

自分で触ってよくわかる

変数分布の可視化の話：

1変数の分布をグラフにする

神戸市立医療センター中央市民病院
臨床研究推進センター

宮越 千智

今回の学習目標

- ✓ ヒストグラムを使って量的変数の分布を可視化できる
- ✓ 箱ひげ図を使って量的変数の分布を可視化できる
- ✓ 棒グラフを使って質的変数の分布を可視化できる

復習:

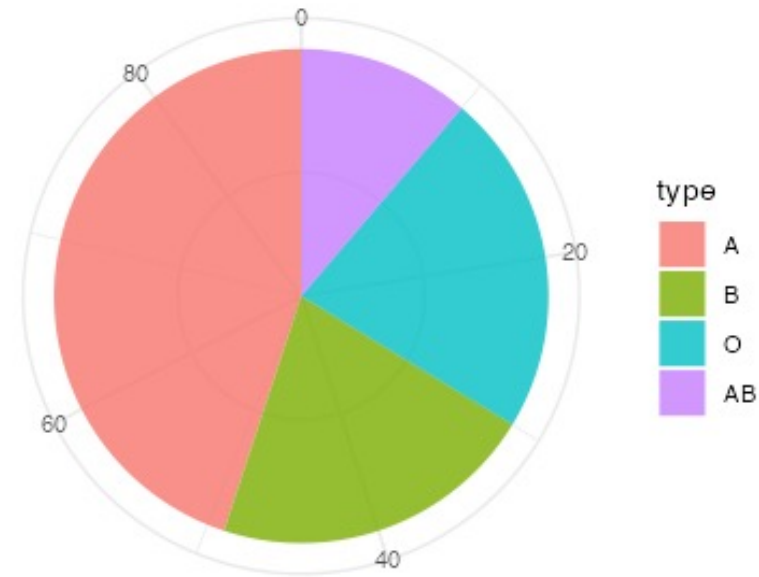
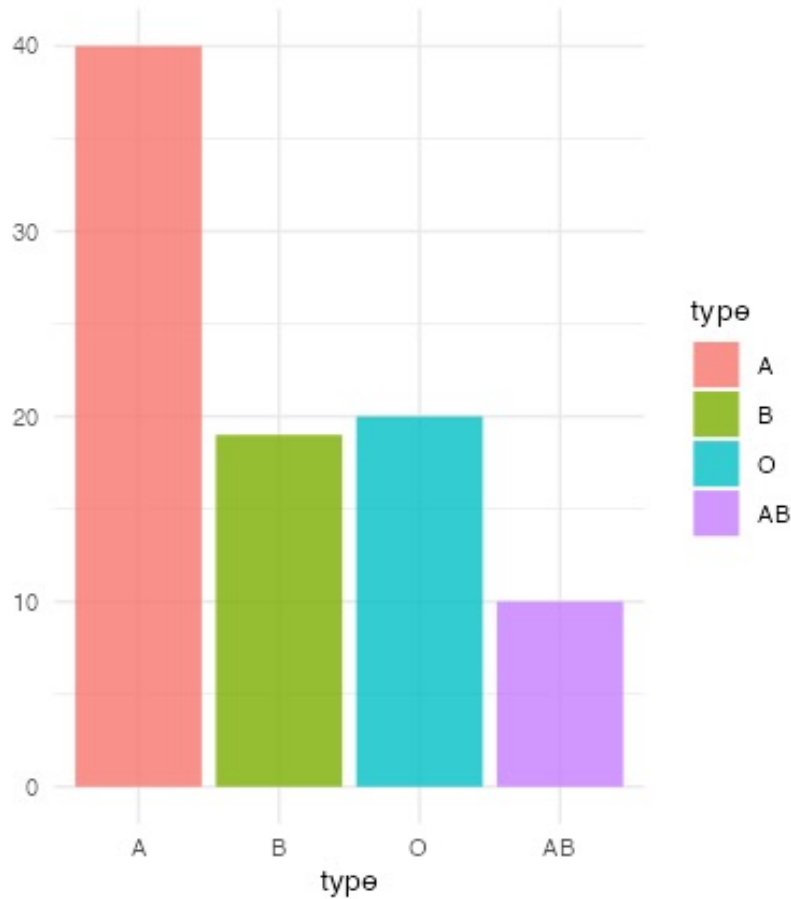
変数の分布を数値で示す方法

変数の種類	示し方	指標		対応する グラフ
質的変数	水準ごとに度数と割合を示す	度数、割合		棒グラフ 円グラフ
量的変数	いくつかの区分に分けて 度数と割合を示す	度数、割合		ヒストグラム
		中心位置	平均値 中央値 最頻値	箱ヒゲ図
		散らばり具合	分散・標準偏差 四分位範囲 範囲	

復習: 質的変数の可視化

棒グラフ(または円グラフ)を用いる

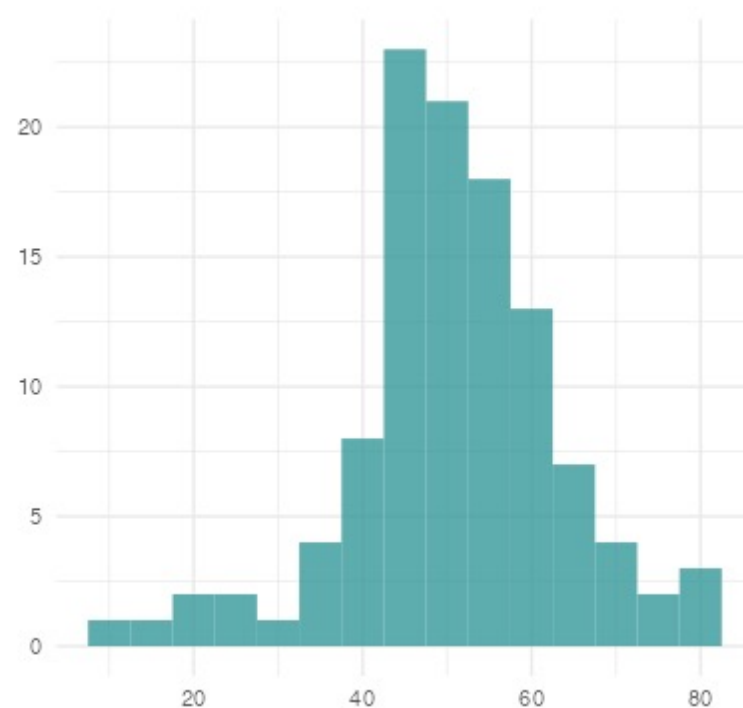
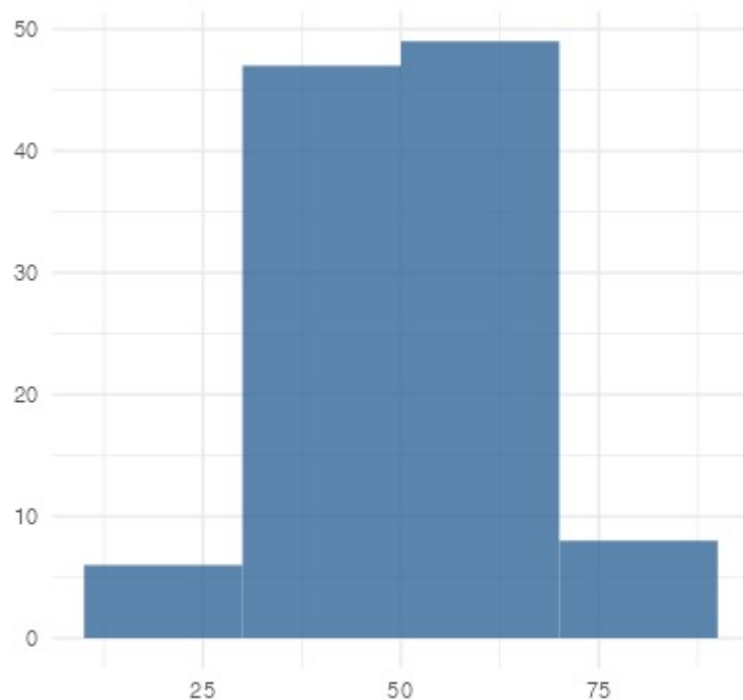
- ・ 微妙な大小関係は棒グラフの方が判断しやすい



復習: 量的変数の可視化

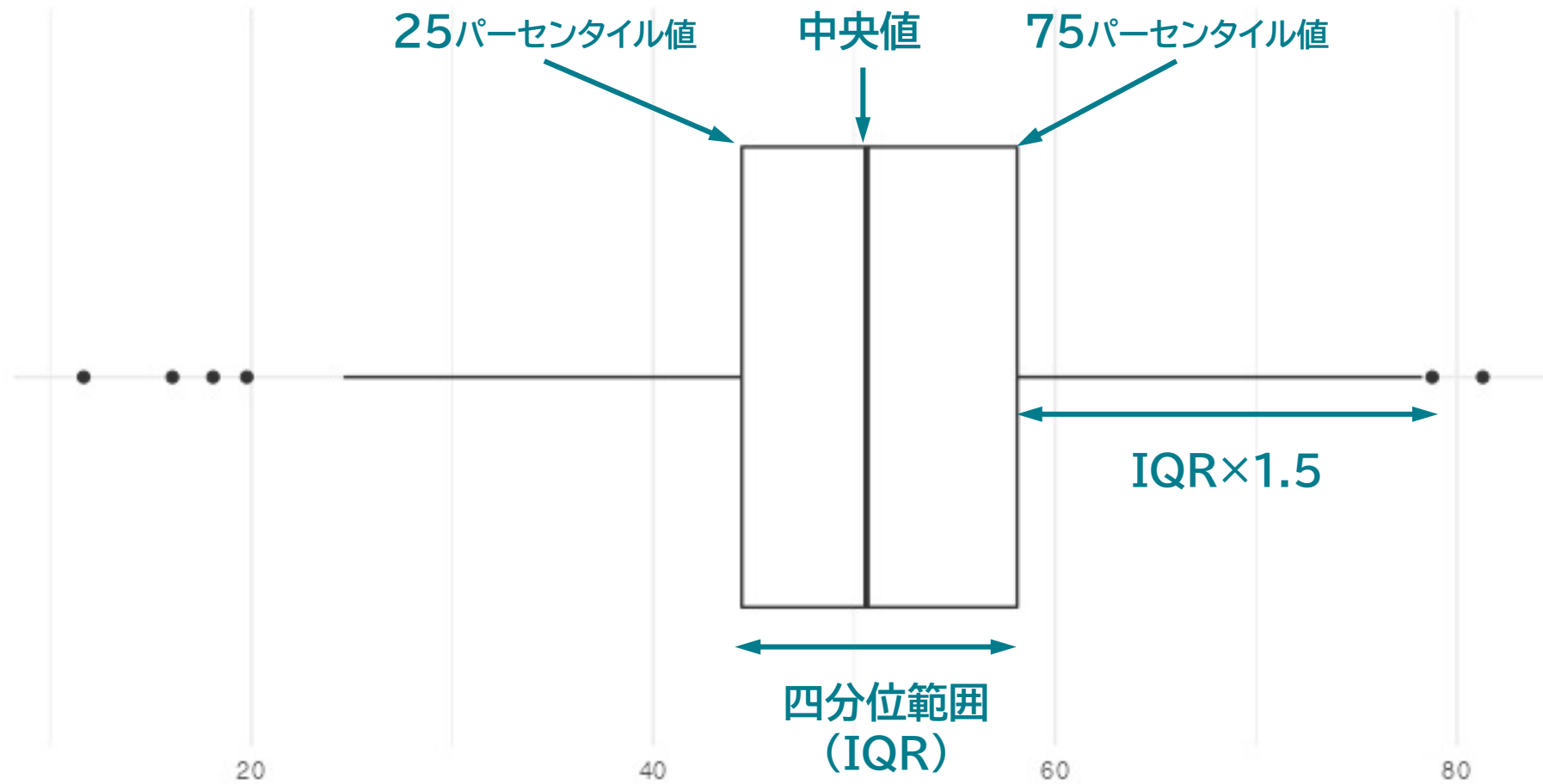
区間に分けてヒストグラムで示す

- 帯の面積が各区分の度数に比例する
- 同じデータでも区切り方で印象が変わる



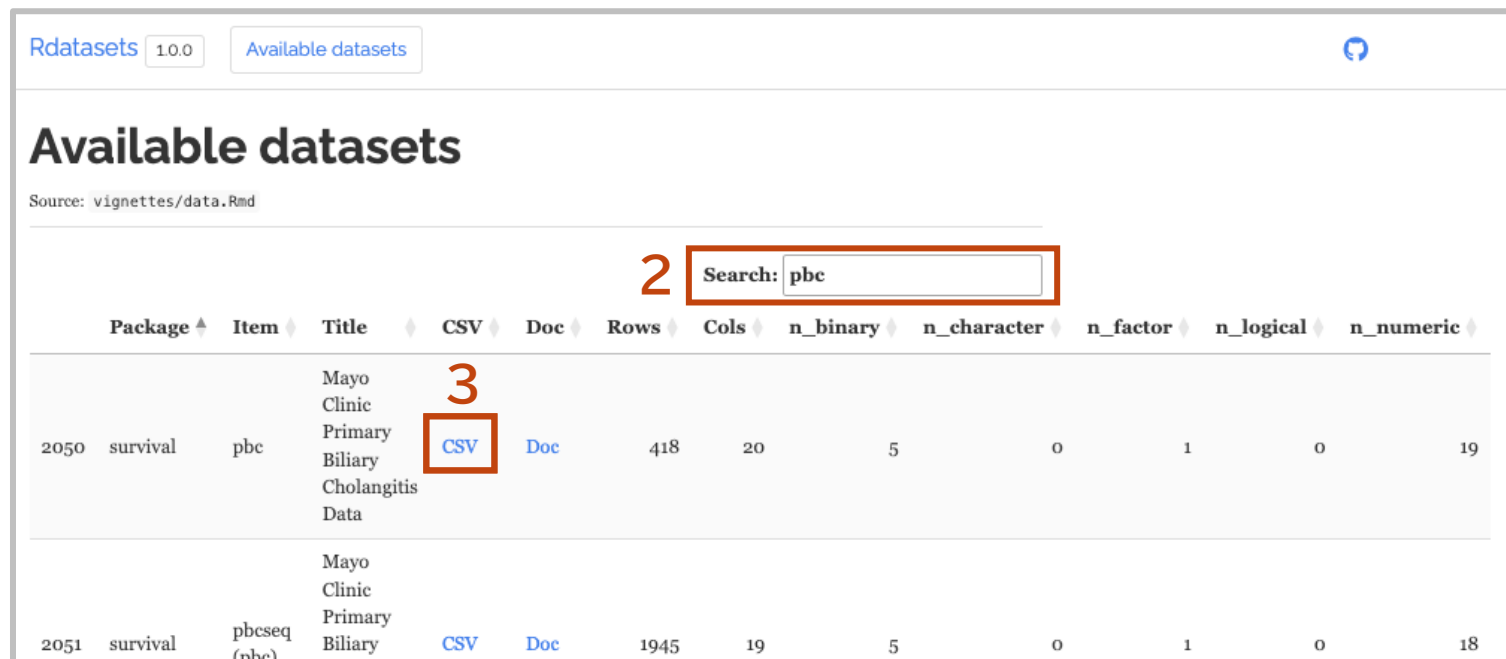
復習: 量的変数の可視化

箱ひげ図で要約値を図示する



用意されている練習用データセットを使う

- Rの豊富なサンプルデータ集(2000種類以上)
 - ✓ <https://vincentarelbundock.github.io/Rdatasets/articles/data.html>
 - ✓ Pythonにもサンプルデータはあるが、Rの方が豊富
- 今回もsurvivalパッケージのpbcデータを使う
 1. 上記のリンクにアクセス
 2. データセット名で検索
 3. csvファイルをダウンロード



The screenshot shows the Rdatasets website interface. At the top, there's a header with 'Rdatasets 1.0.0' and a link to 'Available datasets'. Below this is a search bar with the text 'Search: pbc' and a red box around it, labeled with a red '2'. The main content area is titled 'Available datasets' and lists search results. The first result is for the 'pbc' dataset, which is highlighted with a red box around the 'CSV' link, labeled with a red '3'. The table columns include Package, Item, Title, CSV, Doc, Rows, Cols, n_binary, n_character, n_factor, n_logical, and n_numeric.

Package	Item	Title	CSV	Doc	Rows	Cols	n_binary	n_character	n_factor	n_logical	n_numeric
2050	survival	pbc Mayo Clinic Primary Biliary Cholangitis Data	CSV	Doc	418	20	5	0	1	0	19
2051	survival	pbcseq (pbc) Mayo Clinic Primary Biliary	CSV	Doc	1945	19	5	0	1	0	18

Rを使いたい人 

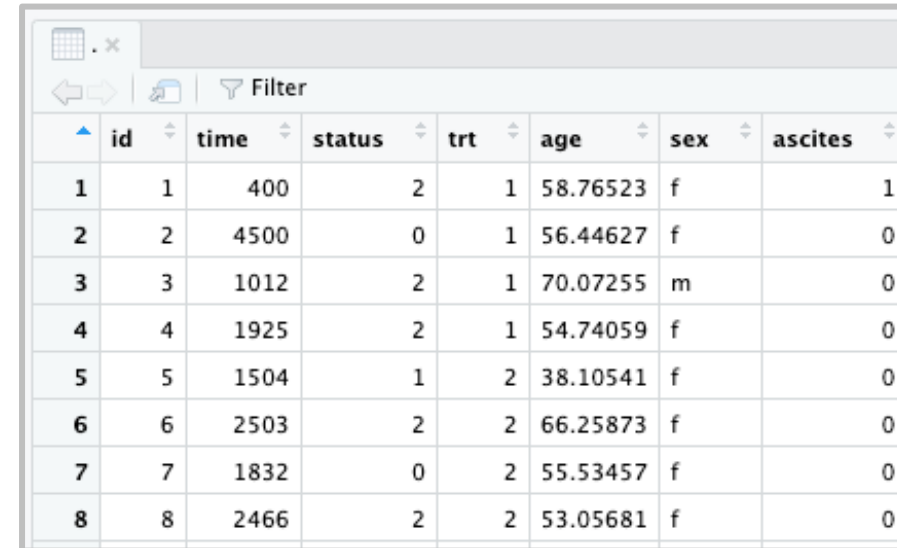
Rを使いたい人: サンプルデータを読み込んで確認する

1. survialパッケージを読み込む
(tidyverseパッケージも読み込んでおく)
2. data()関数でデータセットを読み込む

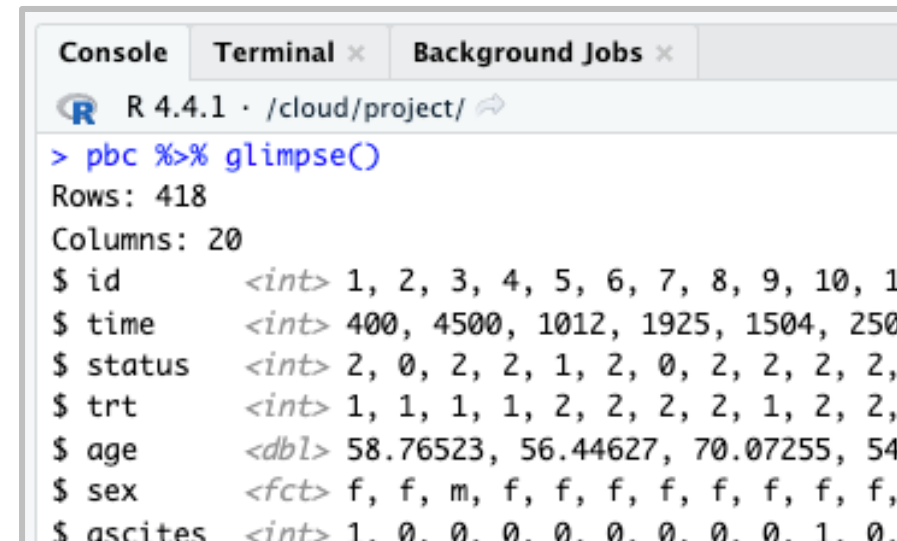
```
data(pbc)
```

3. 下のコードを実行して、データの全体像をつかんでおく

```
pbc %>% View()      #データセットを別タブで表示  
pbc %>% glimpse()   #変数一覧を表示
```



	id	time	status	trt	age	sex	ascites
1	1	400	2	1	58.76523	f	1
2	2	4500	0	1	56.44627	f	0
3	3	1012	2	1	70.07255	m	0
4	4	1925	2	1	54.74059	f	0
5	5	1504	1	2	38.10541	f	0
6	6	2503	2	2	66.25873	f	0
7	7	1832	0	2	55.53457	f	0
8	8	2466	2	2	53.05681	f	0



```
R 4.4.1 · /cloud/project/  
> pbc %>% glimpse()  
Rows: 418  
Columns: 20  
$ id      <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 1  
$ time    <int> 400, 4500, 1012, 1925, 1504, 250  
$ status  <int> 2, 0, 2, 2, 1, 2, 0, 2, 2, 2, 2  
$ trt     <int> 1, 1, 1, 1, 2, 2, 2, 2, 1, 2, 2  
$ age     <dbl> 58.76523, 56.44627, 70.07255, 54  
$ sex     <fct> f, f, m, f, f, f, f, f, f, f, f  
$ ascites <int> 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0
```

Rを使いたい人: 量的変数の分布をグラフで確認する

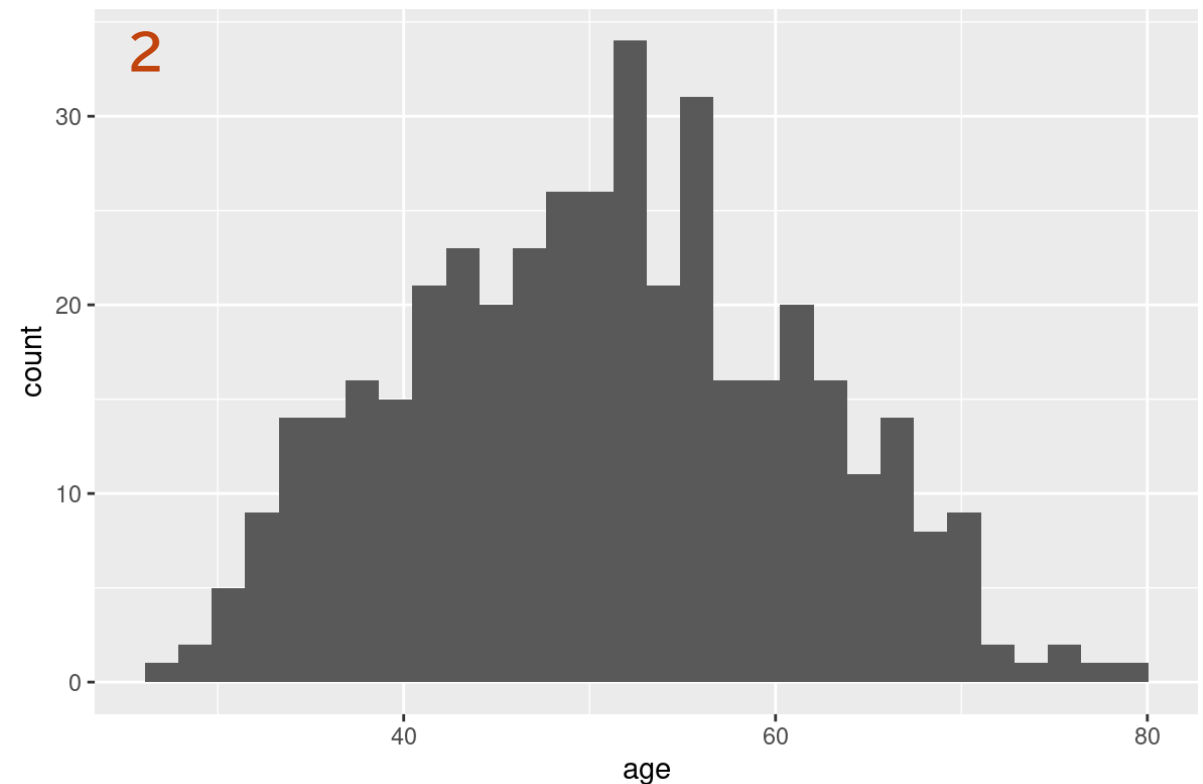
1. 下のコードを実行してみる

```
ggplot(data = pbc, aes(x = age)) +  
  geom_histogram()
```

2. 右下の [Plot] タブにグラフが表示される

3. 変えたい箇所があれば、
Chat GPTやGeminiに質問する！

- ✓ ヒストグラムの色を紺色(navy)にするには？
- ✓ 背景を白色にするには？
- ✓ X軸のラベルを“年齢(歳)”にするには？
- ✓ ヒストグラムの帯の幅(binwidth)を変えるには？
- ✓ 箱ヒゲ図を確認するには？



Rを使いたい人: ggplotの基本構文

- “gg”とは: Grammar of Graphics(作図の文法)
- グラフを描くために必要な要素をレイヤー(層)として重ねていく



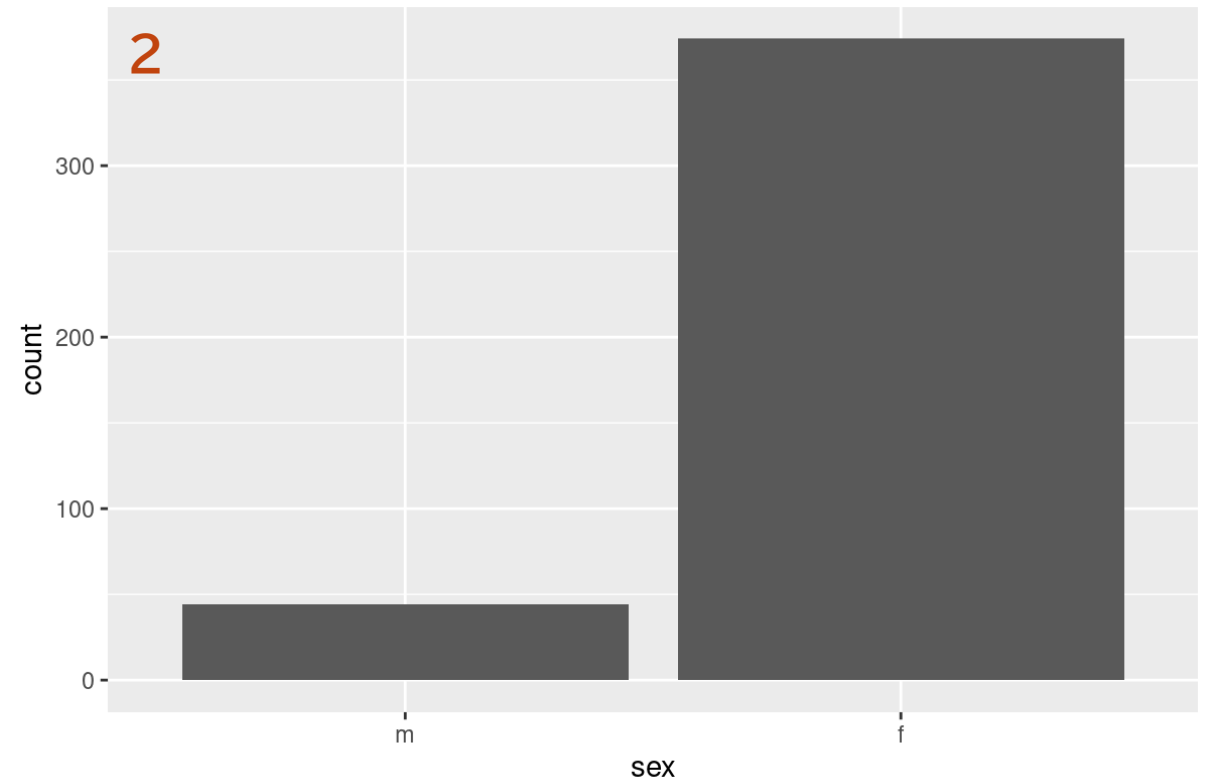
Theme	データに関与しない見た目
Coordinates	軸
Statistics	要約統計量を追加する
Facets	複数のグラフを並べる
Geometrics	どんなグラフのタイプを使うか
Aesthetics	どの変数をプロットするか どの変数で色分け・線種分けするか
Data	プロットしたいデータ

Rを使いたい人: 質的変数の分布をグラフで確認する

1. 下のコードを実行してみる

```
ggplot(data = pbc, aes(x = sex)) +  
  geom_bar()
```

2. 右下の [Plot] タブにグラフが表示される
3. 変えたい箇所があれば、
Chat GPTやGeminiに質問する！



Pythonを使いたい人 



Pythonを使いたい人: サンプルデータを読み込んで確認する

1. 使いたいサンプルデータのパッケージ名とデータセット名をメモしておく
2. 以下のように、statsmodelsパッケージのget_rdataset()を使う

```
import statsmodels.api as sm
dataset = sm.datasets.get_rdataset("データセット名", "パッケージ名")
df = dataset.data
```

✓ datasetには、データ本体(.data)のほか、データセットのタイトル(.title)やデータセットに関する説明(._doc_)も含まれているので、.dataという属性のみ取り出してdfと名前をつけた

3. 下のコードを実行して、データの全体像をつかんでおく

```
print(df)
```

	id	time	status	trt	age	sex	ascites	hepato	spiders	edema
0	1	400	2	1.0	58.765229	f	1.0	1.0	1.0	1.0
1	2	4500	0	1.0	56.446270	f	0.0	1.0	1.0	0.0
2	3	1012	2	1.0	70.072553	m	0.0	0.0	0.0	0.5
3	4	1925	2	1.0	54.740589	f	0.0	1.0	1.0	0.5
4	5	1504	1	2.0	38.105407	f	0.0	1.0	1.0	0.0
...
413	414	681	2	NaN	67.000684	f	NaN	NaN	NaN	0.0
414	415	1103	0	NaN	39.000684	f	NaN	NaN	NaN	0.0
415	416	1055	0	NaN	56.999316	f	NaN	NaN	NaN	0.0
416	417	601	0	NaN	58.001360	f	NaN	NaN	NaN	0.0



Pythonを使いたい人: 量的変数の分布をグラフで確認する

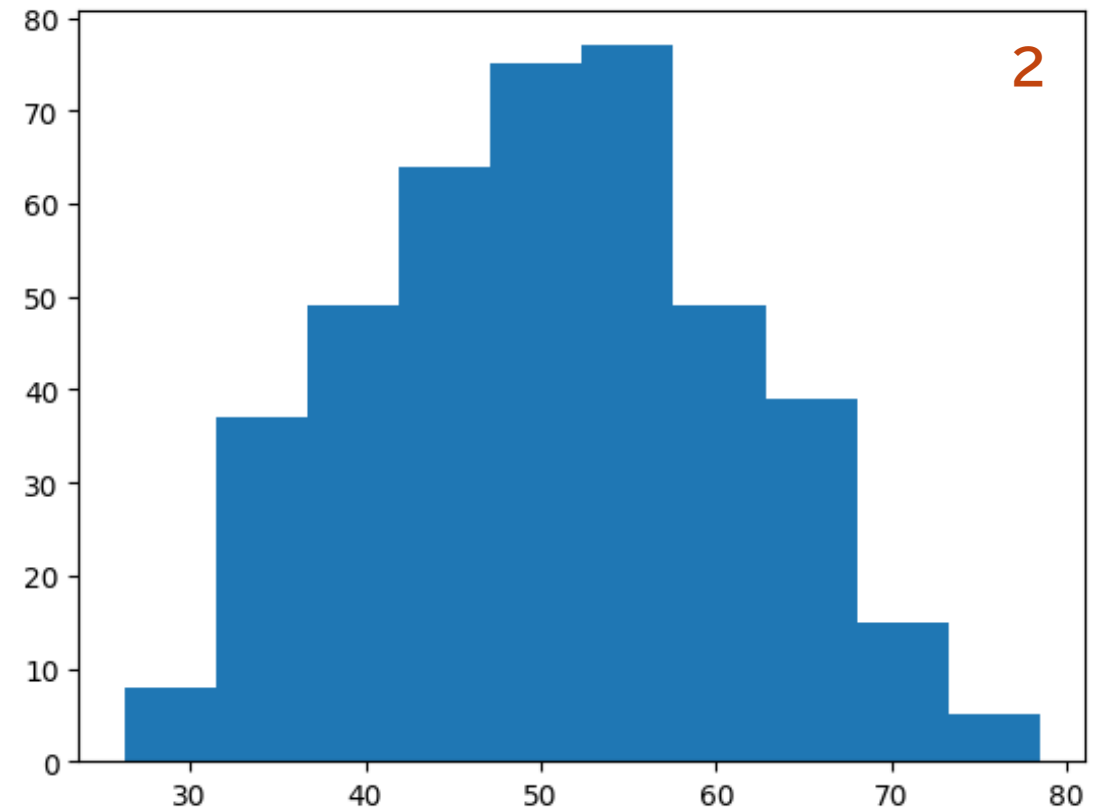
1. pandasとmatplotlibパッケージを読み込んでおく

```
import pandas as pd  
import matplotlib.pyplot as plt
```

2. 下のコードを実行してみる

```
plt.hist(df['age'])  
plt.show()
```

3. 変えたい箇所があれば、
Chat GPTやGeminiに質問する！
 - ✓ 軸にラベルをつけるには？
 - ✓ ヒストグラムの色を緑色(green)にするには？
 - ✓ ヒストグラムの帯の数(bins)を変えるには？
 - ✓ 箱ヒゲ図を確認するには？





Pythonを使いたい人: 質的変数の分布をグラフで確認する

1. pandasとmatplotlibパッケージを読み込んでおく

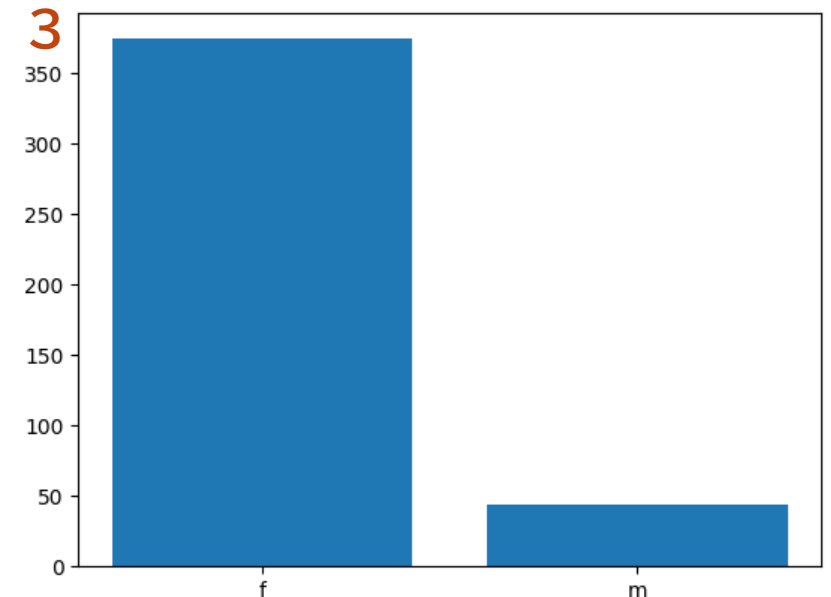
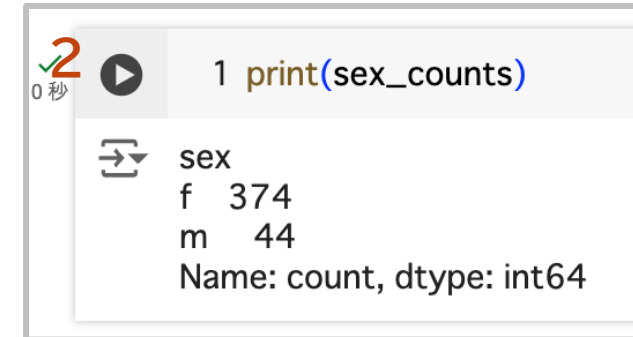
```
import pandas as pd  
import matplotlib.pyplot as plt
```

2. 下のコードで集計して、内容を確認する

```
sex_counts = df['sex'].value_counts()  
print(sex_counts)
```

3. 下のコードでグラフを描く

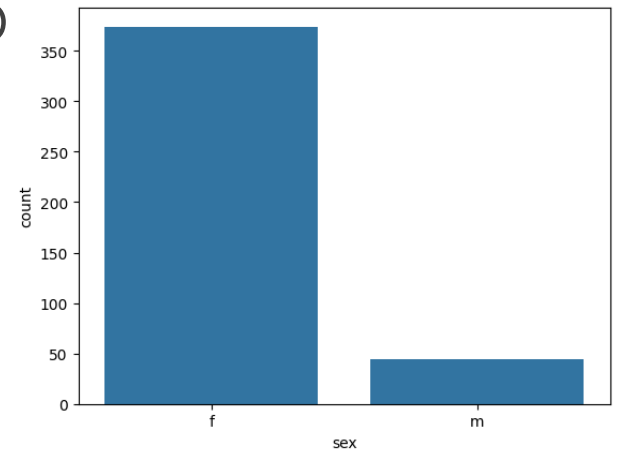
```
plt.bar(sex_counts.index, sex_counts)  
plt.show()
```



質的変数は集計が必要

- 質的変数をグラフにするときには、
カテゴリー数を集計した結果を作図関数に渡す必要がある
 - ✓ Pythonでは`value_counts()`を使って各カテゴリー数を集計した
 - ✓ Rの`ggplot`では質的変数であることを認識して自動で集計してくれる
(Rでも標準パッケージの`barplot()`関数を使うときは`table()`関数で集計必要)
- Pythonでも`seaborn`パッケージを使うとシンプルなコードになる
 - ✓ `seaborn`パッケージ: `matplotlib`を使いやすくしてくれたもの

```
import matplotlib.pyplot as plt
import seaborn as sns
sns.countplot(x='sex', data=df)
plt.show()
```



課題6：箱ヒゲ図

- Rのsurvivalパッケージにあるpbcデータについて、アルブミン値の分布を箱ヒゲ図で表してみましょう
 - ✓ 治療方法を表す変数: albumin
 - ✓ R(tidyverse): geom_boxplot() を使う
 - ✓ Python(matplotlib): .boxplot()を使う

今回のまとめ

- ✓ Rのggplotを使うときは、
　　プラス記号(+)が行末になるように改行します
- ✓ Chat GPTやGeminiを活用してください