“考古学のためのデータビジュアライゼーション”

“@ISHIIJUNPEI”

覚えるべき用語

連続量

- 数字で表される属性
- 土器の口径、器高、石器の刃部長や重量

離散量

- 何らかの分類
- 記号で表される属性
- 土器の分類、石器の器種

変数の性質と可視化手法

変数1	変数2	可視化手法
連続量		ヒストグラム
離散量		棒グラフ
離散量	連続量	ファセットヒストグラム、密度図、箱ひげ図
離散量	離散量	積上げ棒グラフ、ファセット棒グラフ
連続量	連続量	散布図

ヒストグラムで連続量を可視化する

連続量のデータ

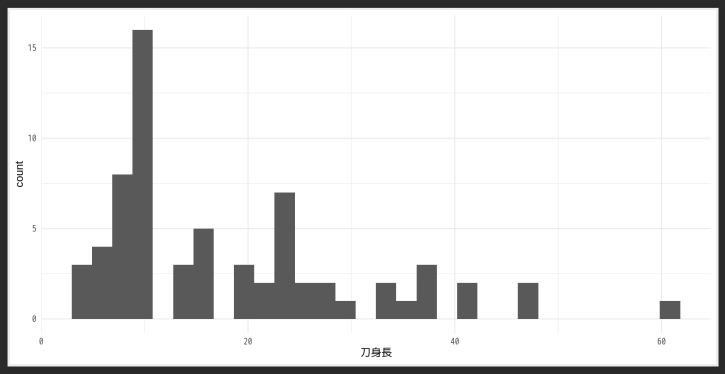
- 「分布の形」を確認する
- ヒストグラムを描く

刀身長の分布

北海道恵庭市西島松5遺跡出土の奈良時代の刀剣類のデータ

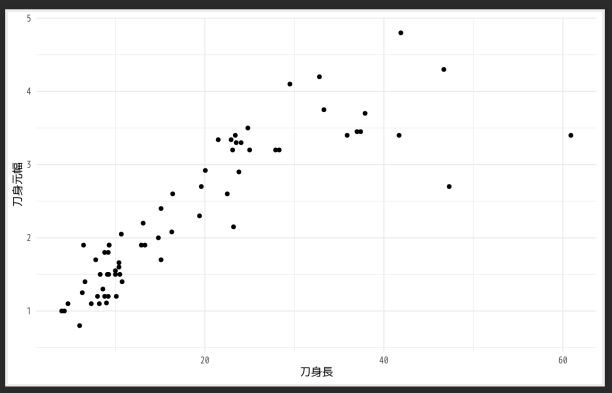
全長	刀身長	茎長	刀身先幅	刀身元幅	刀身元厚	茎先幅	茎元幅	茎先厚
6.2	4.00	2.20	0.80	1.00	0.40	0.60	0.80	0.30
9.2	4.30	4.90	0.90	1.00	0.30	0.40	1.05	0.30
6.9	4.70	2.20	1.00	1.10	0.25	0.65	0.80	0.20
8.2	6.00	2.20	0.65	0.80	0.30	0.80	1.05	0.30
11.8	6.30	5.50	0.60	1.25	0.30	0.60	1.05	0.30
12.0	6.44	5.56	1.40	1.90	0.40	0.65	1.25	0.34

刀身長は複数にサイズ分化がありそうだ



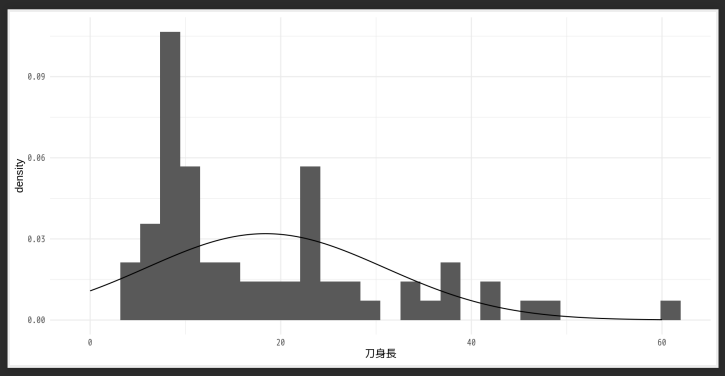
散布図ではだめなのか？

分布の形を観察するには適切でない



ヒストグラムを使うべき理由

分布の形状を数的モデルで近似



敬遠されがちヒストグラム

- 「数的モデルとの近似が容易である」という特性を活かせていない
- 正規分布で近似できることの意味がわかりにくい

エクセルでヒストグラム

- エクセルでヒストグラムを作りにくい
- 度数分布表から棒グラフを作成
- 度数分布表から作り直さなければならない

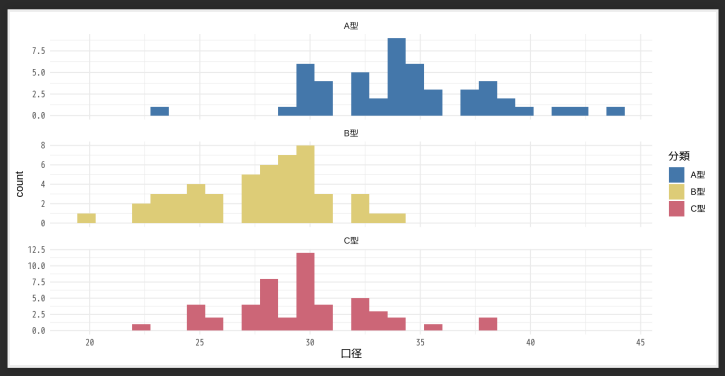
箱ひげ図を用いた比較

連続量と離散量の組み合わせ

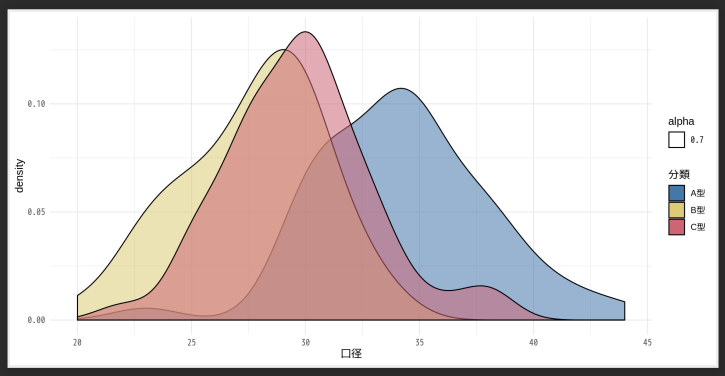
- 分類ごとにサイズの分布を確認する
- 有力な方法が複数存在
- 本命は**箱ひげ図**

ヒストグラム

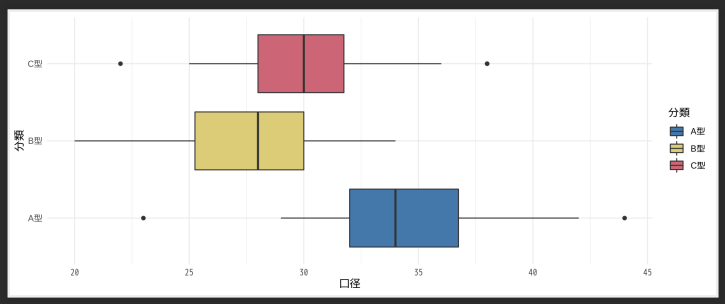
ヒストグラムを離散量でファセット



密度図



箱ひげ図



可視化手法とその性質

- 分類ごとの差を可視化する目的には**箱ひげ図**がもっとも有効
- 分布の形状に注目したい場合はヒストグラムや密度図

棒グラフをかしこく使う

近世陶磁器組成

北海道内近世後期の陶磁器組成データ

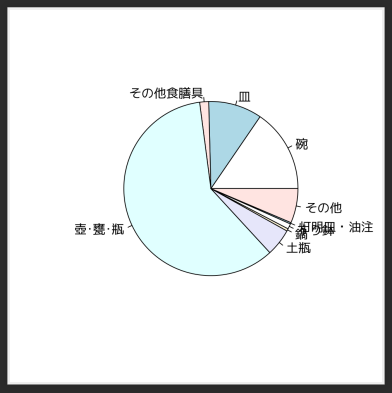
遺跡名	器種	点数
弁天貝塚	碗	134
弁天貝塚	皿	84
弁天貝塚	その他食膳具	34
弁天貝塚	土瓶	6
弁天貝塚	鍋	0
弁天貝塚	すり鉢	46

棒グラフの特性

- 離散量を表現する手法
- 遺跡ごとあるいは住居跡ごとに出土遺物の構成比を調べる
- 離散量（遺跡名）と離散量（器種）の組み合わせによるデータの可視化

円グラフは使わない

人間の目は面積の大小を認識するのは苦手



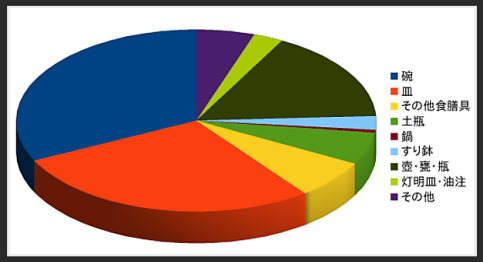
Rのヘルプでも・・・

円グラフは不適切な可視化手法です。人間の目は直線的な形状の判断には優れていますが、面の比較は苦手です。円グラフで表現できるデータは棒グラフやドットチャートで表現するべきです。

「円グラフで表示できるデータは全てドットチャートで表現できます。円の内角による不正確な判断ではなく、誰もが判断できるモノサシを用いるべきであることを意味しています」（Cleveland）

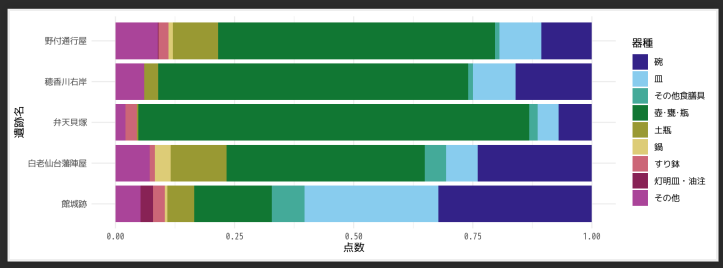
使ってはいけないグラフ

- 3D円グラフは目の錯覚を利用する手法
- 公文書や学術的な報告では**絶対**に使うべきではない



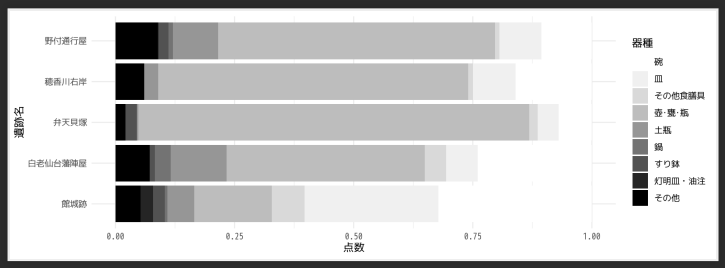
構成比棒グラフ

比率を可視化する優れたグラフ表現



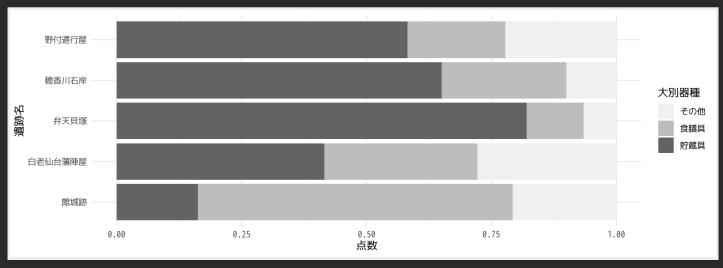
グレースケールの悲哀

- オフセット印刷の場合、グレスケール（網掛け）は20〜30%スパンが識別できる限界
- ぎりぎり4群〜5群



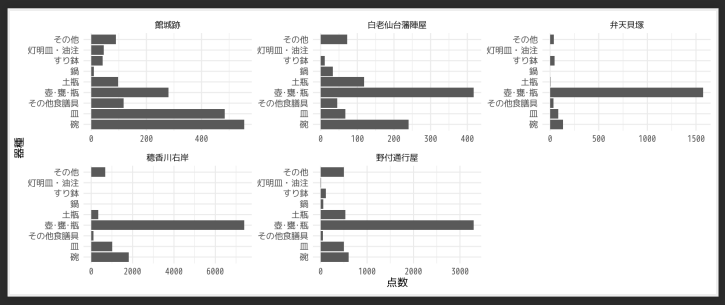
解決法A　カテゴリーを減らす

カテゴリーは3群をめざす



解決法B　ファセットする

花粉分析などでよく見る形のグラフ



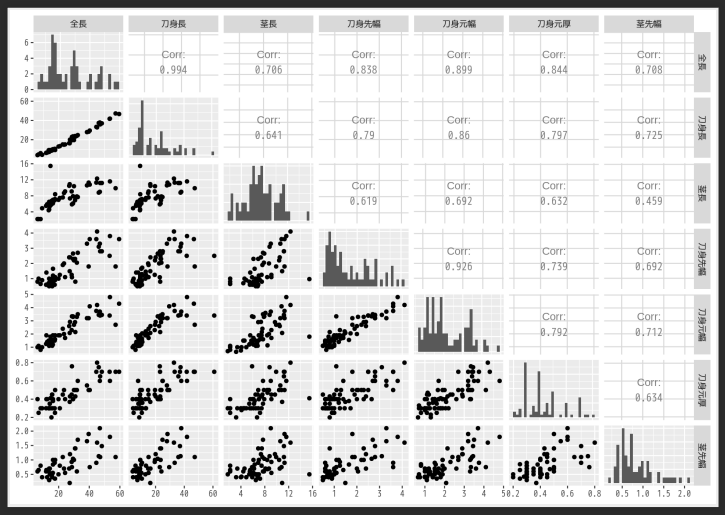
散布図で関係を可視化する

散布図の目的

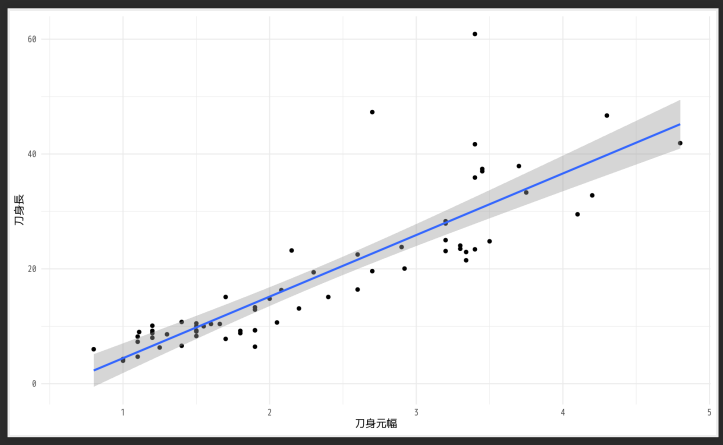
- 連続量×連続量の組み合わせのデータ
- 2変量の関係を可視化する
- つまり因果関係を可視化する

刀身長と他の属性の関係

- 恵庭西島松5遺跡出土の古代刀剣データ
- 追求すべきテーマは「刀身長と他の属性との因果関係」
- 刀身長を予測可能な変量はなにか？



刀身元幅から刀身長を予測する



刀身長の予測式

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-6.280888	1.9780720	-3.175257	0.0022892
刀身元幅	10.723991	0.7889999	13.591878	0.0000000

y=10.72x-6.28

まとめ

データの型をみきわめる

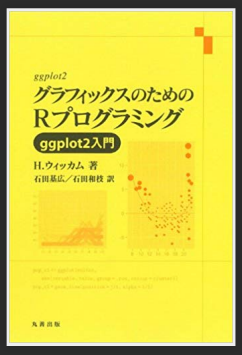
- データの型と組み合わせ
- 連続量と離散量

辞書代わり

Rグラフィックスクックブック ggplot2によるグラフ作成のレシピ集

（Winston Chang,オライリージャパン）

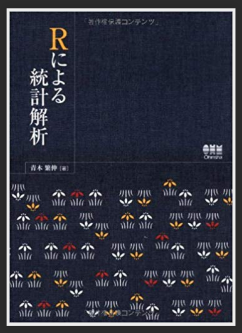
読み物として



グラフィックスのためのRプログラミング

（Hadley Wickham,丸善出版）

基本操作と統計手法



Rによる統計解析

（青木繁伸,オーム社）