e., it just turned out that 28 people chose the puppy). To use the terminology from my mathematical statistics textbook (**Hogg2005**), I should technically refer to this situation as a chi-square test of homogeneity and reserve the term chi-square test of independence for the situation where both the row and column totals are random outcomes of the experiment. In the initial drafts of this book that's exactly what I did. However, it turns out that these two tests are identical, and so I've collapsed them together.

Now, this is all very well and good, but we have a problem. Unlike the situation we had with the goodness-of-fit test, the null hypothesis doesn't actually specify a particular value for P_i. It's something we have to estimate (Chapter **??**) from the data! Fortunately, this is pretty easy to do. If 28 out of 180 people selected the flowers, then a natural estimate for the probability of choosing flowers is 28/180, which is approximately .16. If we phrase this in mathematical terms, what we're saying is that our estimate for the probability of choosing option i is just the row total divided by the total sample size:

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

Therefore, our expected frequency can be written as the product (i.e. multiplication) of the row total and the column total, divided by the total number of observations:[*10]

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

Now that we've figured out how to calculate the expected frequencies, it's straightforward to define a test statistic, following the exact same strategy that we used in the goodness-of-fit test. In fact, it's pretty much the *same* statistic.

For a contingency table with r rows and c columns, the equation that defines our X^2 statistic is

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

The only difference is that I have to include two summation signs (i.e., $\sum$) to indicate that we're summing over both rows and columns.

As before, large values of X^2 indicate that the null hypothesis provides a poor description of the data, whereas small values of X^2 suggest that it does a good job of accounting for the data. Therefore, just like last time, we want to reject the null hypothesis if X^2 is too large.

Not surprisingly, this statistic is χ^2 distributed. All we need to do is figure out how many degrees of freedom are involved, which actually isn't too hard. As I mentioned before, you can (usually) think of the degrees of freedom as being equal to the number of data points that you're analysing, minus the number of constraints. A contingency table with r rows and c columns contains a total of $r \times c$ observed frequencies, so that's the total number of observations. What about the constraints? Here, it's slightly trickier. The answer is always the same

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

[*10] Technically, E_{ij} here is an estimate, so I should probably write it $\hat{E}_{ij}$. But since no-one else does, I won't either.

but the explanation for *why* the degrees of freedom takes this value is different depending on the experimental design. For the sake of argument, let's suppose that we had honestly intended to survey exactly 87 robots and 93 humans (column totals fixed by the experimenter), but left the row totals free to vary (row totals are random variables). Let's think about the constraints that apply here. Well, since we deliberately fixed the column totals by Act of Experimenter, we have c constraints right there. But, there's actually more to it than that. Remember how our null hypothesis had some free parameters (i.e., we had to estimate the P_i values)? Those matter too. I won't explain why in this book, but every free parameter in the null hypothesis is rather like an additional constraint. So, how many of those are there? Well, since these probabilities have to sum to 1, there's only $r - 1$ of these. So our total degrees of freedom is:

$$\begin{aligned} df &= \text{(number of observations)} - \text{(number of constraints)} \\ &= (rc) - (c + (r - 1)) \\ &= rc - c - r + 1 \\ &= (r - 1)(c - 1) \end{aligned}$$

Alternatively, suppose that the only thing that the experimenter fixed was the total sample size N. That is, we quizzed the first 180 people that we saw and it just turned out that 87 were robots and 93 were humans. This time around our reasoning would be slightly different, but would still lead us to the same answer. Our null hypothesis still has $r - 1$ free parameters corresponding to the choice probabilities, but it now *also* has $c - 1$ free parameters corresponding to the species probabilities, because we'd also have to estimate the probability that a randomly sampled person turns out to be a robot.[*11] Finally, since we did actually fix the total number of observations N, that's one more constraint. So, now we have rc observations, and $(c-1)+(r-1)+1$ constraints. What does that give?

$$\begin{aligned} df &= \text{(number of observations)} - \text{(number of constraints)} \\ &= rc - ((c - 1) + (r - 1) + 1) \\ &= rc - c - r + 1 \\ &= (r - 1)(c - 1) \end{aligned}$$

Amazing.

1.2.2 Doing the test in JASP

Okay, now that we know how the test works let's have a look at how it's done in JASP. As tempting as it is to lead you through the tedious calculations so that you're forced to learn it the long way, I figure there's no point. I already showed you how to do it the long way for the

[*11] A problem many of us worry about in real life.

goodness-of-fit test in the last section, and since the test of independence isn't conceptually any different, you won't learn anything new by doing it the long way. So instead I'll go straight to showing you the easy way. After you have run the test in JASP ('Frequencies' - 'Contingency Tables'), all you have to do is look underneath the contingency table in the JASP results window and there is the χ^2 statistic for you. This shows a χ^2 statistic value of 10.722, with 2 d.f. and p-value = 0.005.

That was easy, wasn't it! You can also ask JASP to show you the expected counts - just click on the check box for 'Counts' - 'Expected' in the 'Cells' options and the expected counts will appear in the contingency table. And whilst you are doing that, an effect size measure would be helpful. We'll choose Cramer's V, and you can specify this from a check box in the 'Statistics' options, and it gives a value for Cramer's V of 0.244. We will talk about this some more in just a moment.

This output gives us enough information to write up the result:

> Pearson's χ^2 revealed a significant association between species and choice ($\chi^2(2) = 10.7$, $p < .01$). Robots appeared to be more likely to say that they prefer flowers, but the humans were more likely to say they prefer data.

Notice that, once again, I provided a little bit of interpretation to help the human reader understand what's going on with the data. Later on in my discussion section I'd provide a bit more context. To illustrate the difference, here's what I'd probably say later on:

> The fact that humans appeared to have a stronger preference for raw data files than robots is somewhat counter-intuitive. However, in context it makes some sense, as the civil authority on Chapek 9 has an unfortunate tendency to kill and dissect humans when they are identified. As such it seems most likely that the human participants did not respond honestly to the question, so as to avoid potentially undesirable consequences. This should be considered to be a substantial methodological weakness.

This could be classified as a rather extreme example of a reactivity effect, I suppose. Obviously, in this case the problem is severe enough that the study is more or less worthless as a tool for understanding the difference preferences among humans and robots. However, I hope this illustrates the difference between getting a statistically significant result (our null hypothesis is rejected in favour of the alternative), and finding something of scientific value (the data tell us nothing of interest about our research hypothesis due to a big methodological flaw).

1.2.3 Postscript

I later found out the data were made up, and I'd been watching cartoons instead of doing work.

1.3

The continuity correction

Okay, time for a little bit of a digression. I've been lying to you a little bit so far. There's a tiny change that you need to make to your calculations whenever you only have 1 degree of freedom. It's called the "continuity correction", or sometimes the **Yates correction**. Remember what I pointed out earlier: the χ^2 test is based on an approximation, specifically on the assumption that the binomial distribution starts to look like a normal distribution for large N. One problem with this is that it often doesn't quite work, especially when you've only got 1 degree of freedom (e.g., when you're doing a test of independence on a 2×2 contingency table). The main reason for this is that the true sampling distribution for the X^2 statistic is actually discrete (because you're dealing with categorical data!) but the χ^2 distribution is continuous. This can introduce systematic problems. Specifically, when N is small and when $df = 1$, the goodness-of-fit statistic tends to be "too big", meaning that you actually have a bigger α value than you think (or, equivalently, the p values are a bit too small).

Yates1934 suggested a simple fix, in which you redefine the goodness-of-fit statistic as:

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

Basically, he just subtracts off 0.5 everywhere.

As far as I can tell from reading Yates' paper, the correction is basically a hack. It's not derived from any principled theory. Rather, it's based on an examination of the behaviour of the test, and observing that the corrected version seems to work better. You can specify this correction in JASP from a check box in the 'Statistics' options, where it is called 'χ^2 continuity correction'.

1.4

Effect size

As we discussed earlier (Section **??**), it's becoming commonplace to ask researchers to report some measure of effect size. So, let's suppose that you've run your chi-square test, which turns out to be significant. So you now know that there is some association between your variables (independence test) or some deviation from the specified probabilities (goodness-of-fit test). Now you want to report a measure of effect size. That is, given that there is an association or deviation, how strong is it?

There are several different measures that you can choose to report, and several different tools that you can use to calculate them. I won't discuss all of them but will instead focus on the most commonly reported measures of effect size.

By default, the two measures that people tend to report most frequently are the ϕ statistic and the somewhat superior version, known as Cramér's V.

Mathematically, they're very simple. To calculate the ϕ statistic, you just divide your X^2 value by the sample size, and take the square root:

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

The idea here is that the ϕ statistic is supposed to range between 0 (no association at all) and 1 (perfect association), but it doesn't always do this when your contingency table is bigger than 2×2, which is a total pain. For bigger tables it's actually possible to obtain $\phi > 1$, which is pretty unsatisfactory. So, to correct for this, people usually prefer to report the V statistic proposed by **Cramer1946**. It's a pretty simple adjustment to ϕ. If you've got a contingency table with r rows and c columns, then define $k = \min(r, c)$ to be the smaller of the two values. If so, then **Cramér's** V statistic is

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

And you're done. This seems to be a fairly popular measure, presumably because it's easy to calculate, and it gives answers that aren't completely silly. With Cramer' s V, you know that the value really does range from 0 (no association at all) to 1 (perfect association).

1.5

Assumptions of the test(s)

All statistical tests make assumptions, and it's usually a good idea to check that those assumptions are met. For the chi-square tests discussed so far in this chapter, the assumptions are:

- *Expected frequencies are sufficiently large*. Remember how in the previous section we saw that the χ^2 sampling distribution emerges because the binomial distribution is pretty similar to a normal distribution? Well, like we discussed in Chapter **??** this is only true when the number of observations is sufficiently large. What that means in practice is that all of the expected frequencies need to be reasonably big. How big is reasonably big? Opinions differ, but the default assumption seems to be that you generally would like to see all your expected frequencies larger than about 5, though for larger tables you would probably be okay if at least 80% of the the expected frequencies are above 5 and none of them are below 1. However, from what I've been able to discover (**Cochran1954**) these seem to have been proposed as rough guidelines, not hard and fast rules, and they seem to be somewhat conservative (**Larntz1978**).
- *Data are independent of one another*. One somewhat hidden assumption of the chi-square test is that you have to genuinely believe that the observations are independent. Here's what I mean. Suppose I'm interested in proportion of babies born at a particular hospital that are boys. I walk around the maternity wards and observe 20 girls and only 10 boys. Seems like a pretty convincing difference, right? But later on, it turns out that I'd actually walked into the same ward 10 times and in fact I'd only seen 2 girls and 1 boy. Not as convincing, is it? My original 30 *observations* were massively non-independent, and were only in fact equivalent to 3 independent observations. Obviously this is an extreme (and extremely silly) example, but it illustrates the basic issue. Non-independence "stuffs things up". Sometimes it causes you to falsely reject the null, as the silly hospital example illustrates, but it can go the other way too. To give a slightly less stupid example, let's consider what would happen if I'd done the cards experiment slightly differently Instead of asking 200 people to try to imagine sampling one card at random, suppose I asked 50 people to select 4 cards. One possibility would be that *everyone* selects one heart, one club, one diamond and one spade (in keeping with the "representativeness heuristic"; Tversky & Kahneman 1974). This is highly non-random behaviour from people, but in this case I would get an observed frequency of 50 for all four suits. For this example the fact that the observations are non-independent (because the four cards that you pick will be related to each other) actually leads to the opposite effect, falsely retaining the null.

If you happen to find yourself in a situation where independence is violated, it may be possible to use the nonparametric tests, such as the McNemar test or the Cochran test. Similarly, if your expected cell counts are too small, check out the Fisher exact test. At present, JASP does not implement these tests, but check back later! For now, we'll just mention that these tests exist, but describing them is beyond the scope of this book.

1.6 Summary

The key ideas discussed in this chapter are:

- The χ^2 (chi-square) goodness-of-fit test (Section **??**) is used when you have a table of observed frequencies of different categories, and the null hypothesis gives you a set of "known" probabilities to compare them to.
- The χ^2 (chi-square) test of independence (Section **??**) is used when you have a contingency table (cross-tabulation) of two categorical variables. The null hypothesis is that there is no relationship or association between the variables.
- Effect size for a contingency table can be measured in several ways (Section **??**). In particular we noted the Cramér's V statistic.
- Both versions of the Pearson test rely on two assumptions: that the expected frequencies are sufficiently large, and that the observations are independent (Section **??**). Various nonparametric tests can be used for certain kinds of violations of independence or count assumptions.

If you're interested in learning more about categorical data analysis a good first choice would be **Agresti1996** which, as the title suggests, provides an *Introduction to Categorical Data Analysis*. If the introductory book isn't enough for you (or can't solve the problem you're working on) you could consider **Agresti2002**, *Categorical Data Analysis*. The latter is a more advanced text, so it's probably not wise to jump straight from this book to that one.