考古学のためのデータビジュアライゼーション

@ishiijunpei

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

覚えるべき用語

連続量

数字で表される属性です。土器の口径、器高、石器の刃部長や重量などです。

離散量

何らかの分類がなされ、記号で表される属性です。土器の分類、石器の器種などです。

連続量と離散量の組み合わせによる可視化手法

- 連続量　　ヒストグラム
- 連続量 × 連続量　散布図
- 離散量　　棒グラフ
- 離散量 × 離散量　積上げ棒グラフ、ファセット棒グラフ
- 離散量 × 連続量　ファセットヒストグラム、箱ひげ図

連続量と離散量〜法量と土器分類の関係〜

以下の手順でダミーデータを生成します。

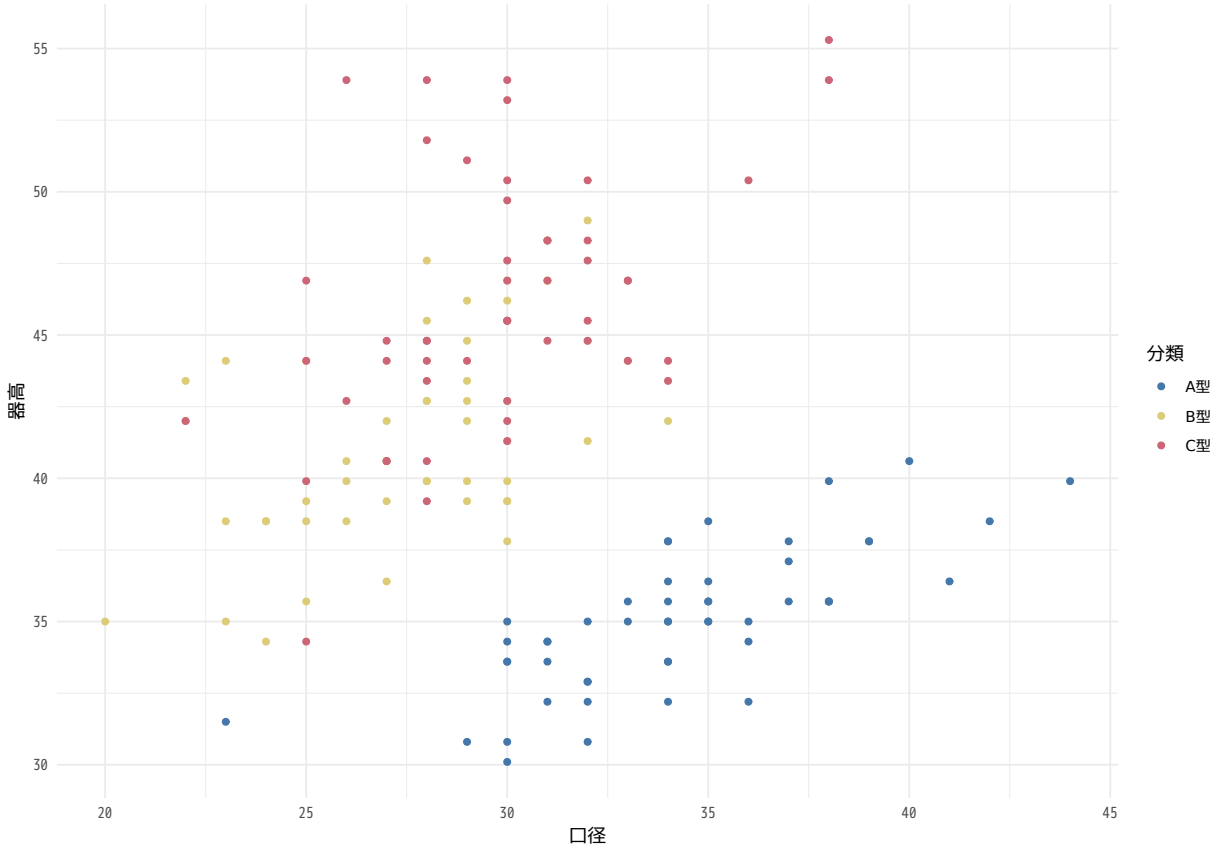
```
# iris データ読み込み
data<-iris
#ダミーデータ生成
pot<-data[,c(1,2,5)]
colnames(pot)<-c("器高","口径","分類")
pot$分類<-factor(pot$分類,levels=c("setosa","versicolor","virginica"),
    labels=c("A 型","B 型","C 型"))
pot$器高<-pot$器高*7
pot$口径<-pot$口径*10
pot%>%head()%>%kable()
```

器高	口径	分類
35.7	35	A 型

器高	口径	分類
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

土器の口径を x 軸、器高を y 軸にとって散布図を描き、法量と土器分類の関係を調べるケースを想定しています。下のような散布図により、「A 型」は「B 型」や「C」型と区別できそうだ、などと判断するわけです。

```
pot%>%
    ggplot(aes(x=口径,y=器高,colour=分類))+
        geom_point()+
        scale_colour_ptol()+
        theme_minimal()
```



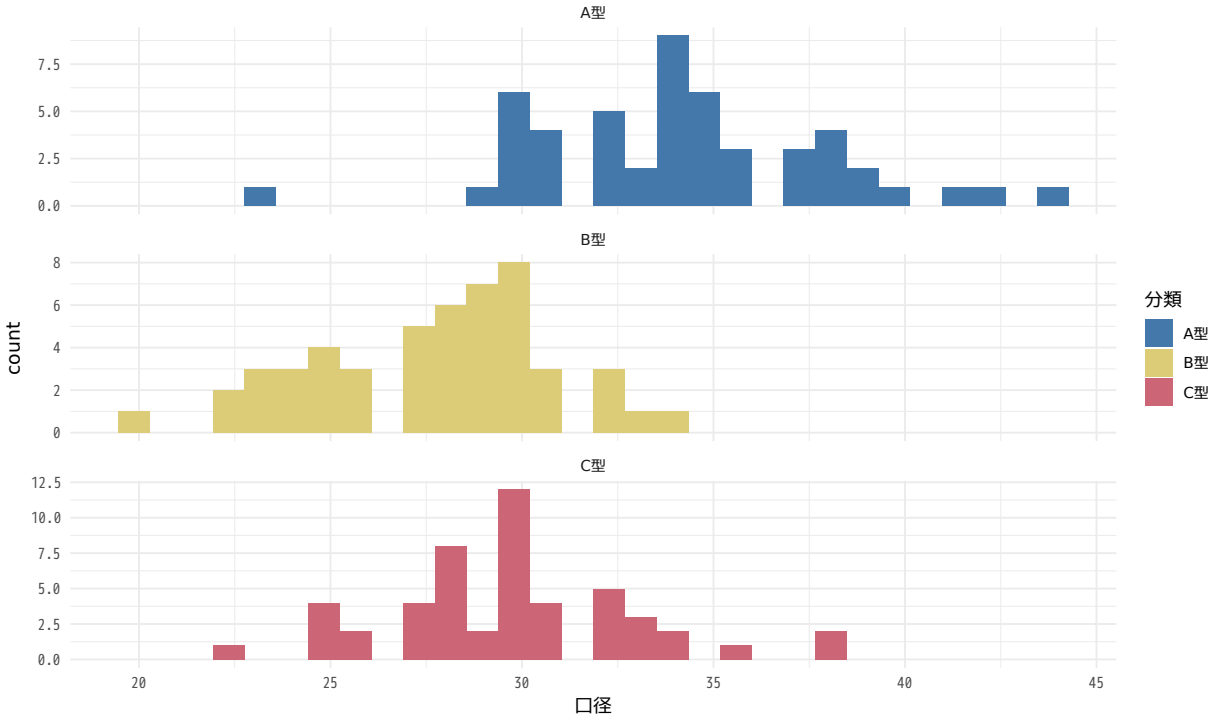
一変量ずつ分析する

複数の連続量の分布を散布図を用いて可視化する前に、一変量ずつ分布を確認することが必要です。連続量と離散量の組み合わせで可視化する場合には以下のような方法があります。

ヒストグラム

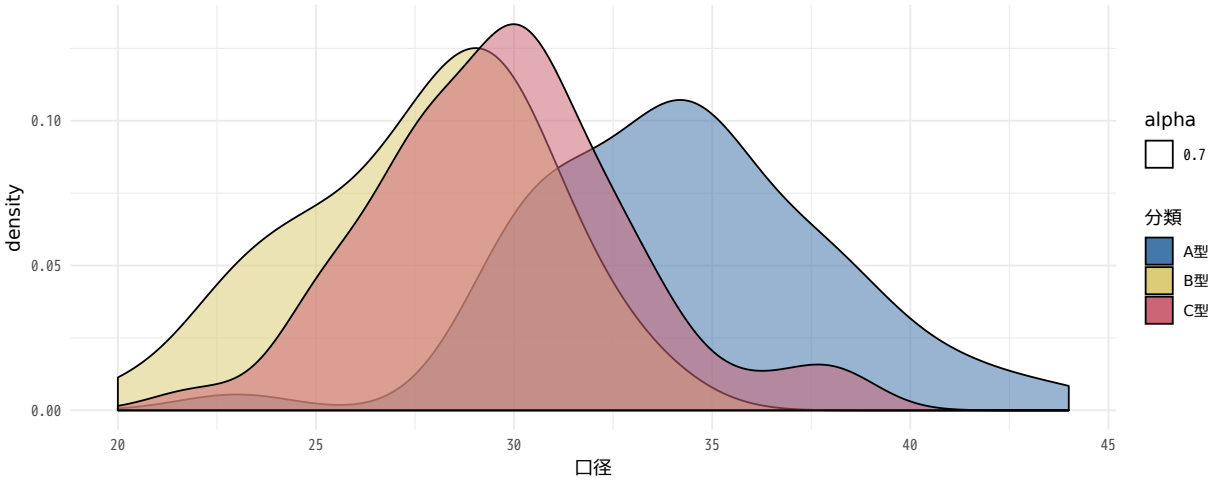
```
p<-pot%>%
    ggplot(aes(x=口径,fill=分類))+
        geom_histogram()+
        scale_fill_ptol()+
        facet_wrap(~分類,ncol=1,scales="free_y")+
        theme_minimal()
print(p)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



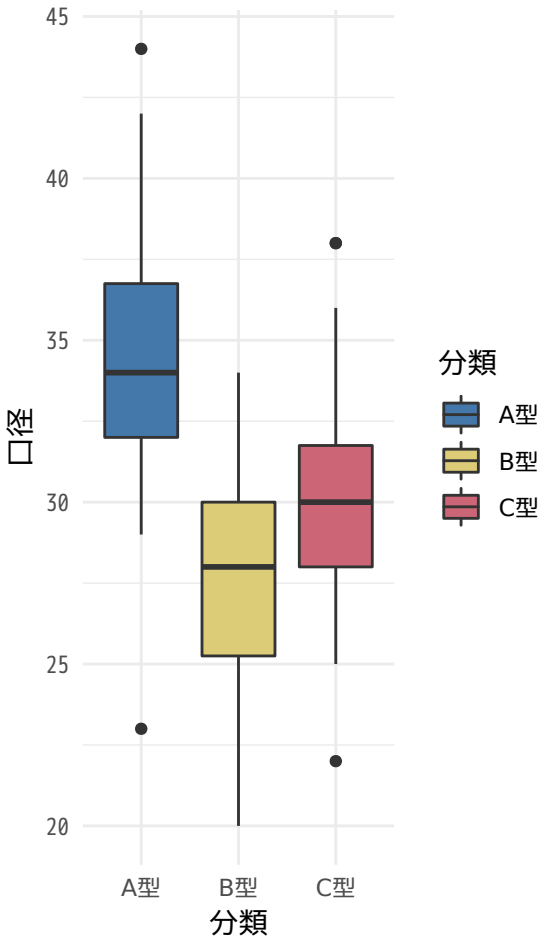
密度図

```
p<-pot%>%
    ggplot(aes(x=口径,fill=分類,alpha=0.7))+
        geom_density()+
        scale_fill_ptol()+
        theme_minimal()
print(p)
```



箱ひげ図

```
p<-pot%>%
    ggplot(aes(x=分類,y=口径,fill=分類))+
        geom_boxplot()+
        scale_fill_ptol()+
        theme_minimal()
print(p)
```



密度図や箱ひげ図では分類ごとの差がよくわかります。散布図では「B 型」と「C 型」の違いは明確ではありませんでしたが、1 変量ずつ比較することで分類ごとの差が明確になりました。。

まとめ

- 安易な散布図で納得しない。
- 連続量の分布を知りたければヒストグラム library(tidyverse) library(ggthemes) library(knitr) library(rmarkdown)
- 離散量ごとの分布の違いを知りたければ箱ひげ図

多重比較

箱ひげ図によって、土器の口径は分類によって差がありそうだということがわかりました。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います。

この場合、3 つの群に分類されていますので、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ります。多群の差の検定手法の一つである「多重比較」を行います。

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します。 p 値が 2.2e-16 と極めて小さい値をとることから、品種によって差があることがわかります。

```
# aov 関数の結果を anova 関数に渡します。
# 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を行います。いずれ分類でも有意な差が確認できます。

```
# aov 関数の結果を TukeyHSD 関数に渡す
tkh<-aov(口径~分類,data=pot)%>%TukeyHSD()
tkh$分類%>%kable(format="markdown")
```

	diff	lwr	upr	p adj
B 型-A 型	-6.58	-8.1885528	-4.971447	0.0000000
C 型-A 型	-4.54	-6.1485528	-2.931447	0.0000000
C 型-B 型	2.04	0.4314472	3.648553	0.0087802

まとめ

多重比較は強力な手法ですが、統計的に有意であることが考古学的に有意であることを保証するわけではありません。検定で有意差を証明することは大切ではありますが、必須の作業ではありません。適切な手法でデータの分布を可視化するだけで、多くの場合は十分です。

ヒストグラムを活用する

連続量のデータがあった場合、まず何をするのが適切か？と問われれば、「分布の形を確認する」と答えることになります。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す。

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します。

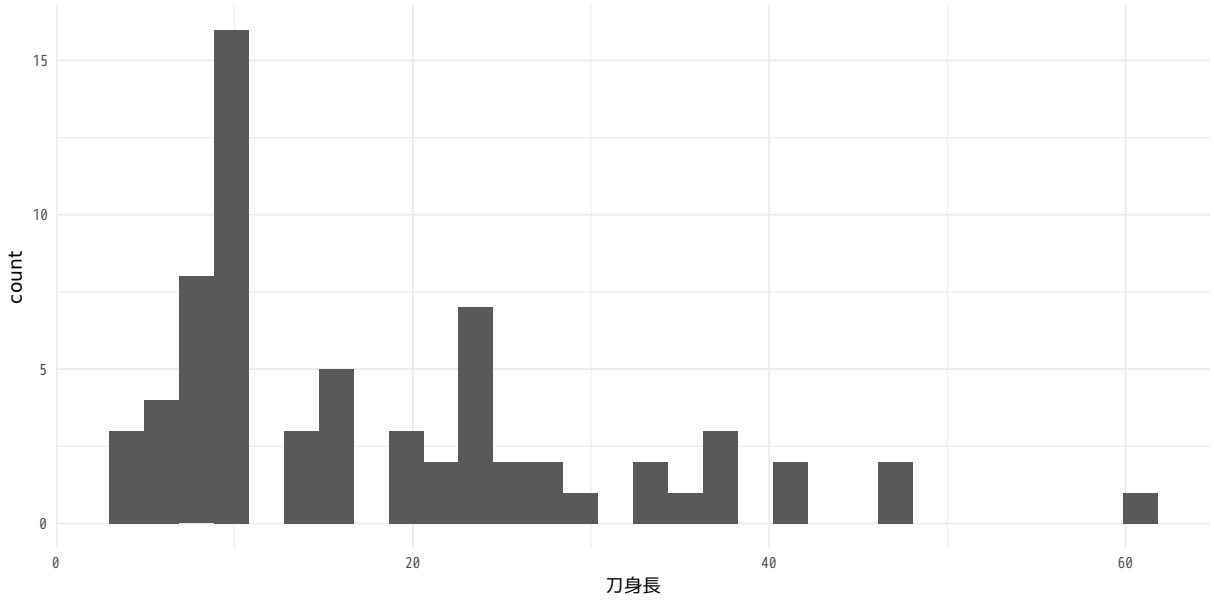
```
iron<-read.csv("data/iron.csv")
iron%>%head()%>%kable()
```

遺構番号	名称	分類	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
P125	小刀・刀子	刀子小	6.2	4.00	2.20	0.80	1.00	0.40	0.60	0.80	0.30
P103	小刀・刀子	刀子小	9.2	4.30	4.90	0.90	1.00	0.30	0.40	1.05	0.30
P121	小刀・刀子	刀子小	6.9	4.70	2.20	1.00	1.10	0.25	0.65	0.80	0.20
P97	小刀・刀子	刀子小	8.2	6.00	2.20	0.65	0.80	0.30	0.80	1.05	0.30
P125	小刀・刀子	刀子小	11.8	6.30	5.50	0.60	1.25	0.30	0.60	1.05	0.30
P207	小刀・刀子	刀子小	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ますが、そうした予備知識をいったん忘れてデータを観察します。

刀身長の分布は 10cm、20cm 超、40cm 前後に峰をもつ 3 峰分布といえるでしょうか？私たちの予備知識に照らし合わせると、刀子、短刀、短めの太刀に相当する刀身サイズの分化があると推測できます。ここではこれ以上踏み込みませんが、「分布の形はヒストグラム」というのが鉄則です。

```
p<-iron%>%
    ggplot(aes(x=刀身長))+
    geom_histogram()+
    theme_minimal()
print(p)
```

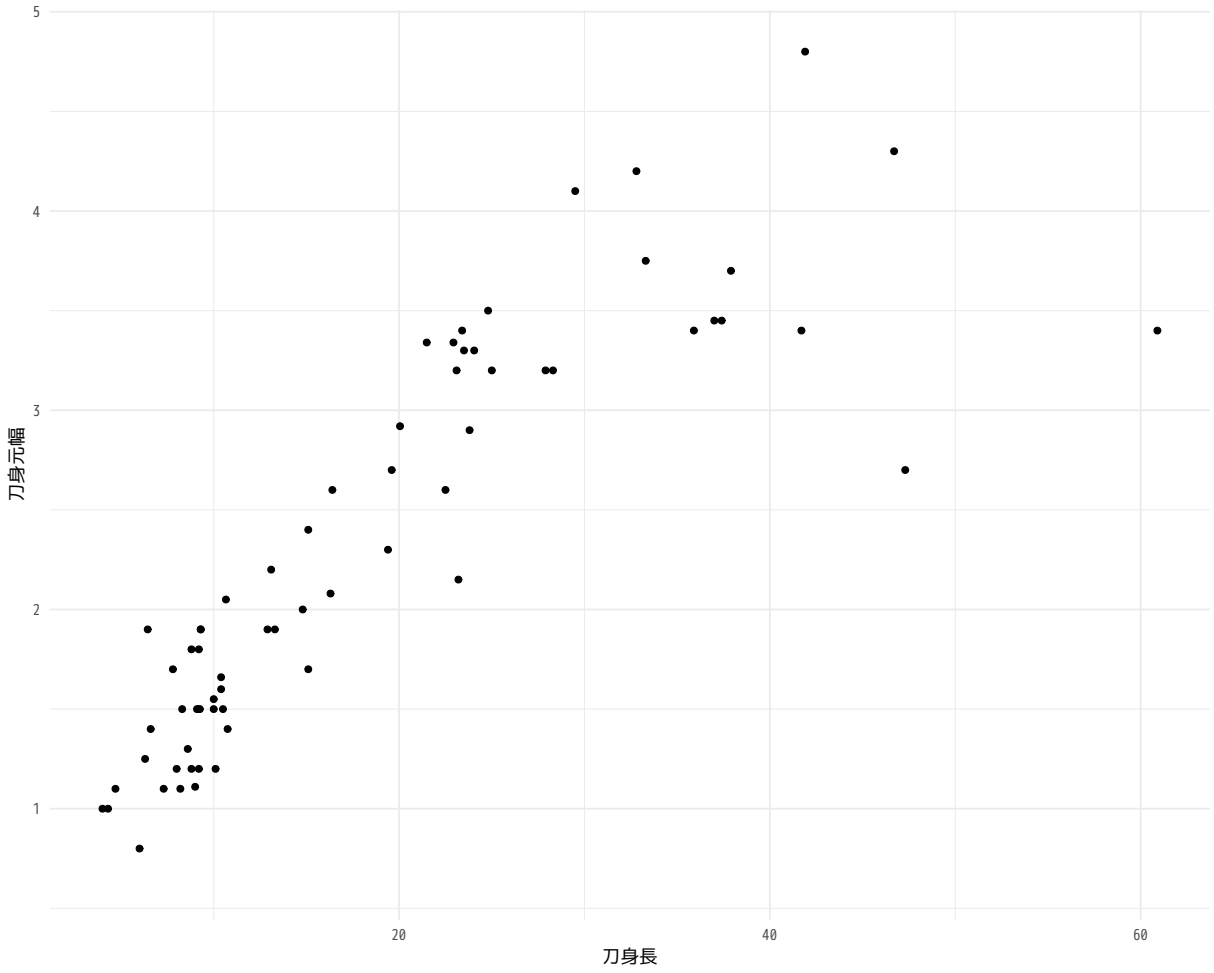


散布図ではだめなのか？

2 変量が用意できる場合は散布図で比較することも可能と思われるかもしれません。実際に考古学の論文や発掘調査報告書では、分布の可視化に散布図を用いているケースが非常に多いと感じます。

下の図は、刀身長と刀身元幅の散布図です。この図が間違いとは言いませんが、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でしょうか？下の散布図から分布に関して何らかの結論を出すのは難しいのではないでしょうか。

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



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す。下の図は正規曲線を重ねた刀身長のヒストグラムです。正規化しているので、Y 軸は密度になっていますが、分布の形は変化しません。

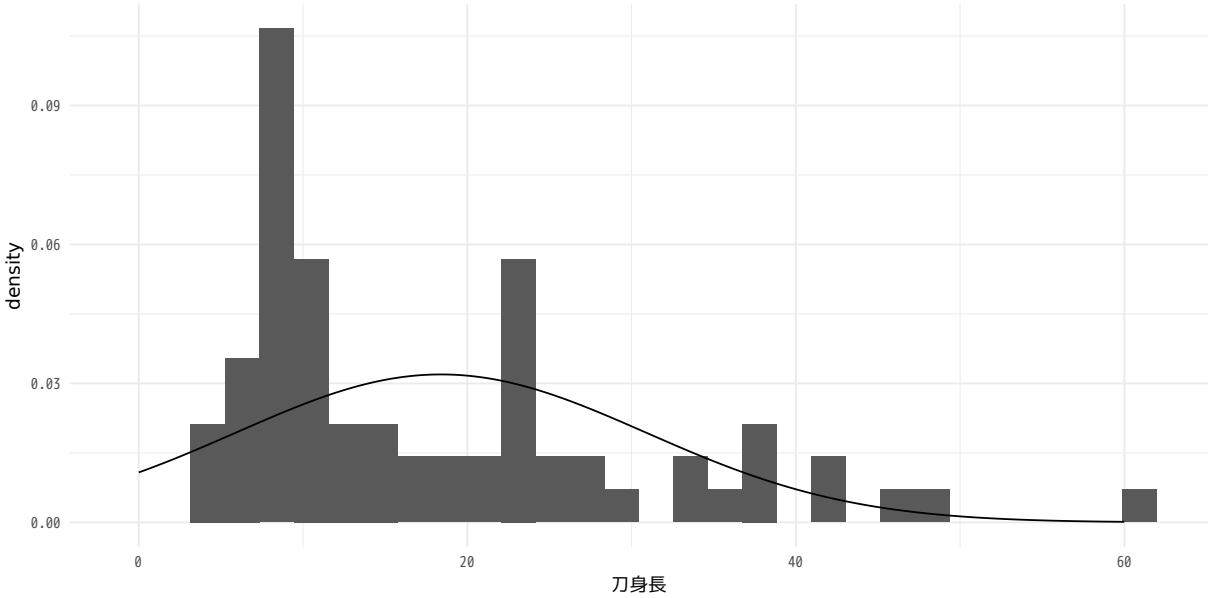
```
# 正規曲線作成のための統計量
iron%>%
    summarise(mean=mean(刀身長,na.rm=T),sd=sd(刀身長,na.rm=T))%>%
    kable()
```

mean	sd
18.4003	12.49029

```
# 正規曲線作成
x<-seq(0, 60, 0.1)
nom <- x%>%dnorm(mean=18.40, sd=12.49)
nom2<-data.frame(X=x,Y=nom)
#正規曲線付きヒストグラム
p<-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()
print(p)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



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します。これは、散布図では絶対に表現できません。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ことは確信できそうです。

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ます。正規分布とは何か、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います。

エクセルでヒストグラム

もう一つの大きな理由は「エクセルでヒストグラムを作りにくい」ということかもしれません。エクセルでヒストグラムを作れないわけではないのですが、度数分布表から棒グラフとして作成することになるので、一手間かかります。

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す。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ます。

構成比のグラフ表現

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ます。

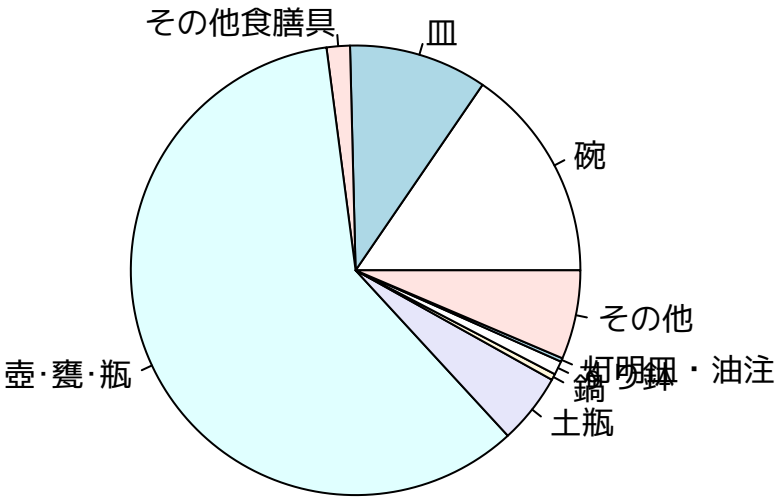
```
# データ読み込み
toj<-read.csv("data/pot.csv")
# データの順序定義
toj$器種<-toj$器種%>%
    factor(levels=c("碗","皿","その他食膳具","壺・甕・瓶",
        "土瓶","鍋","すり鉢","灯明皿・油注","その他"))
toj%>%head()%>%kable()
```

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

棒グラフは離散量を表現するために使われます。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際に使用する可視化手法を紹介します。もっとも大切なことは、 **円グラフを使わない**ということです。

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ますが、面積の大小を認識するのは苦手です。円グラフは面積で比率を表すので比率の比較には向いていないのです。

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



特に 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）見せるための手法です。公文書や学術的な報告では**絶対に**使うべきものではありません。

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ます。

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

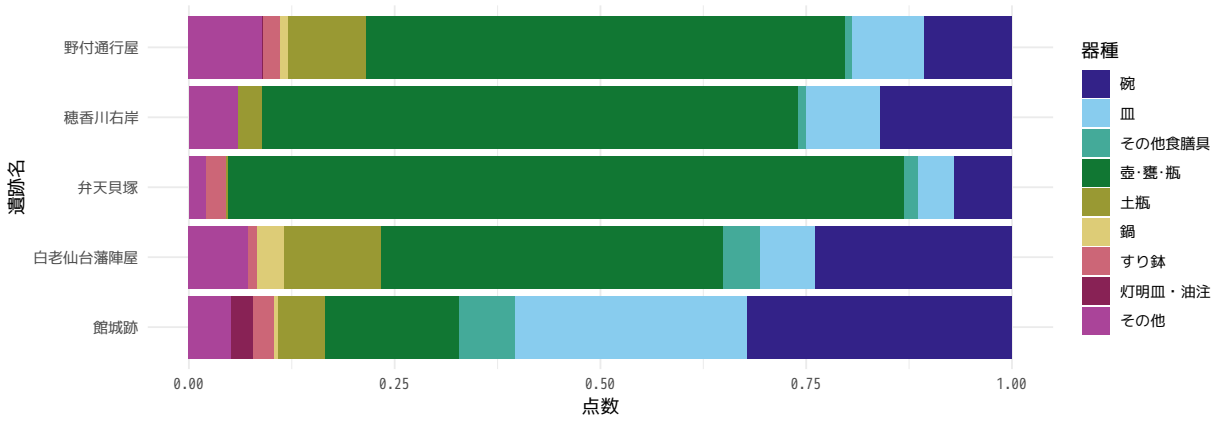
円グラフは不適切な可視化手法です。人間の目は直線的な形状の判断には優れていますが、面の比較は苦手です。円グラフで表現できるデータは棒グラフやドットチャートで表現するべきです。

「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 1985,p264）

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す。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す。構成比棒グラフは比率を比較するには非常に優れたグラフ表現です。

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

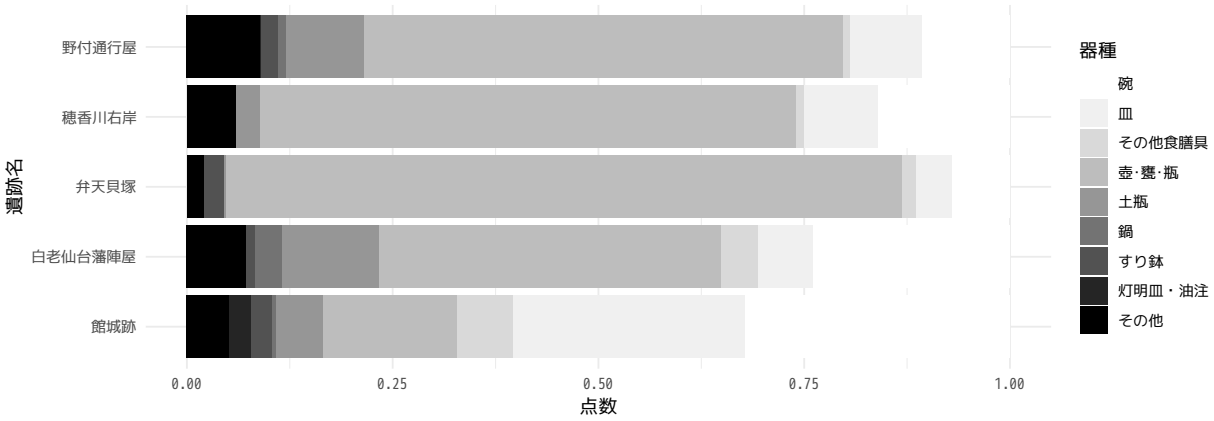


ただし、発掘調査報告書でカラーグラフが掲載できるケースは稀で、大半はグレースケールで表現されることになります。下のグラフはモニター上ではなんとか識別できますが、オフセット印刷の仕上がりでこれを識別することは不可能です。凡例との対比は特に困難です。

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ぎりぎりなので、構成比では 4 群～5 群が限界となります。

```
p<-toj%>%
    ggplot(aes(x=遺跡名,y=点数,fill=器種))+
```

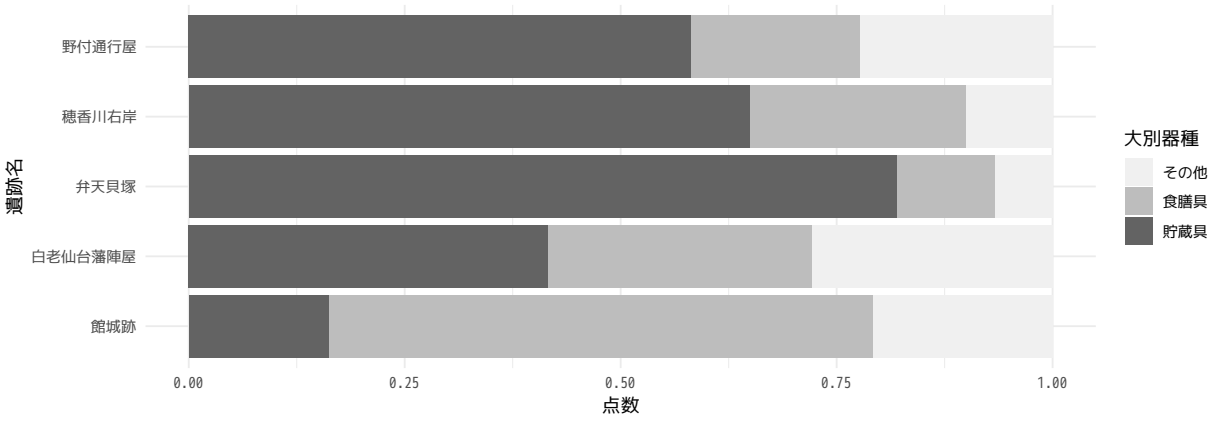
```
        geom_bar(stat="identity",position="fill")+
        coord_flip()+
        scale_fill_brewer(palette="Greys")+
        theme_minimal()
print(p)
```



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す。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考えるべきです。３群まで減らせばオフセット印刷でも対応可能なグレースケールのグラフになります。

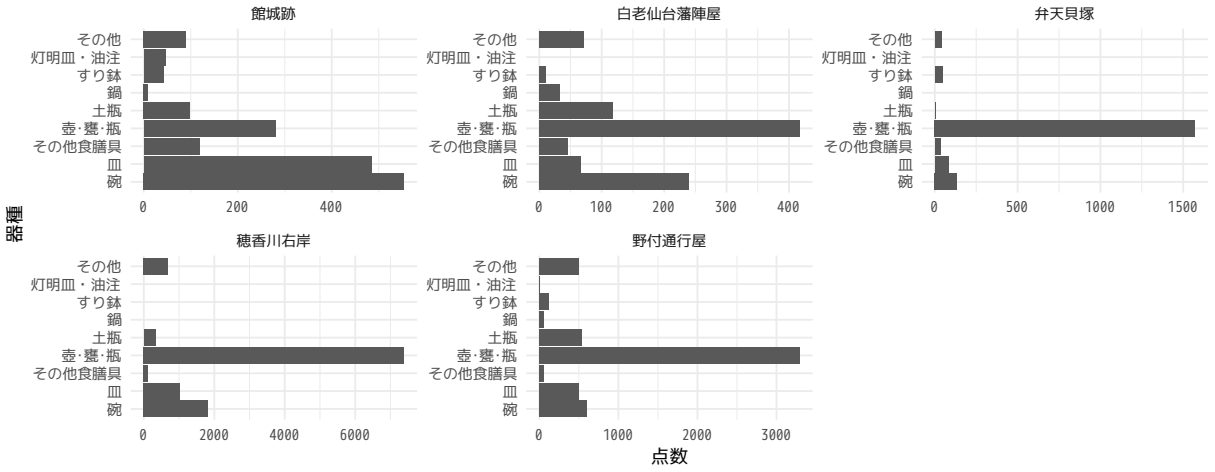
```
# 食膳具、貯蔵具、その他に区分
toj2<-toj%>%
    mutate(大別器種 = case_when(
        grepl("碗", 器種)|grepl("皿", 器種)|grepl("その他の食膳具", 器種)  == TRUE ~ "食膳具",
        grepl("壺· 甕· 瓶", 器種) == TRUE ~ "貯蔵具",
        grepl("灯明皿・油注", 器種)|grepl("その他", 器種)|grepl("すり鉢", 器種)|
            grepl("鍋", 器種)|grepl("土瓶", 器種) == TRUE ~ "その他",
        ))
# ３区分の構成比棒グラフ
p<-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()
print(p)
```



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します。花粉分析などの分析結果でよく見る形のグラフです。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ますし、オフセット印刷原稿としても対応可能です。

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



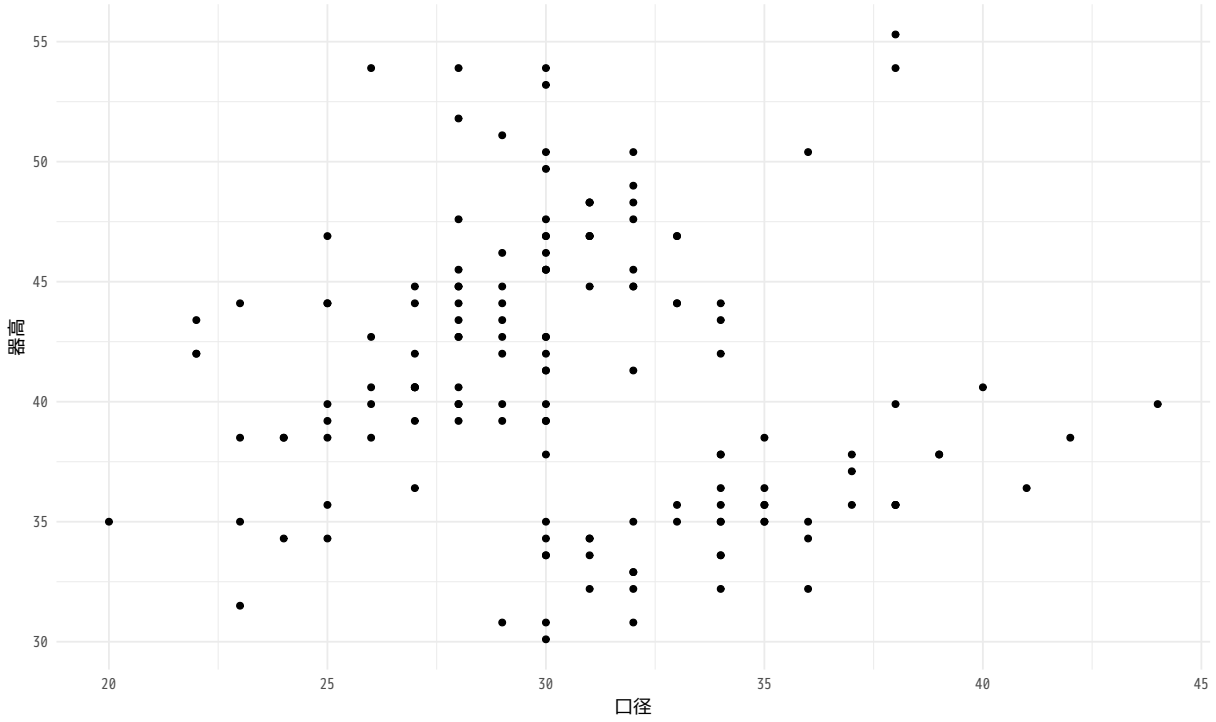
散布図

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ます。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ません。しかし、散布図が最も得意とする「二変量の関係を可視化する」という用途に使われることが案外少ないように思います。

ヒストグラムの項で用いた土器のデータを使用します。分析の目的は、口径と器高の関係を調べることです。

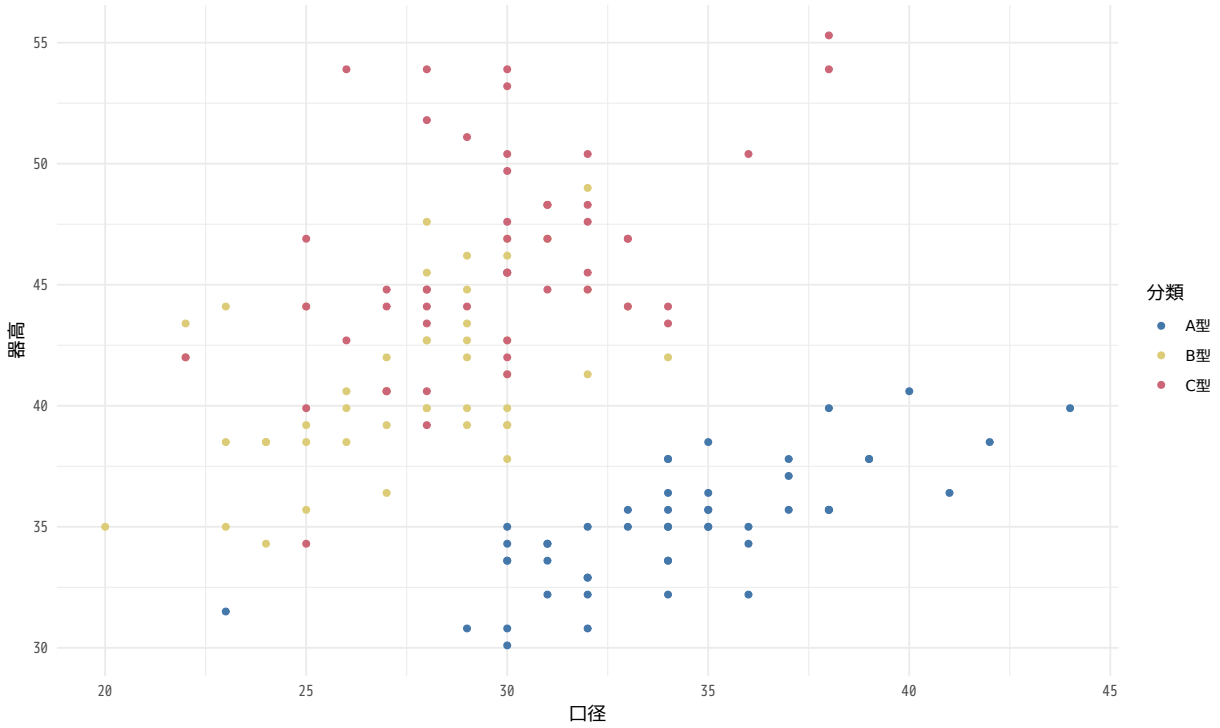
```
p<-pot%>%
    ggplot(aes(x=口径,y=器高))+
```

```
        geom_point()+
        theme_minimal()
plot(p)
```

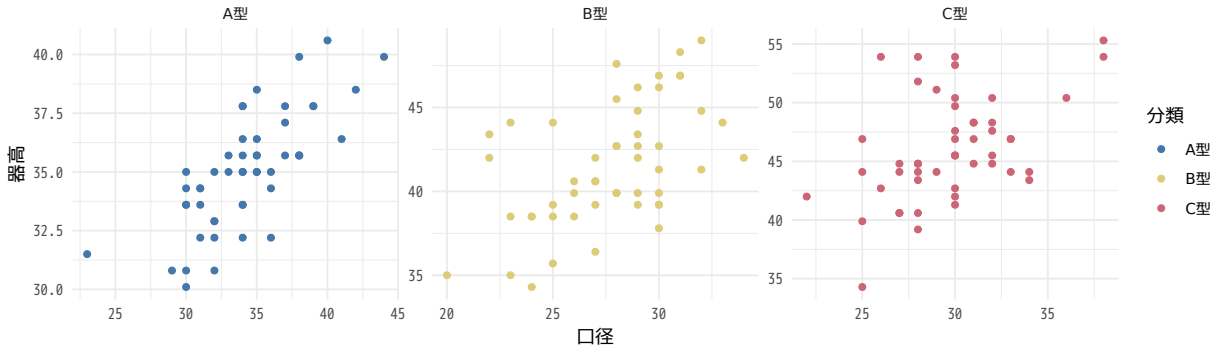


土器の分類によって口径と器高が異なると予想されるため、分類群ごとに調べた方が良さそうです。

```
# 同じグラフに描画
p<-pot%>%
    ggplot(aes(x=口径,y=器高,colour=分類))+
        geom_point()+
        scale_colour_ptol()+
        theme_minimal()
plot(p)
```




```
# 分類群ごとにファセットして描画
p<-pot%>%
    ggplot(aes(x=口径,y=器高,colour=分類))+
        geom_point()+
        facet_wrap(~分類,scales="free")+
        scale_colour_ptol()+
        theme_minimal()
plot(p)
```



散布図を描く場合に同一の領域に色やシェープを変えて描画する場合が多いように感じます。もちろん間違いではありませんが、「二変量の関係を可視化する」という目的では群ごとにファセットする方が特徴を明確に捉えることができます。特に発掘調査報告書や雑誌掲載原稿ではカラー図版を使えないケースが多いと思いますので、群ごとにファセットしたほうが読み取りやすいグラフ表現になります。

「二変量の関係」とは？

散布図の評価が難しいのは、「二変量の関係」を何らかの数学的なモデルに近似して表現しなければならないからです。日頃から数式に慣れている研究者ならともかく、私たち考古学研究者は数学的な訓練をあま

り受けていません。そうしたことから、散布図→二変量の関係の読み解き→モデル式の当てはめ　という手順を踏んで統計処理を進めていないケースが多いように感じています。

「二変量の関係」を記述する

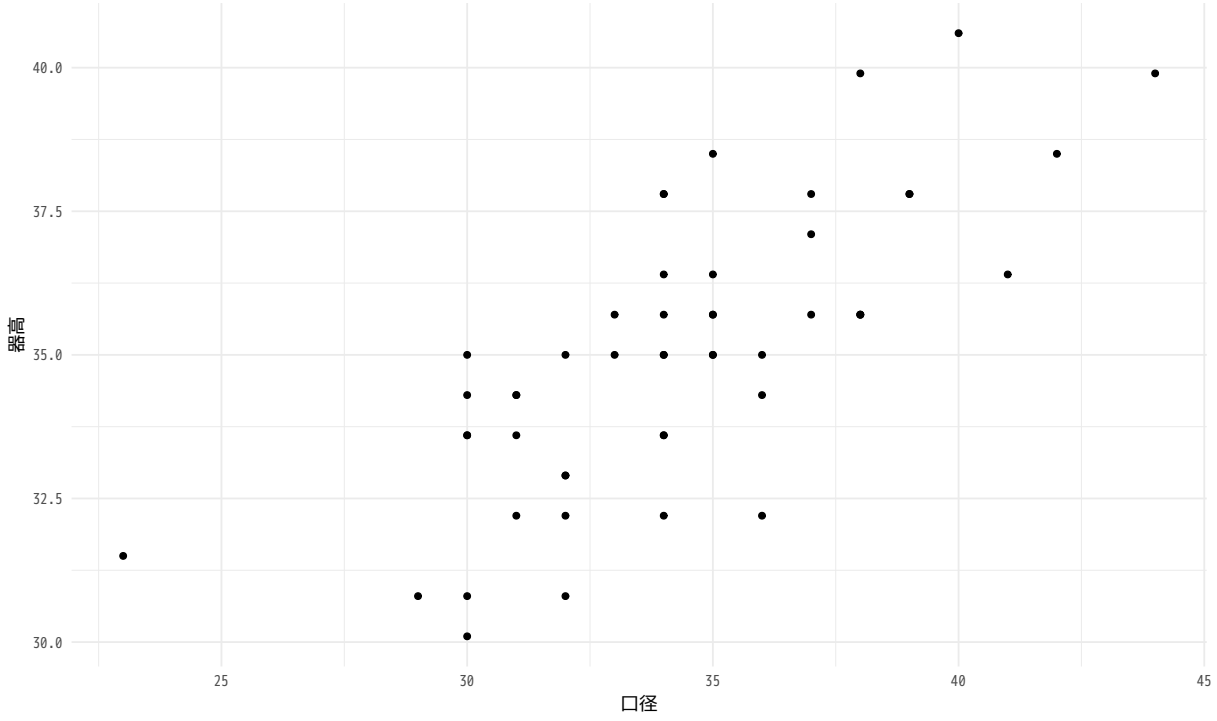
まず、考えるべきことは**二変量の関係が一次式に当てはまるか否か**ということです。

二変量の関係を考えるとき、少し統計に詳しい方は「相関係数」を算出しようと考えると思います。しかし、相関係数は二変量の関係を一次式をモデルとして一次式との一致度合いを計るものです。二変量の関係が２次式や反比例式で近似できる場合にはあてはまりません。

散布図を読み解く

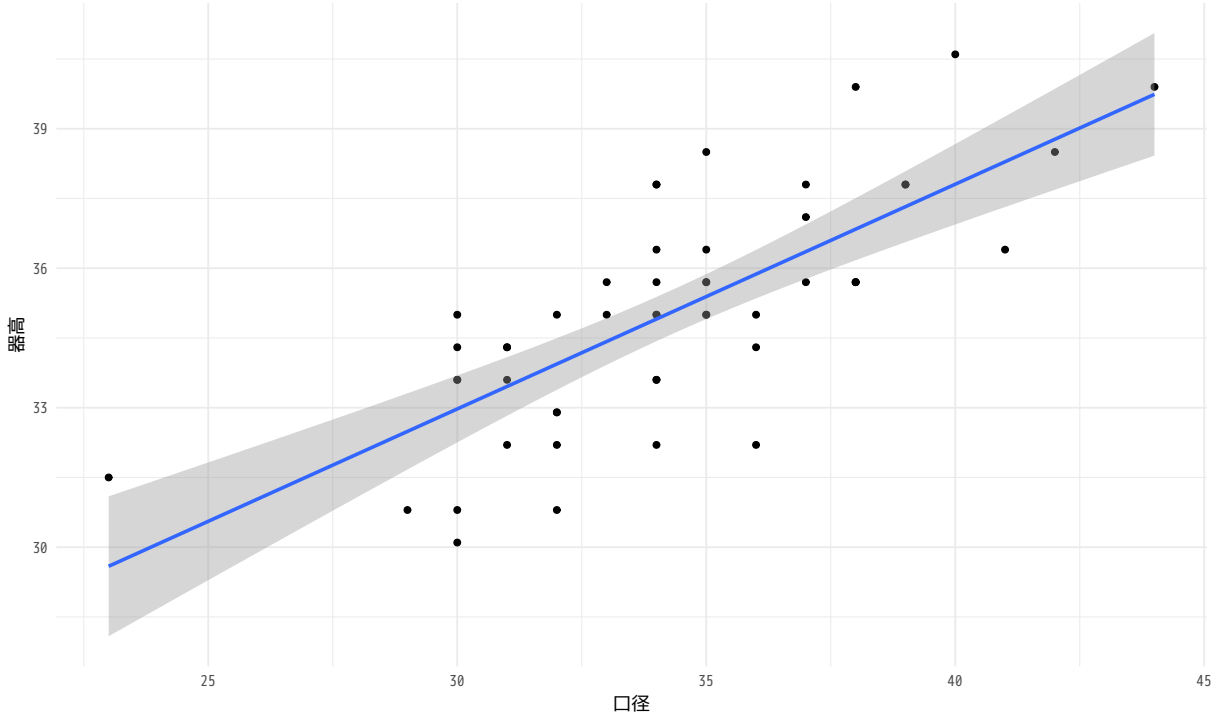
あらためて口径と器高の散布図を読み解きます。下図は分類「A 型」の散布図です。

```
p<-pot%>%filter(分類=="A 型")%>%
    ggplot(aes(x=口径,y=器高))+
        geom_point()+
        theme_minimal()
print(p)
```



どちらかといえば、直線的な分布に見えます。一次式を当てはめると以下のようになります。

```
p<-pot%>%filter(分類=="A 型")%>%
    ggplot(aes(x=口径,y=器高))+
        geom_point()+
        geom_smooth(method="lm")+
        theme_minimal()
print(p)
```



上で当てはめた一次式は下のようにして求められます。

```
coe<-lm(器高 ~ 口径,data=subset(pot, 分類=="A 型"))%>%summary()
coe$coefficients%>%kable()
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	18.4730087	2.1701002	8.512514	0
口径	0.4833428	0.0629292	7.680738	0

一次式に当てはめると次の式になります。

$y = 0.48x + 18.47$

因果関係を可視化する

「二変量の関係を可視化する」ということの目的は、究極的には「因果関係の可視化」です。

たとえば、学力と子どもの環境の因果関係を統計的に考える場合を考えると、「学力テストの点数」という変量「果」に対して「因」となる変量は「親の収入」や「TV の視聴時間」、「睡眠時間」などが考えられます。

したがって、散布図を描く前に考えることは「因果」の「因」にあたる変量と「果」に当たる変量が何か、ということです。少なくとも「因」にあたる変量がはっきりしないデータには、そもそも散布図を描く価値はない、と断言できます。

独立変数と従属変数

散布図を描く場合の約束として、因果関係の「果」にあたる y 軸に、「因」にあたる変量を x 軸に割り当てます。 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

独立変数と呼びます。

刀身長はどうやって決まるか

散布図行列

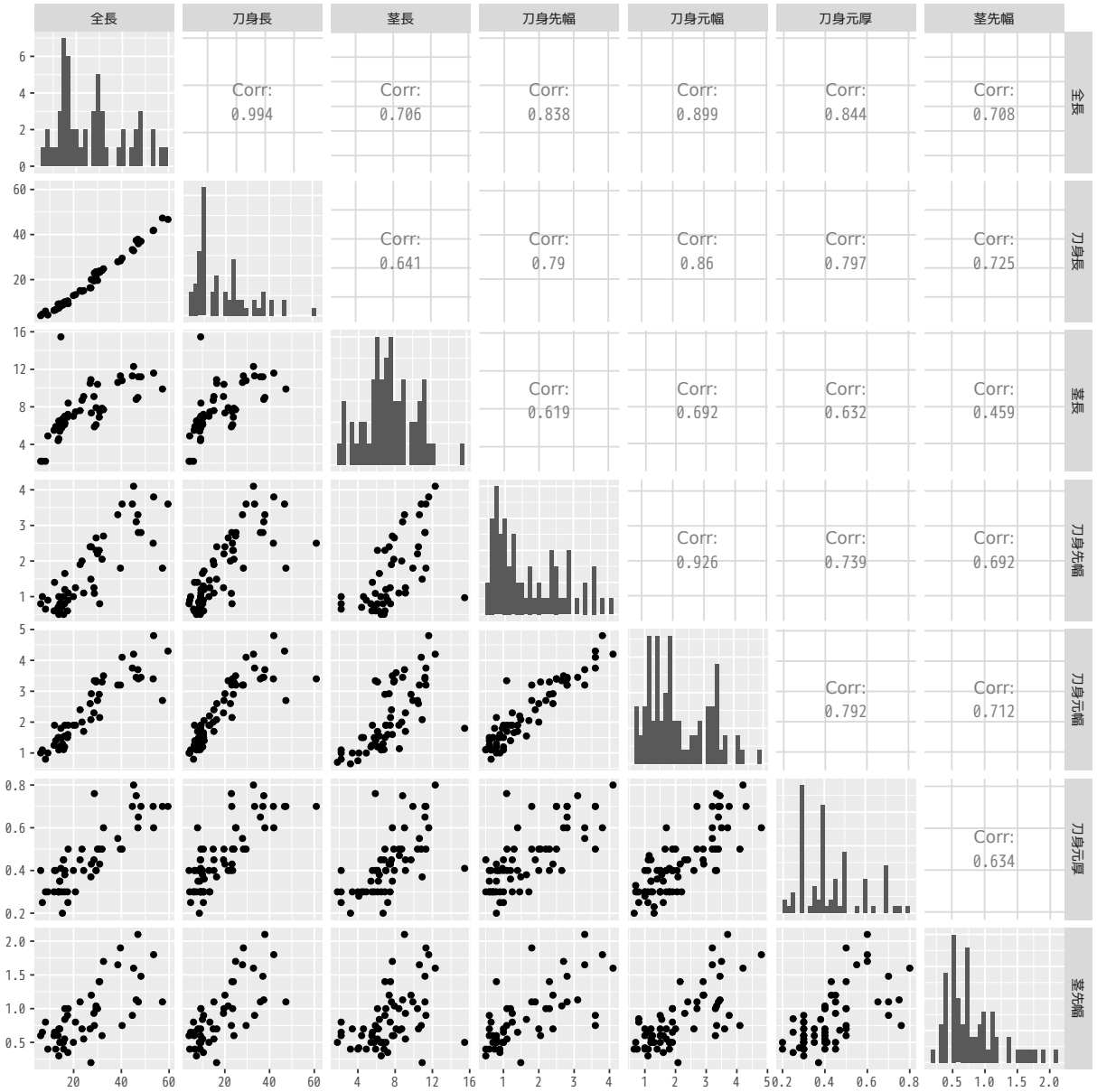
恵庭西島松 5 遺跡出土の古代刀剣を対象としたデータで散布図を作成します。追求すべきテーマは「刀身長と他の属性との因果関係」です。

刀身の長さは端的に刀剣のサイズを示すものです。刀剣をつくるときには、まず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ます。

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ります。

GGally パッケージを利用して散布図行列を描画します。

```
# GGally パッケージ読み込み
library(GGally)
#
p<-iron%>%select(全長, 刀身長, 茎長, 刀身先幅, 刀身元幅, 刀身元厚, 茎先幅)%>%
    ggpairs(diag=list(continuous="barDiag"))
print(p)
```



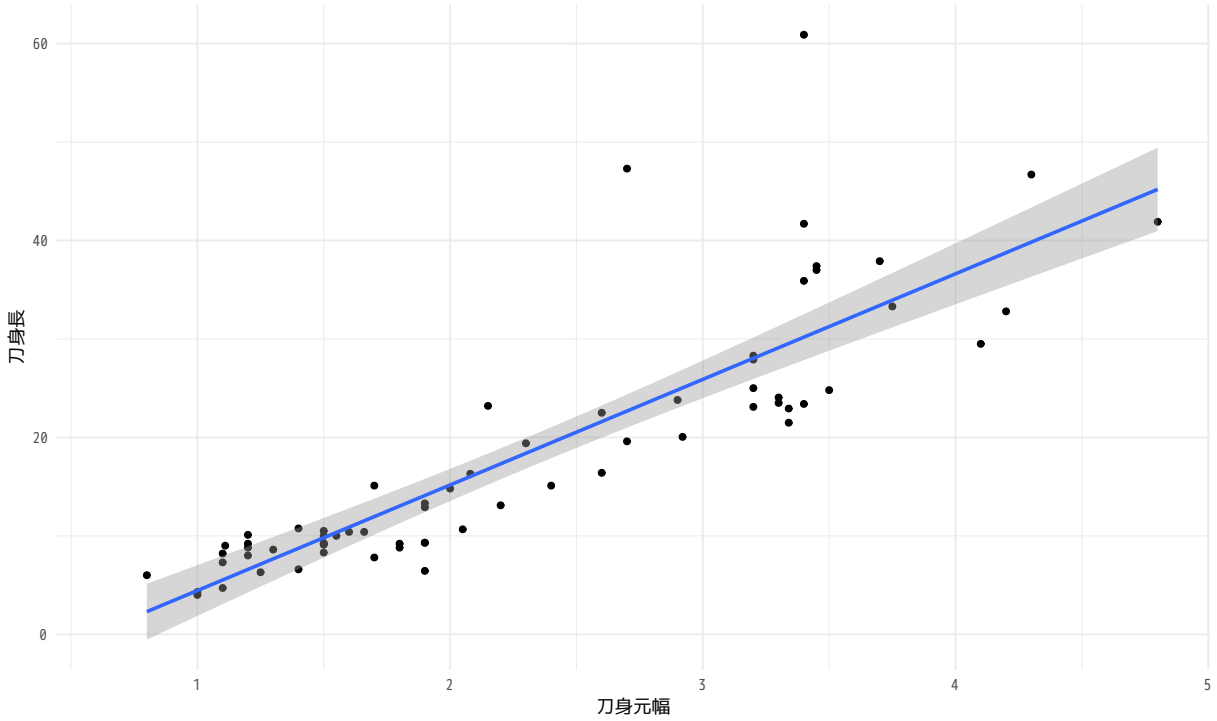
散布図が示すところからは、多くの属性が刀身長と相関関係にあることが読み取れます。一方、刀身元幅のように非常に強い相関を示す変量もあれば、茎先幅のように相関が弱い変量もあります。刀身長との相関の強弱を判断することで、古代剣製作にかかる規範意識を読み取ることが可能かもしれません。

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した。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す。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します。出土刀剣では刀身が破損せずに出土することはまれですから、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す。

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

```
print(p)
```



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す。

```
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

$y = 10.72x - 6.28$

まとめ

- 連続量のデータは、まず**分布**を調べる→ヒストグラム
- 構成比（離散量）のデータに円グラフは使わない
- 離散量は棒グラフ
- 連続量どうしの関係は散布図で可視化する