OpenIntro

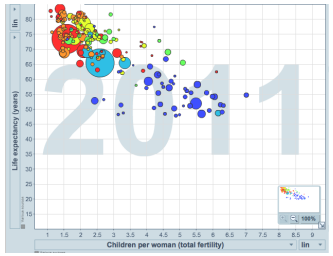
第 2 章：データの要約

OpenIntro Statistics 第 4 版（日本語版）

原著スライド：Mine Çetinkaya-Rundel（OpenIntro）

散布図

散布図は 2 つの数値変数間の関係を可視化するのに有用である。



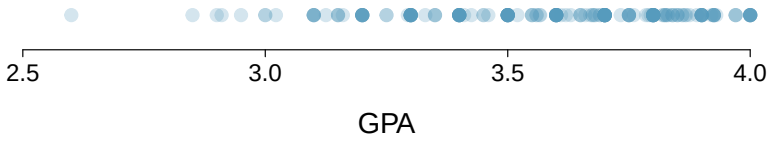
平均寿命と合計特殊出生率は関連があるか、それとも独立か？

この関係は年代を通じて変わらなかったか、それとも変化したか？

http://www.gapminder.org/world

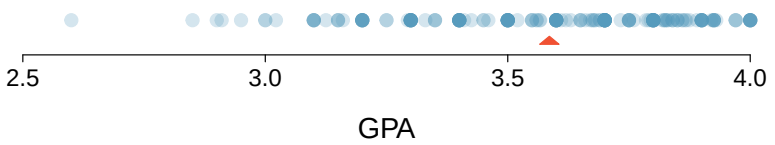
ドットプロット

1 つの数値変数を可視化するのに有用。色が濃い部分はより多くの観測値が集中している。



このデータセットの GPA の分布をどのように説明するか？ 分布の中心・形状・散らばりについて述べること。

ドットプロットと平均



- 平均（上のプロットで三角形で示されている）は、データの分布の中心を測る方法の 1 つである。
- GPA の平均は 3.59 である。

平均

- 標本平均（$\bar{x}$ で表す）は次のように計算できる：

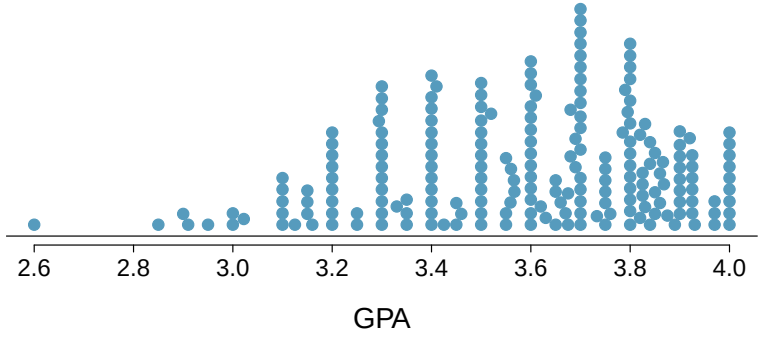
$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n},$$

ここで $x_1, x_2, \cdots, x_n$ は n 個の観測値を表す。

- 母平均も同様に計算されるが μ で表される。母集団データが得られることはまれなため、μ を計算することはしばしば不可能である。
- 標本平均は標本統計量であり、母平均の点推定値として機能する。この推定は完璧ではないかもしれないが、標本が良好（母集団を代表している）であれば、通常は十分な推定値となる。

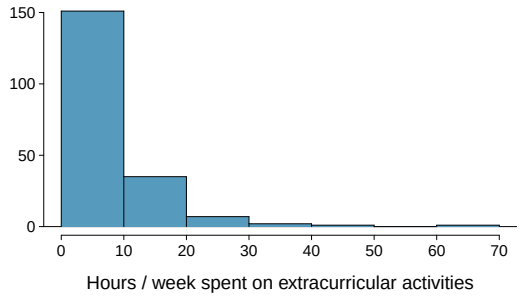
積み上げドットプロット

棒が高い部分はより多くの観測値が集中しており、分布の中心と形状をやや判断しやすくなる。



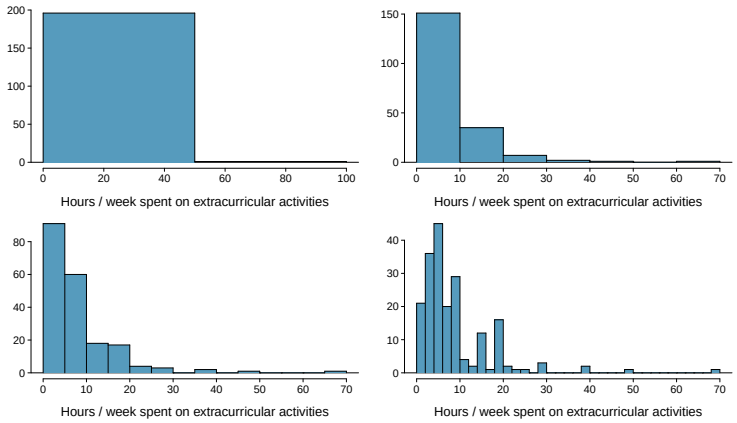
ヒストグラム—課外活動時間

- ヒストグラムはデータの密度を表す。棒が高い部分はデータが相対的に多い。
- ヒストグラムはデータ分布の形状を説明するのに特に便利である。
- 選択する階級幅によってヒストグラムが示す内容が変わる。



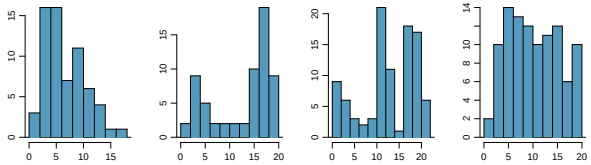
階級幅

これらのヒストグラムのうち有用なものはどれか？ データを明かしすぎているのはどれか？ 隠しすぎているのはどれか？



分布の形状：最頻値の個数

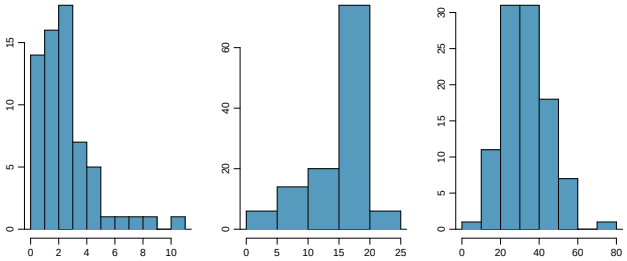
ヒストグラムに１つの顕著な山（単峰型）、複数の顕著な山（二峰型・多峰型）、または顕著な山がない（一様型）か？



注：最頻値の個数を判断するには、ヒストグラムの上に滑らかな曲線を描くことを想像しよう——棒を木のブロックと見立て、そこにスパゲッティを垂らしたときの形が滑らかな曲線になる。

分布の形状：歪み

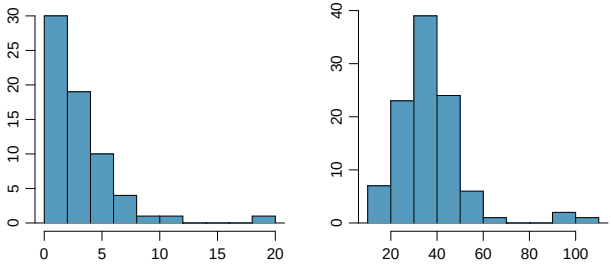
ヒストグラムは右歪みか、左歪みか、それとも対称か？



注：ヒストグラムは長い裾の側に歪んでいると言う。

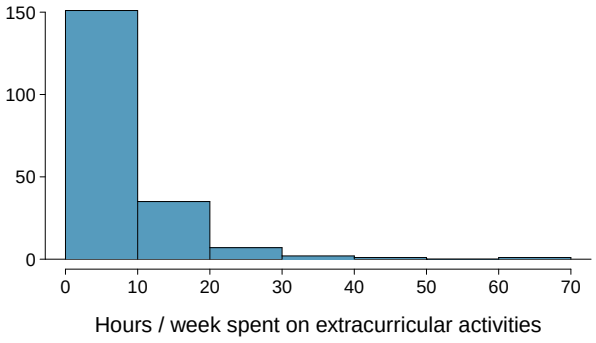
分布の形状：異常値

異常な観測値や潜在的な外れ値はあるか？



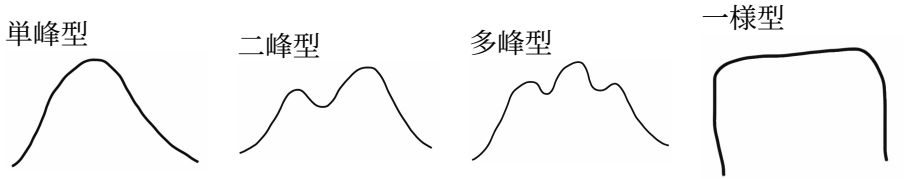
課外活動

学生が課外活動に費やす週当たりの時間の分布の形状をどのように説明するか？

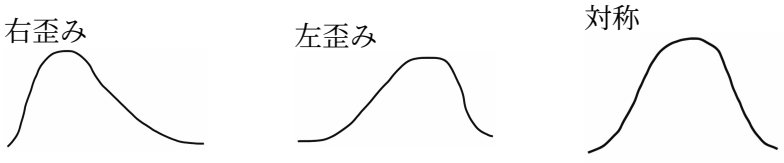


よく見られる分布の形状

- 最頻値の個数



- 歪み



練習問題

次の変数のうち一様分布すると予想されるものはどれか？

(a) 成人女性の体重
(b) ノースカロライナ州の無作為標本の給与
(c) 住宅価格
(d) クラスメートの誕生日（日付）

応用活動：分布の形状

次の変数の予想される分布をスケッチせよ：

- ピアスの数
- 試験の得点
- IQ スコア

任意の変数の予想分布を判断する方法を簡潔に（1〜2 文で）説明せよ。

あなたは「典型的な」人か？

http://www.youtube.com/watch?v=4B2xOvKFFz4

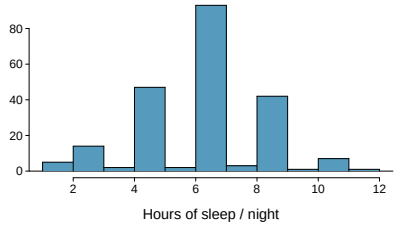
分布の真の特徴を伝えるうえで、中心だけがどれほど有用か？

分散

分散は平均からの偏差の 2 乗のおおよその平均である。

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

- 標本平均は $\bar{x} = 6.71$、標本サイズは $n = 217$ である。
- 学生が 1 晩に睡眠をとる時間の分散は次のように計算できる：



$$s^2 = \frac{(5-6.71)^2 + (9-6.71)^2 + \cdots + (7-6.71)^2}{217-1} = 4.11 \text{ 時間}^2$$

分散（続き）

なぜ分散の計算に偏差の 2 乗を使うのか？

標準偏差

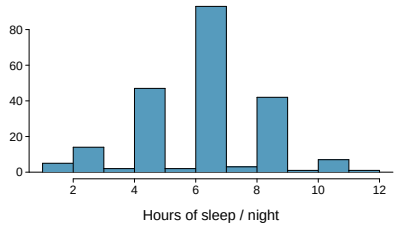
標準偏差は分散の平方根であり、データと同じ単位を持つ。

$$s = \sqrt{s^2}$$

- 学生が 1 晩に睡眠をとる時間の標準偏差は次のように計算できる：

$$s = \sqrt{4.11} = 2.03 \text{ 時間}$$

- すべてのデータが平均の 3 標準偏差以内にあることが分かる。



中央値

- 中央値は、データを昇順に並べたときに真ん中を 2 分割する値である。

$$0, 1, \textit{2}, 3, 4$$

- 観測値の数が偶数の場合、中央値は真ん中 2 つの値の平均である。

$$0, 1, \underline{2, 3}, 4, 5 \rightarrow \frac{2+3}{2} = \textit{2.5}$$

- 中央値はデータの中間点であるため、値の 50%がその下にある。したがって、第 *50* パーセンタイルとも呼ばれる。

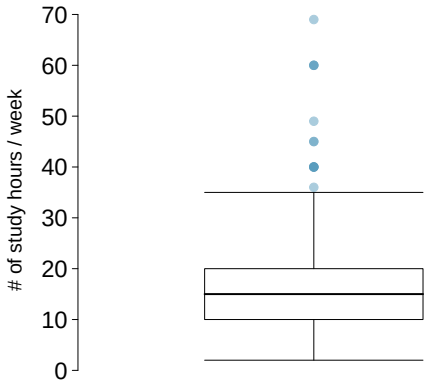
Q1、Q3、IQR

- 第 25 パーセンタイルは第 *1* 四分位数（*Q1*）とも呼ばれる。
- 第 50 パーセンタイルは中央値とも呼ばれる。
- 第 75 パーセンタイルは第 *3* 四分位数（*Q3*）とも呼ばれる。
- Q1 と Q3 の間にはデータの中央 50%がある。このデータが示す範囲を四分位範囲（*IQR*）と呼ぶ。

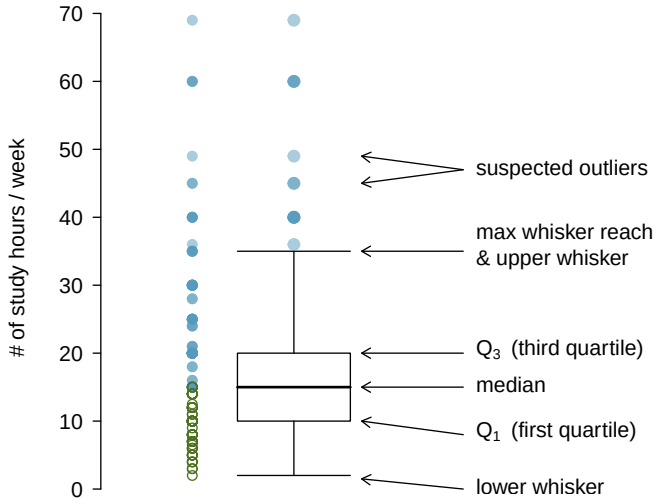
$$IQR = Q3 - Q1$$

箱ひげ図

箱ひげ図の箱はデータの中央 50%を表し、箱内の太線は中央値である。



箱ひげ図の構造



ひげと外れ値

- 箱ひげ図のひげは四分位数から最大 $1.5 \times IQR$ まで伸びる。

$$\text{上ひげの最大リーチ} = Q3 + 1.5 \times IQR$$
$$\text{下ひげの最大リーチ} = Q1 - 1.5 \times IQR$$

$$IQR : 20 - 10 = 10$$
$$\text{上ひげの最大リーチ} = 20 + 1.5 \times 10 = 35$$
$$\text{下ひげの最大リーチ} = 10 - 1.5 \times 10 = -5$$

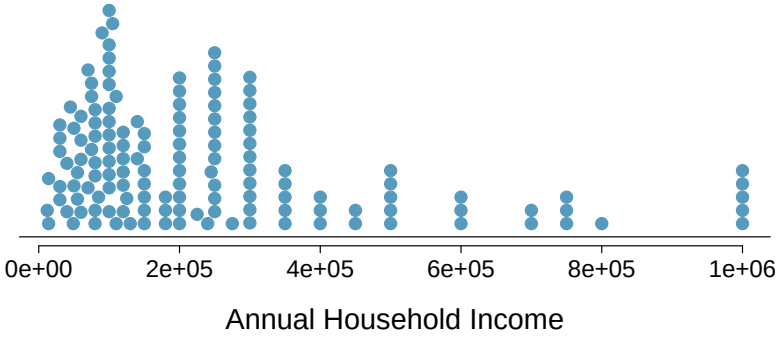
- 潜在的な外れ値とは、ひげの最大リーチを超えた観測値である。残りのデータに対して極端に見える観測値である。

外れ値（続き）

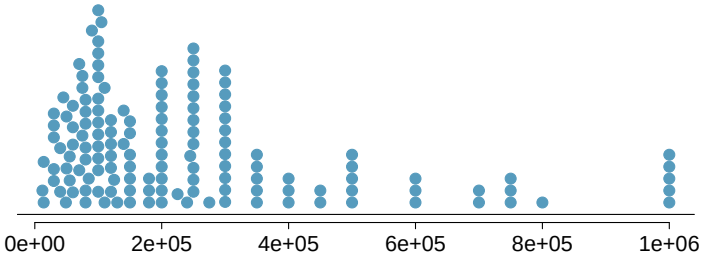
外れ値を探すことがなぜ重要か？

極端な観測値

世帯収入の平均・中央値・標準偏差・IQR は、最大値を 1,000 万ドルに置き換えた場合どのように変化するか？ 最小値を 1,000 万ドルに置き換えた場合はどうか？



頑健な統計量



	頑健		非頑健	
シナリオ	中央値	IQR	$\bar{x}$	s
元のデータ	19 万	20 万	24.5 万	22.6 万
最大値を 1,000 万ドルに変更	19 万	20 万	30.9 万	85.3 万
最小値を 1,000 万ドルに変更	20 万	20 万	31.6 万	85.4 万

頑健な統計量

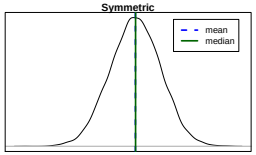
中央値と IQR は、平均と標準偏差よりも歪みや外れ値に対して頑健である。したがって、

- 歪んだ分布では、中心と散らばりを説明するために中央値と IQR を使う方が有用なことが多い
- 対称な分布では、中心と散らばりを説明するために平均と標準偏差を使う方が有用なことが多い

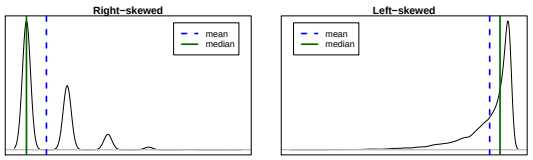
学生の典型的な世帯収入を推定したい場合、平均と中央値のどちらに注目するか？

平均と中央値

- 分布が対称な場合、中心はしばしば平均として定義される：平均 ≈ 中央値

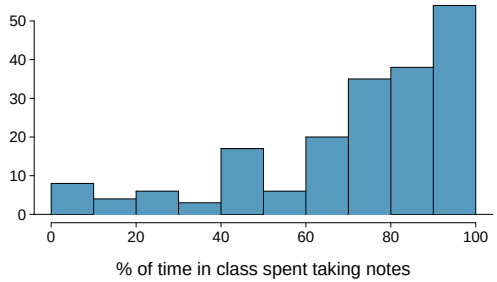


- 分布が歪んでいるか極端な外れ値がある場合、中心はしばしば中央値として定義される
  - 右歪み：平均 > 中央値
  - 左歪み：平均 < 中央値



練習問題

授業中にノートを取ることに費やした時間の割合（Facebook や Twitter などと比較した）の分布について最も正しいと思われるのはどれか？

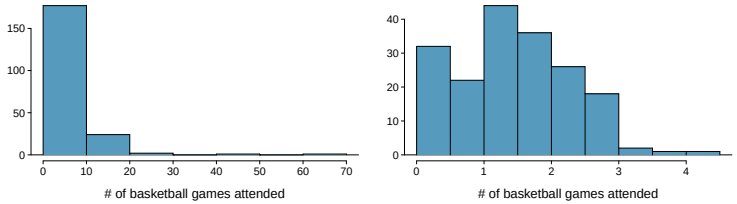


中央値：*80%*
平均：*76%*

(a) 平均 > 中央値
(b) 平均 < 中央値
(c) 平均 ≈ 中央値
(d) 判断不可能

極端に歪んだデータ

データが極端に歪んでいる場合、変換すると分析しやすくなることがある。一般的な変換は対数変換である。
左のヒストグラムは学生が観戦したバスケットボールの試合数の分布を示す。右のヒストグラムは試合数の対数の分布を示す。



変換の利点と欠点

- 歪んだデータは適切な変換によって外れ値が目立たなくなるため、変換後の方が分析しやすい。

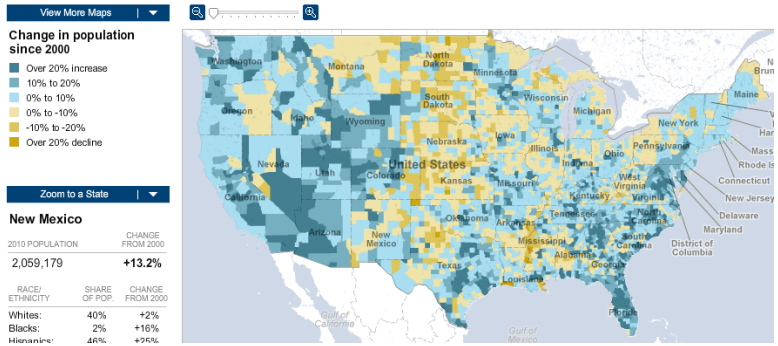
試合数	70	50	25	···
log（試合数）	4.25	3.91	3.22	···

- ただし、測定変数の対数単位での分析結果は解釈が難しいことがある。

他にどのような変数が極端に歪むと予想されるか？

強度マップ

2000 年から 2010 年にかけての人口変化において、どのようなパターンが見られるか？



http://projects.nytimes.com/census/2010/map

分割表

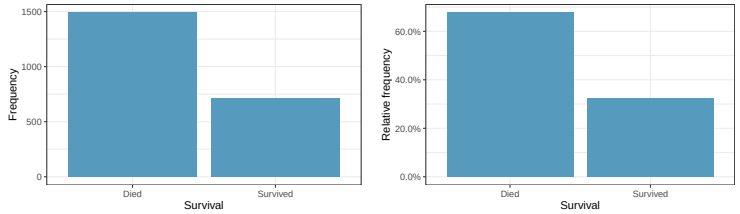
2 つのカテゴリ変数のデータをまとめた表を分割表と呼ぶ。

下の分割表はタイタニック号の乗客の生存と年齢の分布を示す。

		生存		
		死亡	生存	合計
年齢	大人	1438	654	2092
	子供	52	57	109
	合計	1490	711	2201

棒グラフ

棒グラフは 1 つのカテゴリ変数を表示する一般的な方法である。
頻度の代わりに比率を示す棒グラフを相対頻度棒グラフと呼ぶ。



棒グラフとヒストグラムの違いは何か？

適切な比率の選択

タイタニック号の乗客において、年齢と生存の間に関係があるように見えるか？

		生存		
		死亡	生存	合計
年齢	大人	1438	654	2092
	子供	52	57	109
	合計	1490	711	2201

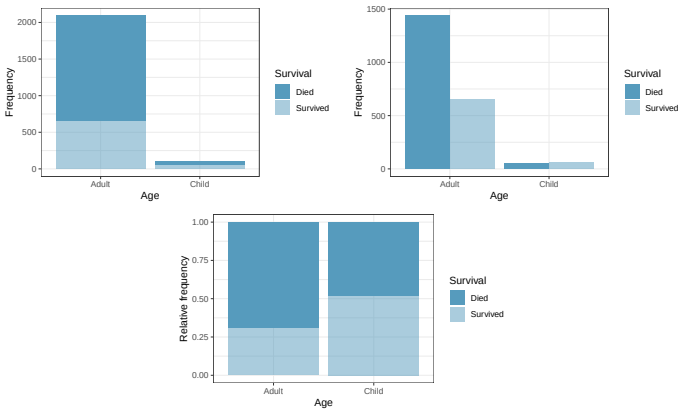
この問いに答えるために行比率を調べる：

- 生存した大人の割合：654 / 2092 ≈ 0.31
- 生存した子供の割合：57 / 109 ≈ 0.52

2変数の棒グラフ

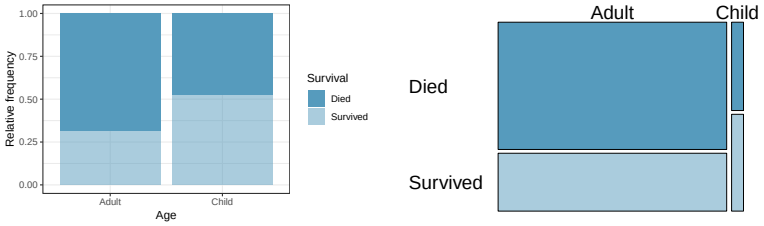
- 積み上げ棒グラフ：分割表の情報を度数で表示するグラフ。
- 並列棒グラフ：同じ情報を棒を積み上げる代わりに横に並べて表示する。
- 正規化積み上げ棒グラフ：分割表の情報を比率で表示するグラフ。

下の3つの図の違いは何か？



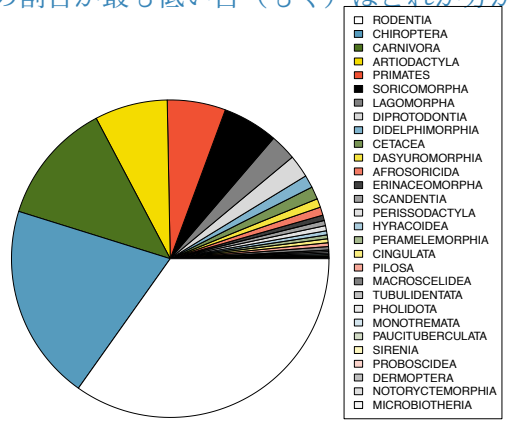
モザイクプロット

下の2つの図の違いは何か？



円グラフ

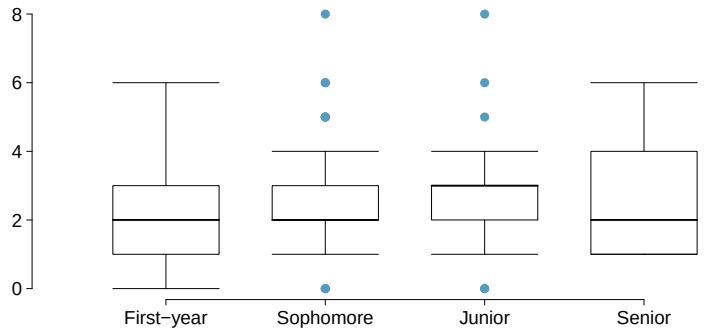
哺乳類の種の割合が最も低い目（もく）はどれか分かるか？



データ出典：*http://www.bucknell.edu/msw3.*

並列箱ひげ図

学年と所属クラブ数の間に関係があるように見えるか？



性差別

- 1972 年、性差別研究の一環として、48 名の男性銀行管理職がそれぞれ同じ人事ファイルを渡され、その人物を「通常業務」と説明された支店長職に昇進させるべきかどうかを判断するよう求められた。
- ファイルは同一だったが、管理職の半数は男性を示すファイルを、残り半数は女性を示すファイルを受け取った。
- どの管理職が「男性」応募書類を受け取り、どの管理職が「女性」応募書類を受け取るかはランダムに決定された。
- 審査された 48 件のファイルのうち、35 件が昇進を推薦された。
- この研究は女性が不当に差別されているかどうかを検証している。

これは観察研究か実験か？

B.Rosen and T. Jerdee (1974), "Influence of sex role stereotypes on personnel decisions", J.Applied Psychology, 59:9-14.

データ

一見して、昇進と性別の間に関係があるように見えるか？

		昇進		
		昇進あり	昇進なし	合計
性別	男性	21	3	24
	女性	14	10	24
	合計	35	13	48

昇進した男性の割合： $21/24 = 0.875$
昇進した女性の割合： $14/24 = 0.583$

練習問題

昇進推薦された男性と女性ファイルの割合に約 30%（正確には 29.2%）の差があることが分かった。この情報に基づいて、以下のうち正しいものはどれか？

(a) 実験を繰り返せば、必ずより多くの女性ファイルが昇進推薦される。今回はまぐれだった。
(b) 昇進は性別に依存しており、男性の方が昇進しやすく、昇進決定において女性への性差別がある。かもしれない
(c) 昇進した男性と女性ファイルの割合の差は偶然によるものであり、昇進決定における女性への性差別の証拠ではない。かもしれない
(d) 女性は男性より能力が低く、そのためより少ない女性が昇進する。

2つの競合する主張

1. 「何も起きていない。」
   昇進と性別は独立であり、性差別はなく、割合の差は単に偶然による。→ 帰無仮説
2. 「何かが起きている。」
   昇進と性別は従属であり、性差別があり、割合の差は偶然ではない。→ 対立仮説

仮説検定としての裁判

- 仮説検定は裁判によく似ている。
- H_0：被告は無実
  H_A：被告は有罪
- 次に証拠を提示する——データを収集する。
- 次に証拠を判断する——「帰無仮説が真であった場合、これらのデータが偶然に起こりうるか？」
  - それが起こる可能性が非常に低い場合、証拠は帰無仮説について合理的な疑いを超える疑念を生じさせる。
- 最終的には決断を下さなければならない。どれほど低い確率を「起こりにくい」と言うのか？

Image from *http://www.nwherald.com/_internal/cimg!0/oo1il4sf8zzaqbboq25oevvbg99wpot.*

仮説検定としての裁判（続き）

- 証拠が無実の仮定を棄却するのに十分でない場合、陪審員は「無罪」の評決を下す。
  - 陪審員は被告が無実とは言わず、有罪にするための証拠が十分でないと言うだけである。
  - 被告は実際に無実かもしれないが、陪審員には確かめる方法がない。
- 統計的には、**帰無仮説を棄却できない**と言う。
  - 帰無仮説が真かどうかを単純に知ることができないため、帰無仮説が真であるとは決して宣言しない。
  - そのため、「帰無仮説を採択する」とは決して言わない。

仮説検定としての裁判（続き）

- 裁判では、証明責任は検察側にある。
- 仮説検定では、証明責任は異常な主張の側にある。
- 帰無仮説は通常の状態（現状）であるため、異常とみなされるのは対立仮説であり、その証拠を集める必要がある。

まとめ：仮説検定の枠組み

- 現状を表す帰無仮説（H_0）から始める。
- また、研究の問い（何を検証するか）を表す対立仮説（H_A）もある。
- 帰無仮説が真であるという仮定のもとで、シミュレーション（今回）または理論的方法（後の講義）によって仮説検定を行う。
- 検定結果が対立仮説を支持する説得力のある証拠を示さない場合は帰無仮説を維持し、示す場合は帰無仮説を棄却して対立仮説を採択する。

実験のシミュレーション…

…独立性の仮定のもと、つまり偶然に委ねる。

偶然モデルに基づくシミュレーションの結果がデータに似ている場合、男女間の昇進ファイルの割合の差は単に偶然によるものと判断できる（昇進と性別は独立）。

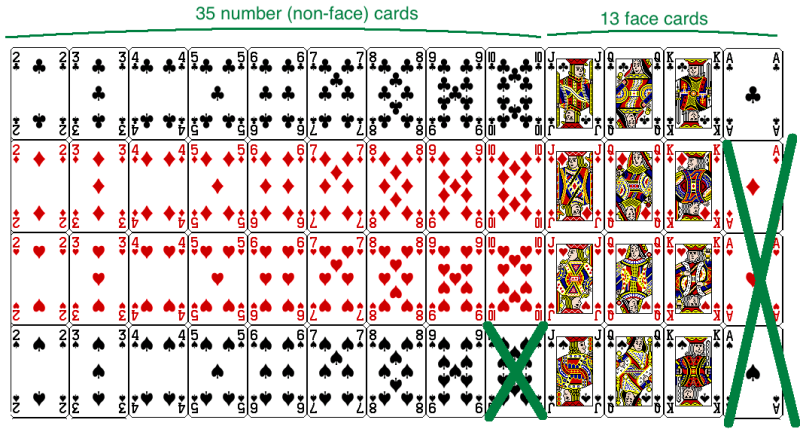
偶然モデルに基づくシミュレーションの結果がデータに似ていない場合、男女間の昇進ファイルの割合の差は偶然ではなく、性別の実際の影響によるものと判断できる（昇進と性別は従属）。

応用活動：実験のシミュレーション

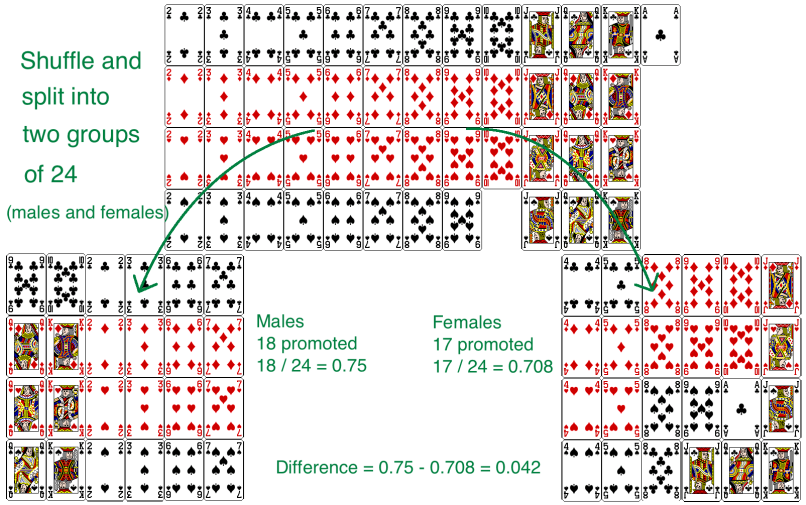
トランプを使ってこの実験をシミュレートせよ。

1. 絵札を「昇進なし」、数札を「昇進あり」とする。エースは絵札として扱う。
   - ジョーカーを取り除く。
   - エース 3 枚を取り除く → デッキに絵札がちょうど 13 枚残る（絵札：A, K, Q, J）。
   - 数札 1 枚を取り除く → デッキに数札がちょうど 35 枚残る（数札：2〜10）。
2. カードをシャッフルし、男性と女性を表す 24 枚ずつの 2 グループに配る。
3. 各グループで昇進推薦された（数札）ファイルの数を数えて記録する。
4. 各グループの昇進推薦ファイルの割合を計算し、差（男性 − 女性）を求めて記録する。
5. ステップ 2〜4 を多数回繰り返す。

ステップ 1



ステップ 2〜4



練習問題

今実施したシミュレーションの結果は、女性への性差別（つまり性別と昇進決定の従属関係）の説得力のある証拠を提供するか？

(a) いいえ。データは対立仮説を支持する説得力のある証拠を提供しないため、性別と昇進決定の独立性という帰無仮説を棄却できない。2 つの割合の差は偶然によるものだった。

(b) はい。データは昇進決定における女性への性差別という対立仮説を支持する説得力のある証拠を提供する。2 つの割合の差は性別の実際の影響によるものだった。

ソフトウェアによるシミュレーション

前述の方法でシミュレーションを実行するのは面倒で時間がかかる。実際にはソフトウェアを使ってシミュレーションを生成する。下のドットプロットは 100 回のシミュレーションに基づく昇進率の差のシミュレーション分布を示す。