考古学のためのデータビジュアライゼーション

石井淳平（厚沢部町）

Data Visualization in Archaeology Using R

Ishii Junpei（Assabu）

・データ解析／Data analysis・可視化／Visualization・ヒストグラム／Histogram
・多重比較／Multiple comparison

はじめに

統計解析環境Rとグラフィックパッケージであるggplot2を用いて、データの分布や構造を可視化する手法の実践例を紹介する。考古学で利用頻度が高い土器の法量や遺物の集計データを実践例として使用しており、日常の業務に本稿のコードがそのまま活用できる内容となっている。なお、本稿で使用したコードとサンプルデータはGitHub（https://github.com/IshiiJunpei/2019datasience）で公開している。

```
# パッケージ読み込み
library(tidyverse)
library(ggthemes)
library(knitr)
library(rmarkdown)
library(revealjs)
```

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などを表す。
・離散量
何らかの分類がなされ、記号で表される属性。土器の分類、石器の器種などを表す。

連続量と離散量の組み合わせによる可視化手法

可視化手法はデータの型とその組み合わせによって決まる。

```
read.csv("data/method.csv") %>% kable()
```

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

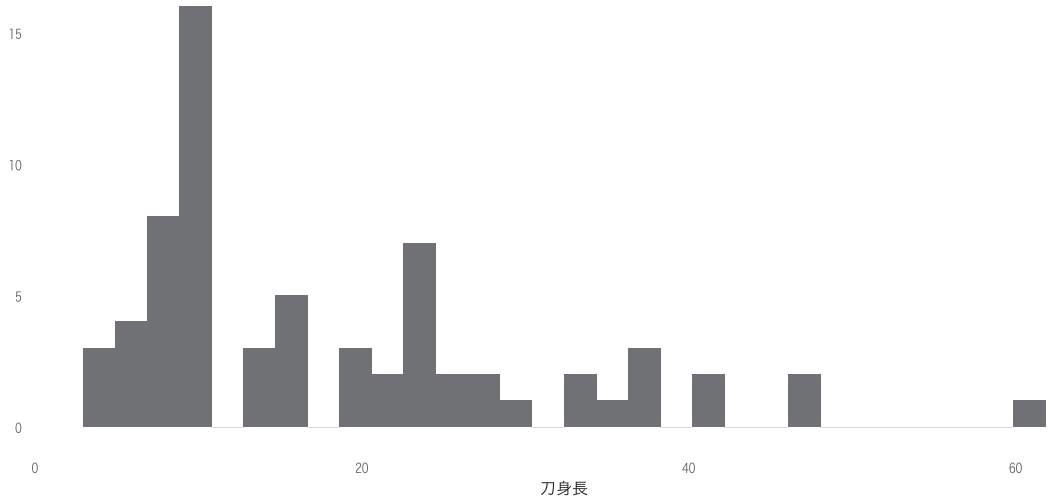
連続量のデータは「分布の形」を確認することが最初の作業となる。分布の形を可視化する最善の方法はヒストグラムを描くことである。

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を使用する。

```
iron <- read.csv("data/iron.csv")
iron[,c(4:12)] %>% head() %>% kable()
```

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



私たちには予備知識として、刀剣には刀子のようなマキリ状の小さなもの、刃渡り30cm前後の短刀、刃渡り60cmを超えるような太刀があることを知っているが、そうした予備知識をいったん忘れてデータを観察する。

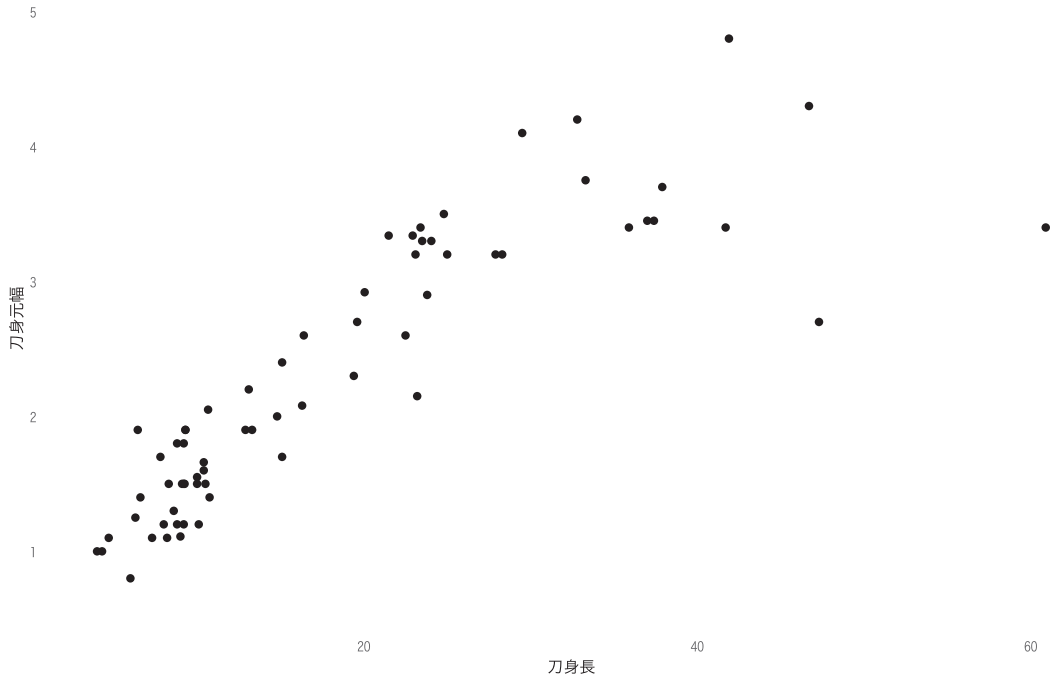
```
iron %>%
  ggplot(aes(x=刀身長)) +
    geom_histogram() +
    theme_minimal()
```

刀身長の分布は10cm、20cm超、40cm前後に峰をもつ3峰分布といえるだろうか？私たちの予備知識に照らし合わせると、刀子、短刀、脇差クラスに相当する刀身サイズの分化があると推測できる。ここではこれ以上踏み込まないが、「分布の形はヒストグラム」というのが鉄則である。

散布図ではだめなのか？

2変量が用意できる場合は散布図を用いることも可能と思われるかもしれない。考古学の論文や発掘調査報告書では、連続量の分布を示す際に散布図を用いているケースが非常に多いと感じる。

下の図は、刀身長と刀身元幅の散布図である。この図が間違いとは言わないが、ヒストグラムと比較して、分布の形がわかりやすいと言えるだろうか？



```
iron %>%
  ggplot(aes(x=刀身長,y=刀身元幅)) +
  geom_point()+
  theme_minimal()
```

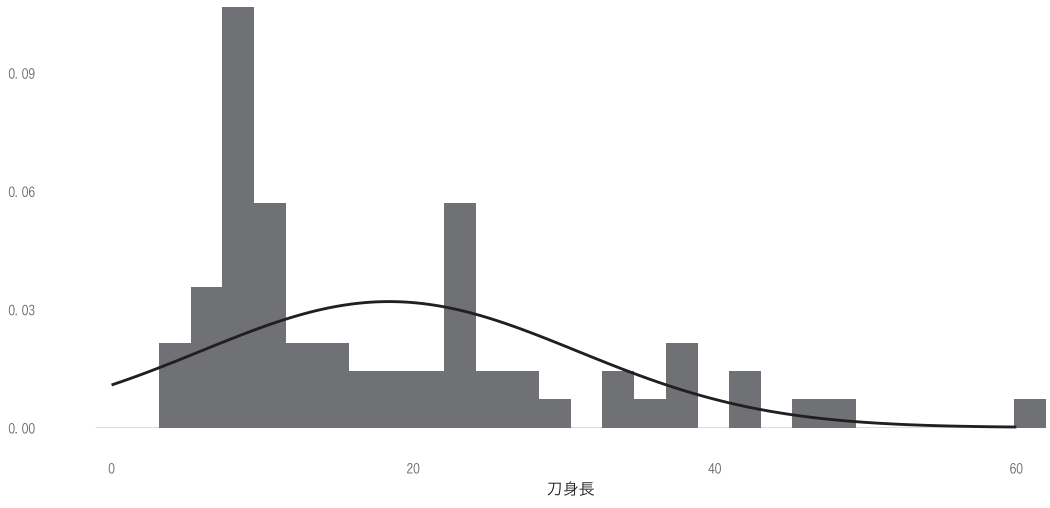
ヒストグラムを使うべき理由

ヒストグラムのもう一つの利点は、分布の形状を数的モデルに近似して比較できることである。次の図は正規曲線を重ねた刀身長のヒストグラムである。

```
# 正規曲線作成のための統計量算出
iron %>%
  summarise(mean = mean(刀身長,na.rm=T),
    sd=sd(刀身長,na.rm=T)) -> s_iron
kable(s_iron)
```

mean	sd
18.4003	12.49029

```
# 正規曲線作成
x<-seq(0, 60, 0.1)
nom <- x %>% dnorm(mean = s_iron$mean,
  sd = s_iron$sd)
nom2<-data.frame(X=x,Y=nom)
#正規曲線付きヒストグラム
iron%>%
  ggplot(aes(x = 刀身長,y = ..density..))+
    geom_histogram() +
    geom_line(data = nom2,aes(x=x,y=Y)) +
    scale_colour_ptol() +
    theme_minimal()
```



刀身長のヒストグラムと正規分布曲線を重ねることによって、刀身長の分布が正規分布から大きく外れていることがはっきりする。これは、散布図では絶対に表現できない。上記のヒストグラムから、古代の刀剣に複数のサイズ規範がある可能性を指摘することはできそうである。

なぜヒストグラムは使われないのだろうか？

理由の一つとして、ヒストグラムのもつ「数的モデルとの近似が容易である」という特性を考古学の研究者が活かせていない、ということが考えられる。例えば正規分布に対する理解や正規分布で近似できるということはどのような意味をもつのか、そのような判断が難しいのだろうと思う。

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」ということも理由の一つかもしれない。エクセルでヒストグラムを作れないわけではないが、度数分布表から棒グラフを作成することになるので、一手間かかる。

ビン幅の調整をするにも、いちいち度数分布表を作り直さないといけない、ということも面倒である。こうした理由でヒストグラムが敬遠されるではないかと感じている。

箱ひげ図を用いた連続量の比較

以下の手順でダミーデータを生成する。

```
# iris データ読み込み
data<-iris
#ダミーデータ生成
pot<-data[,c(1,2,5)]
colnames(pot)<-c( "器高","口径","分類" )
pot$分類<-factor(pot$分類,
  levels = c("setosa","versicolor",
    "virginica"),
  labels = c("A 型","B 型","C 型"))
pot$器高<-pot$器高*7
pot$口径<-pot$口径*10
pot %>% head() %>% kable()
```

器高	口径	分類
35.7	35	A 型
34.3	30	A 型
32.9	32	A 型
32.2	31	A 型
35.0	36	A 型
37.8	39	A 型

連続量の分布や差を、離散量ごとに比較する。連続量と離散量の組み合わせのデータとは、例えば土器分類ごとに口径の分布を確認するようなケースである。次に紹介するように有力な方法がいくつか存在する。本命は**箱ひげ図**であるが、バイオリンプロットも優れた可視化手法である。

ヒストグラム

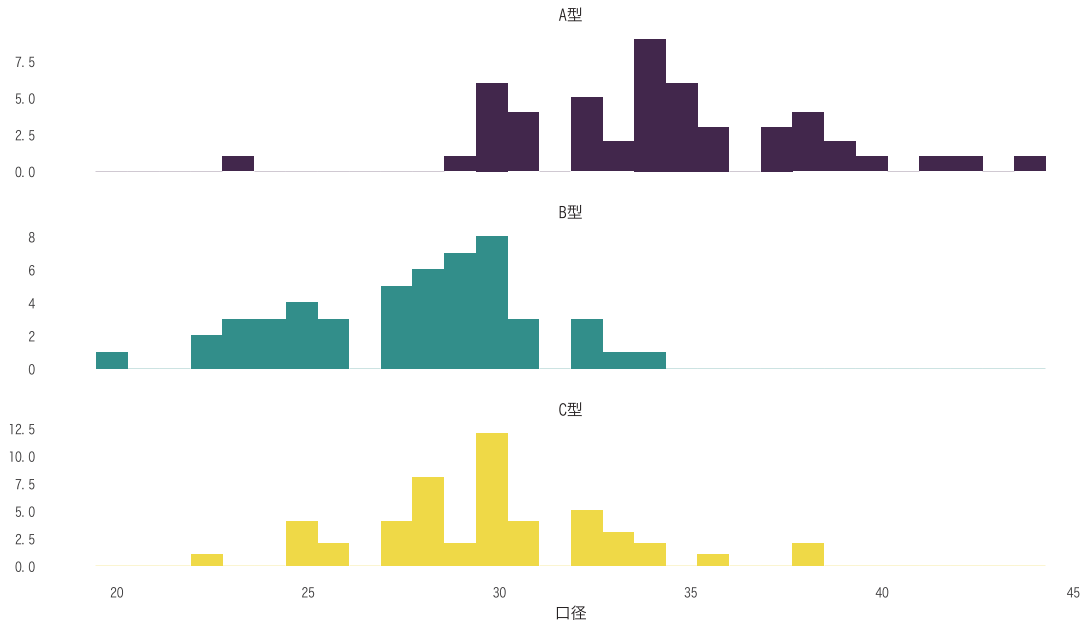
```
pot %>%
  ggplot(aes(x=口径,fill=分類))+
    geom_histogram() +
    scale_fill_viridis_d()+
    facet_wrap(~分類,ncol = 1,
      scales = "free_y" ) +
        theme_minimal()
```

密度図

```
pot %>%
  ggplot(aes(x=口径,fill=分類))+
    geom_density(alpha=0.7)+
    scale_fill_viridis_d()+
    theme_minimal()
```

箱ひげ図

```
library(ggforce)
pot %>%
  ggplot(aes(x=分類,y=口径,fill=分類))+
    #不透明度を 0.2
    geom_boxplot(alpha = 0.2)+
```

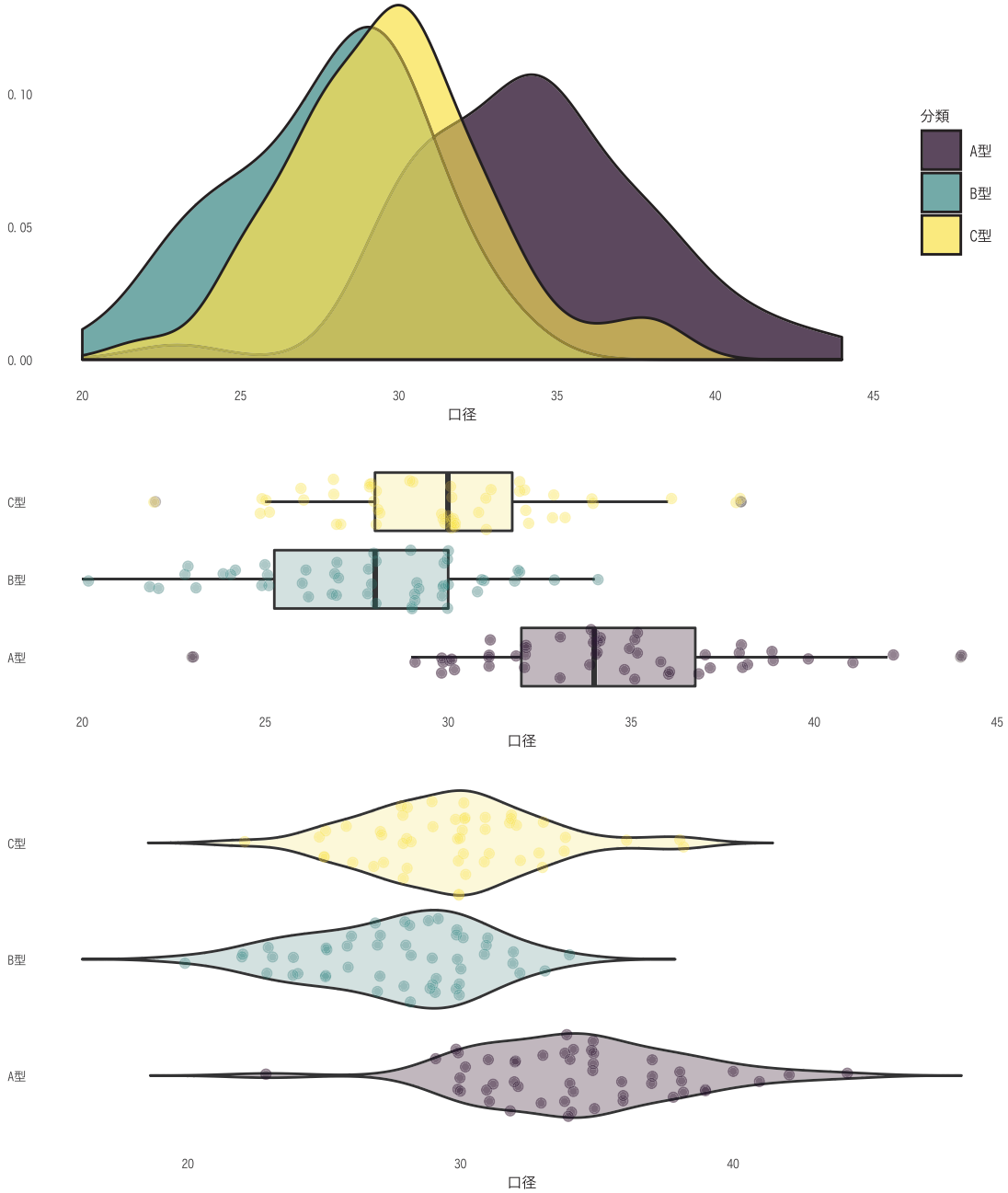


```
#geom_sina()関数で aes()の引数に colour=
  分類を指定
geom_sina(aes(colour = 分類),
  alpha = 0.4, size = 3) +
scale_fill_viridis_d() +
#viridis_d()は連続量、離散量なら
viridis_c()を指定する
scale_colour_viridis_d() +
coord_flip()+
theme_minimal()
```

分類ごとの口径の差を確認する目的では箱ひげ図がもっとも敏感に差を可視化してくれる。分布の形状に注目したい場合はヒストグラムや密度図も有力な手法となる。

多重比較による差の検定

箱ひげ図などの可視化手法によって、土器の口径は分類ごとに差がありそうだということがわかった。差があるかどうかを定量的に判断するために統計的な検定を行う。



この場合、3つの群に分類されているので、3つの群同士に差があるかどうかを統計的に確かめることになる。多群の差の検定手法の一つである「多重比較」を行う。

分散分析

最初に分散分析で品種によって差があるかどうかを確認する。p値が2.2e-16と極めて小さい値をとることから、**分類**によって差があることがわかる。

```
# aov関数の結果をanova関数に渡す。
# aov関数の第一引数は連続量~離散量
aov(口径 ~ 分類,data = pot) %>%
  anova() %>%
  kable(format = "markdown")
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
分類	2	1134.493	567.24667	49.16004	0
Residuals	147	1696.200	11.53878	NA	NA

TukeyHSD関数で多重比較

次にどの分類同士で差があるのかを調べるために「多重比較」という統計手法を用いる。いずれの分類でも有意な差を確認できる。

```
tkh <-
  aov(口径 ~ 分類, data = pot) %>%
  TukeyHSD() %>%
  .$分類 %>%  #TukeyHSD関数の結果から$分類を
              選択
  as_tibble() %>%  #tibble_df形式に変換
  mutate_if(is.numeric, round,3)
#mutate_if ()でnumericクラスのカラムに
round関数を適用する。
tkh %>% kable(format = "markdown")
```

diff	lwr	upr	p adj
-6.58	-8.189	-4.971	0.000
-4.54	-6.149	-2.931	0.000
2.04	0.431	3.649	0.009

棒グラフを賢く使う

器種や分類のような離散量を可視化する場合には棒グラフを用いる。北海道内近世後期の遺跡出土の陶磁器組成のデータを用いて構成比のグラフ表現について考える。

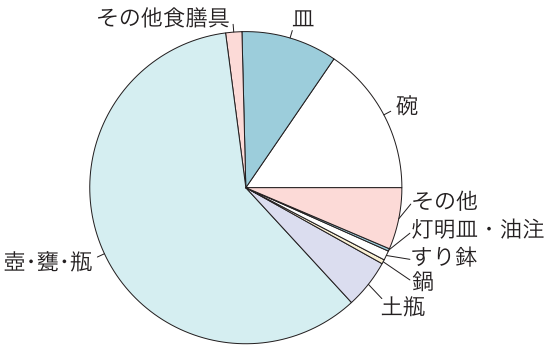
```
# データ読み込み
tojb <- read.csv("data/pot.csv")
# データの順序定義
toj$器種 <- toj$器種 %>%
  factor(levels = c("碗","皿",
    "その他食膳具","壺·甕·瓶", "土瓶","鍋",
    "すり鉢","灯明皿・油注","その他"))
toj %>% head() %>% kable()
```

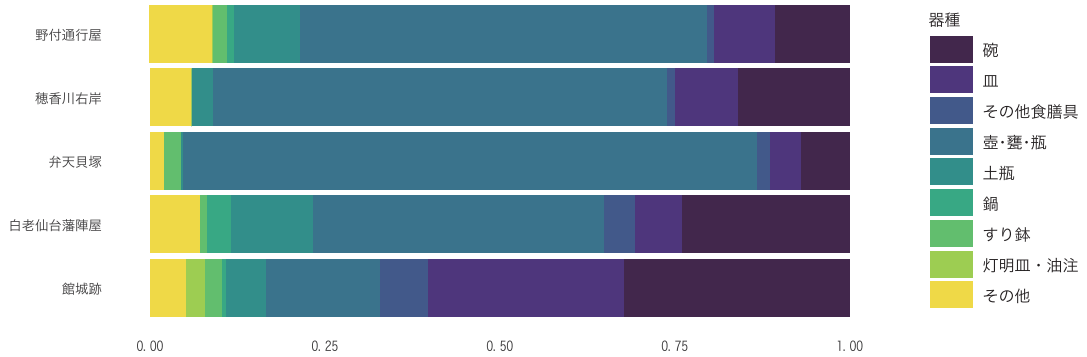
遺跡名	器種	点数
弁天貝塚	碗	134
弁天貝塚	皿	84
弁天貝塚	その他食膳具	34
弁天貝塚	土瓶	6
弁天貝塚	鍋	0
弁天貝塚	すり鉢	46

円グラフは使わない

もっとも大切なことは、円グラフを使わないということである。人間の目は線の長さや点の位置を把握することには長けているが、面積の大小や角度を認識するのは苦手である。円グラフは面積や円の内角で比率を表現することから、適切な可視化手法とはいえない。

```
toj_pie <- toj %>%
  group_by(器種) %>%
  summarise(点数=sum(点数))
pie(toj_pie$点数,labels=toj_pie$器種)
```





なお、Rで円グラフ（Pie charts）のヘルプを表示すると次のように記載されている。

Note:

```
Pie charts are a very bad way of displaying information.  The eye is good at judging linear measures and bad at judging relative areas.  A bar chart or dot chart is a preferable way of displaying this type of data.

Cleveland (1985), page 264: “Data that can be shown by pie charts always can be shown by a dot chart.  This means that judgements of position along a common scale can be made instead of the less accurate angle judgements.” This statement is based on the empirical investigations of Cleveland and McGill as well as investigations by perceptual psychologists.
```

意訳

円グラフは不適切な可視化手法である。人間の目は直線的な形状の判断には優れているが、面の比較は苦手である。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ある。

「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る。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る」（Cleveland 1985,p264）

ダメ！！絶対～3D円グラフ～

3D円グラフは目の錯覚を利用して、特定の値を大きく（小さく）見せる論外な手法である。公文書や学術的な報告では絶対に使うべきものではない。

構成比棒グラフ

構成比を比較するために使われるのが構成比棒グラフである。長さや位置によって視覚化されるため、正確な読み取りが可能である。構成比棒グラフは比率を比較するための優れたグラフ表現である。

```
toj%>%
  ggplot(aes(x=遺跡名,y=点数,fill=器種)) +
  geom_bar(stat="identity",
    position="fill") +
  coord_flip()+
  scale_fill_viridis_d()+
  theme_minimal()
```

モノクログラフの工夫

発掘調査報告書でカラーグラフが掲載できるケースは稀で、大半はグレースケールで表現されることになる。下のグラフはモニター上ではなんとか識別できるが、オフセット印刷の仕上がりでこれを識別することは不可能である。凡例との対比は絶望的である。

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である。したがって、構成比棒グラフでは4群～5群が表現の限界となる。

```
toj%>%
  ggplot(aes(x=遺跡名, y=点数, fill=器種)) +
  geom_bar(stat = "identity",
    position = "fill") +
      coord_flip()+
      scale_fill_brewer(palette="Greys") +
      theme_minimal()
```

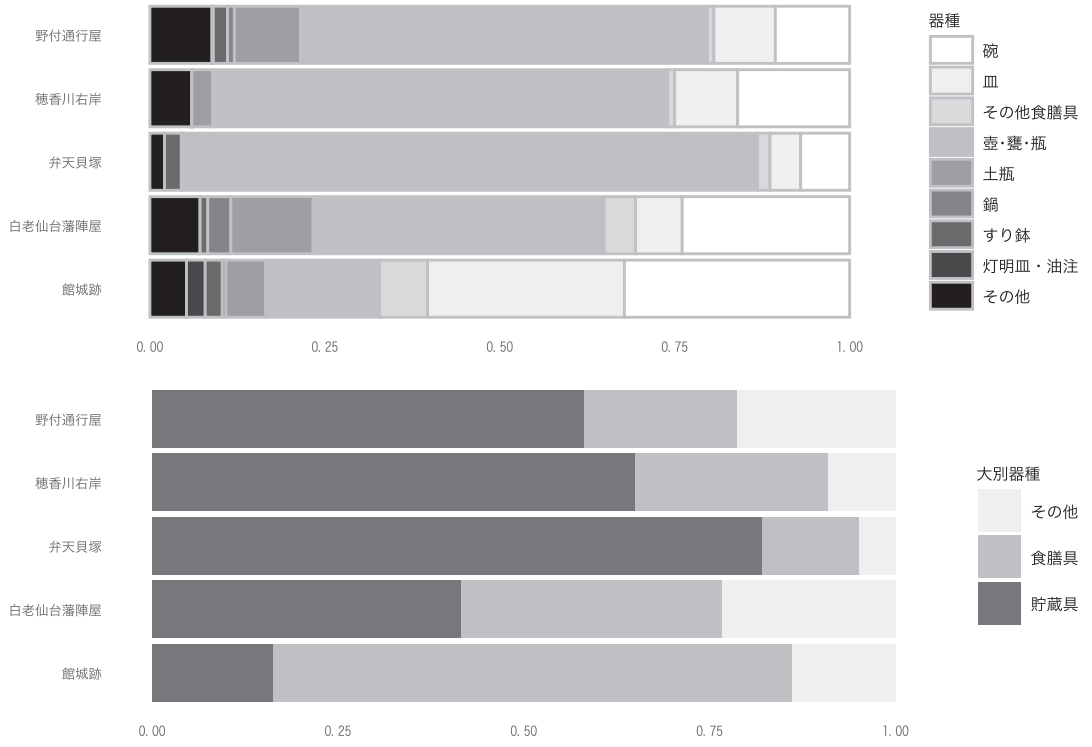
解決法１　カテゴリーを減らす

グラフ表現は複雑な現実をシンプルに割り切って視覚的に表現するためのものである。カテゴリー群が多すぎて識別が困難ならば、カテゴリーを減らすことをまずは考えるべきである。３群まで減らせばオフセット印刷でも識別可能なグレースケールのグラフになる。

```
# 食膳具、貯蔵具、その他に区分
toj2 <- toj %>%
  mutate(
    大別器種 = case_when(
    str_detect(器種,"碗|皿|その他食膳具") ~
      "食膳具",
    str_detect(器種,"壺･甕･瓶") ~
      "貯蔵具",
        str_detect(器種,
          "灯明皿・油注|その他|すり鉢|鍋|
          土瓶") ~ "その他",
    )
  )
# 3区分の構成比棒グラフ
toj2 %>%
  ggplot(aes(x=遺跡名,y=点数,
    fill=大別器種)) +
    geom_bar(stat=“identity”,
      position=“fill”) +
    coord_flip() +
    scale_fill_brewer(palette=“Greys”) +
    theme_minimal()
```

解決法２　ファセットされた棒グラフを使う

どうしてもカテゴリー数を減らしたくない場合は、群変数を器種にとって遺跡ごとにファセットする。花粉分析などの分析結果でよく見る形のグラフ



である。よほどカテゴリーが多くない限り、表現として成立しているし、オフセット印刷原稿としても対応可能である。

```
toj %>%
  ggplot(aes(x=器種,y=点数)) +
  geom_bar(stat="identity") +
  coord_flip() +
  facet_wrap(~遺跡名,scales="free") +
    theme_minimal()
```

散布図で2変量の関係を可視化する

散布図は連続量×連続量の組み合わせのデータで用いられる。考古学の論文・報文でもっとも多く使われるグラフ表現かもしれない。しかし、散布図が最も得意とする「2変量の関係を可視化する」という用途に使われることが意外に少ないように思う。

因果関係を可視化する

「2変量の関係を可視化する」ことの究極の目的は「因果関係の可視化」である。

たとえば、学力と子どもの生活環境の因果関係を統計的に示すなら、「学力テストの点数」という変量「果」（従属変数）に対して「因」となる変量（独立変数）は「親の収入」や「TV の視聴時間」、「睡眠時間」などが考えられる。

したがって、散布図を描く前に考えることは「因果」の「因」にあたる変量（独立変数）と「果」に当たる変量（従属変数）が何か、ということである。少なくとも「果」にあたる変量がはっきりしないデータは散布図を描く価値はない、と断言できる。

刀身長と他の属性の関係

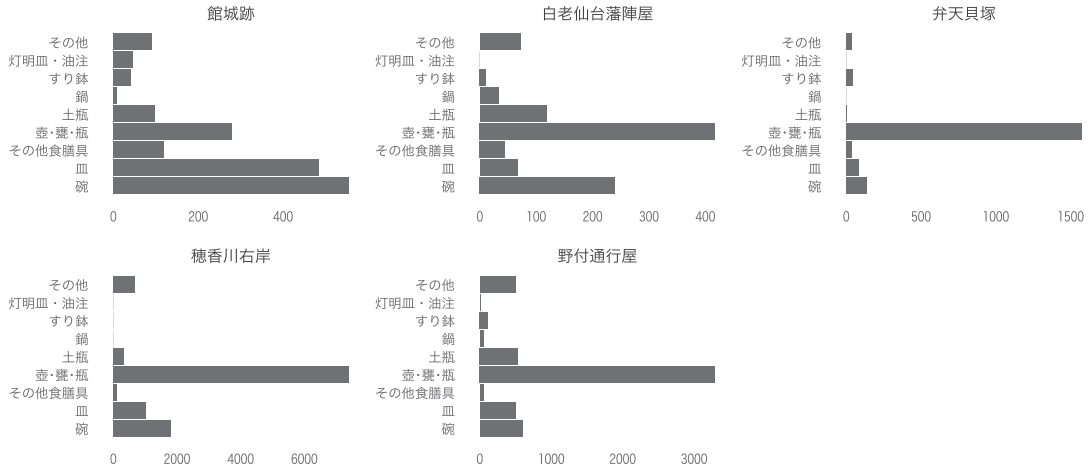
恵庭西島松5遺跡出土の古代刀剣を対象としたデータを再び使用する。追求すべきテーマは「刀身長と他の属性との因果関係」である。

刀身の長さは利用価値に即した刀剣サイズを示すものである。刀剣をつくるときには、刀身長が最初に決まり、刀身長に見合った各部のサイズが決められるものと予想される。この場合、因果関係の「果」にあたる変量が刀身長であり、「因」にあたる変量を探索することとなる。

なお、散布図を描く場合の約束として、因果関係の「果」にあたる変量をy軸に、「因」にあたる変量をx軸に割り当てる。y軸に割り当てられた「果」にあたる変量を従属変数、x軸に割り当てられた「因」にあたる変量を独立変数と呼ぶ。

研究集会ではGGallyパッケージを利用して散布図行列を描画したが、PerformanceAnalyticsパッケージを利用して有意性の評価を示している。

```
library(PerformanceAnalytics)
iron %>%
  select(全長，刀身長，茎長，刀身先幅，
    刀身元幅，刀身元厚，茎先幅）%>%
  chart.Correlation(histogram = TRUE,
    pch = 19)
```



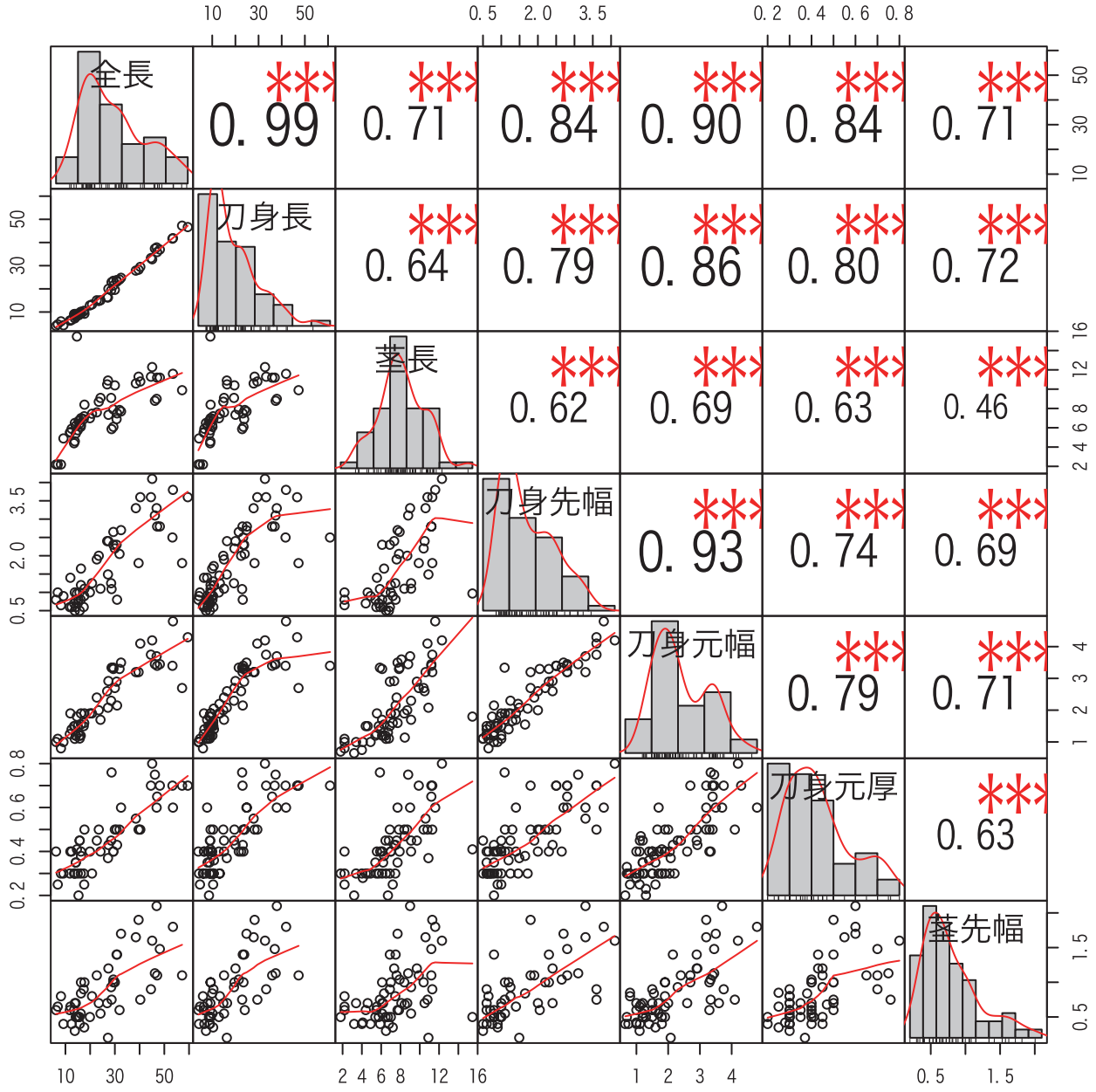
散布図が示すところからは、多くの属性が刀身長と相関関係にあることが読み取れる。一方、「刀身元幅」のように非常に強い相関を示す変量もあれば、「茎先幅」のように相関が弱い変量もある。刀身長との相関の強弱を判断することで、古代剣製作にかかる規範意識を読み取ることが可能かもしれない。

予測する

散布図を作成する目的は２変量の因果関係を考えることであった。因果関係がわかるということは予測ができるということである。次は古代刀剣の刀身元幅から刀身長を予測することを検討する。出土刀剣では刀身が破損せずに出土することはまれであるから、元幅から刀身長を予測できれば、出土刀剣の把握に大きな成果がありそうである。

```
p<-iron%>%
    ggplot(aes(x=刀身元幅,y=刀身長))+
        geom_point()+
        geom_smooth(method="lm")+
        theme_minimal()
```

なお、刀身元幅を独立変数とする刀身長の予測式は次のとおりである。



```
icoe<-lm(刀身長 ~ 刀身元
幅,data=iron)%>%summary()
icoe$coefficients%>%kable()
```

	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

$y=10.72x-6.28$

```
library(ggpmisc)
iron %>%
  ggplot(aes(x=刀身元幅,y=刀身長)) +
    geom_point() +
    geom_smooth(method="lm") +
    theme_minimal() +
    stat_poly_eq(formula = y ~ x,
      eq.with.lhs = "italic(hat(y))~`=`~",
       aes(label = paste(stat(eq.label),
         stat(rr.label), sep = "~~~")
    ), parse = TRUE
  ) +
  stat_fit_glance(label.y = 0.9,
    method = "lm",
    method.args = list(formula = y ~ x),
      aes(label = sprintf(
        '~~italic(P)~"="~~%.25f',
          stat(p.value)
        )
      ),parse = TRUE
    )
```