



データマイニング

Data Mining

3: データの要約・前処理 4: 次元削減

3: Data Summarization, Preprocessing 4: Dimension Reduction

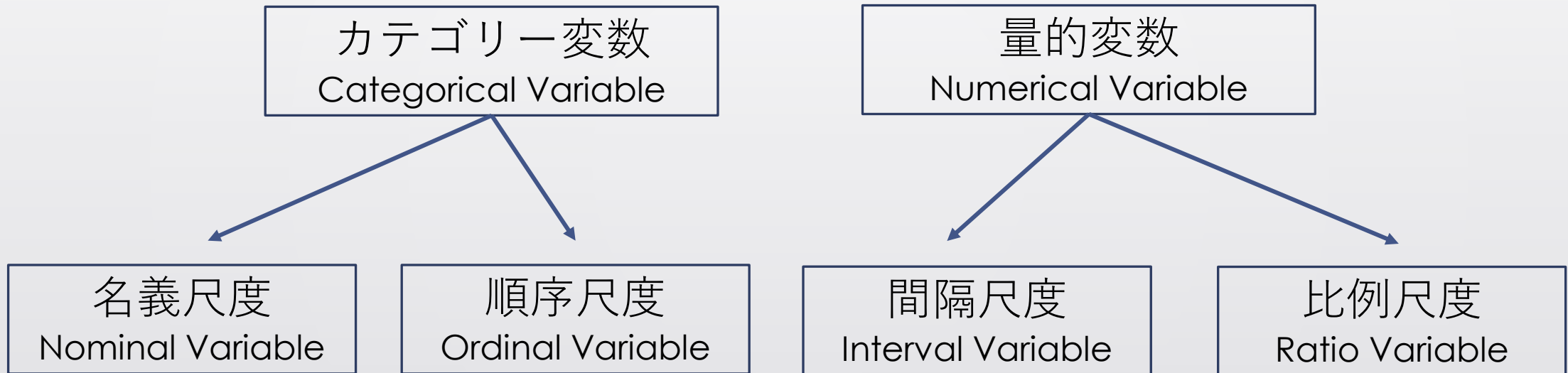
土居 裕和 Hirokazu Doi

長岡技術科学大学 Nagaoka University of Technology



データの要約
Data Summarization
&
Getting to know your data

変数の種類 Types of Variables





変数の種類 Types of Variables

名義尺度
Nominal Variable

あるカテゴリーを、別のカテゴリーと区別するために用いられる、数値自体には意味がない変数

Variables, whose number has no numerical value, often used to discriminate multiple categories

順序尺度
Ordinal Variable

順序を表す変数。変数間の間隔には意味がない。

Variables representing ordering of categories. Intervals between variables do not have any meanings.

変数の種類 Types of Variables

間隔尺度
Interval Variable

変数間の差の値に意味がある変数
Variables whose difference have meanings

比例尺度
Ratio Variable

間隔尺度と似ているが、原点がある点が違う。
Similar to interval variables, but ratio variables have clearly defined point of zero.

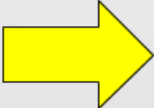
ユーザーID	性別	年齢	年収
sanapon	男	26	411万円
oggi1985	女	33	536万円
murachan	女	39	681万円
shozan.s	男	23	309万円

//////
名義尺度の数値的表現

Numerical representation of nominal variable

ダミー変数 Dummy Variable 男 \Rightarrow 1, 女 \Rightarrow 2

One-hot Encoding

Color		Red	Yellow	Green
Red				
Red		1	0	0
Yellow		1	0	0
Green		0	1	0
Yellow		0	0	1

<https://www.kaggle.com/code/dansbecker/using-categorical-data-with-one-hot-encoding/notebook>



記述統計

Descriptive Statistics

代表値 Representative value

平均 Mean, 中央値 Median, 最頻値 Mode

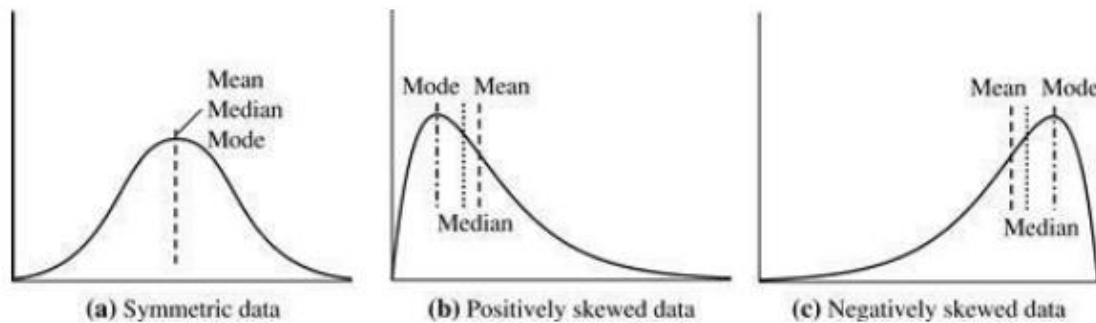


FIGURE 2.1 Mean, median, and mode of symmetric versus positively and negatively skewed data.

算術平均

Arithmetic Mean

$$M = \frac{\sum_{i=1}^n x_i}{n}$$

中央値

順番に並べた時, 中央に位置する数値

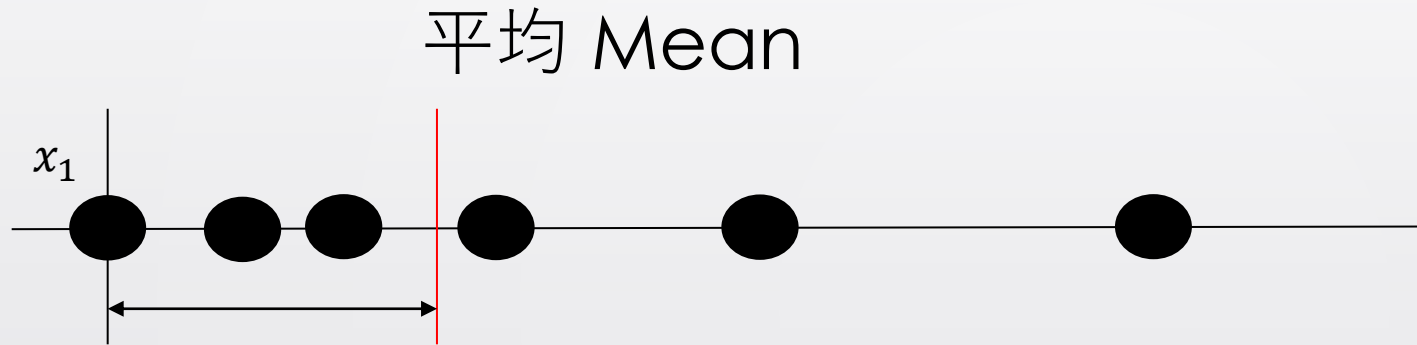
Value lying at the midpoint when a number sequence is ordered in ascending/descending order

最頻値

最も頻繁に現れる数値

The value that appears most frequently in the sequence

分散と標準偏差 Variance and Standard Deviation



Deviation of x_1 from the arithmetic mean

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

標準偏差 = $\sqrt{\text{分散}}$

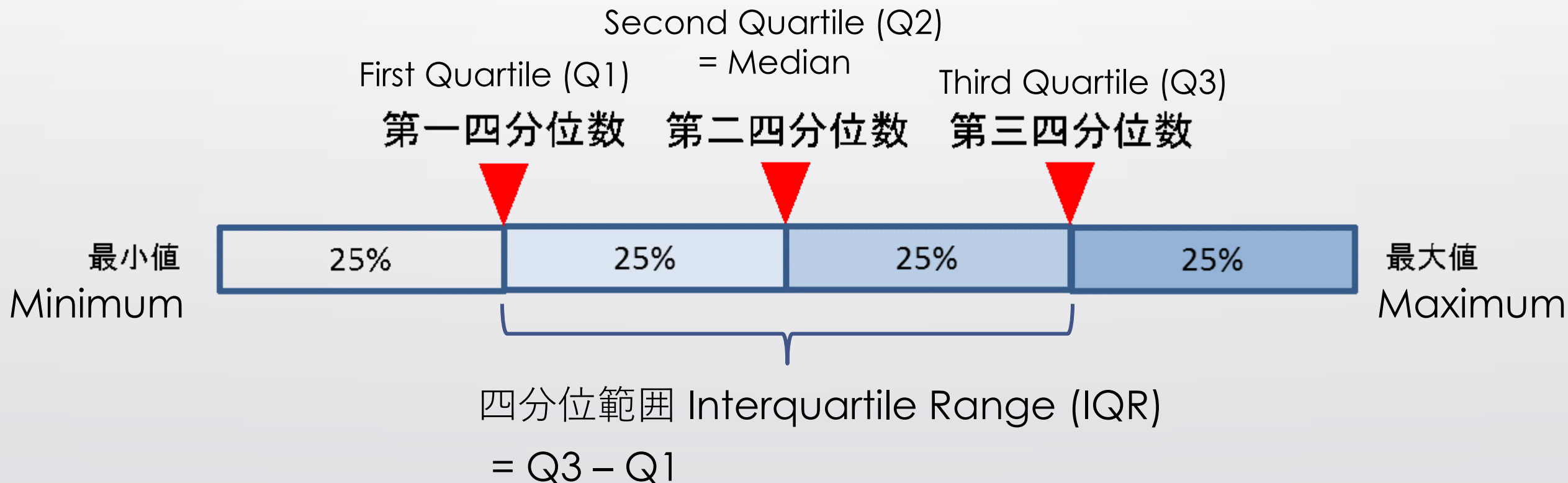
Standard Deviation = $\sqrt{\text{Variance}}$

分散と行列 Variance and Matrix

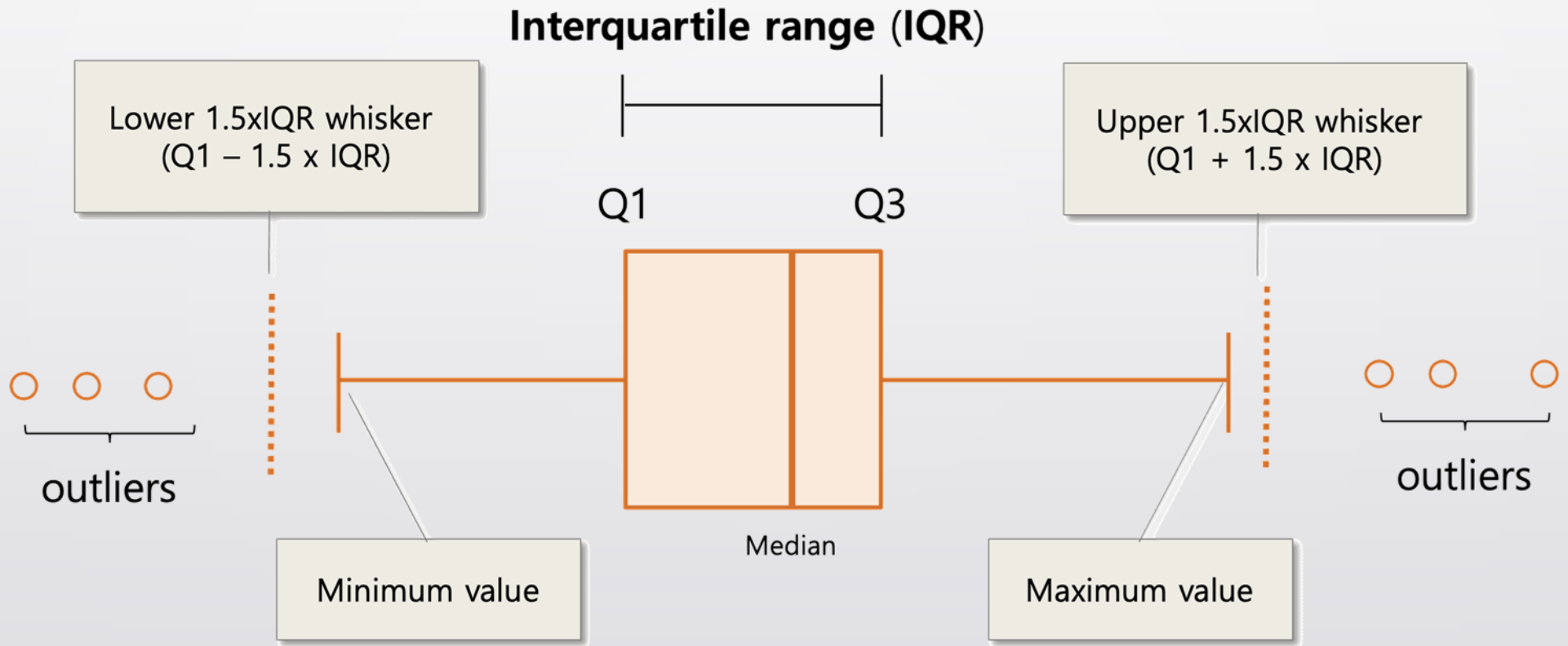
$$\mathbf{X} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_{n-1} \\ x_n \end{bmatrix} \quad \bar{\mathbf{X}} = \begin{bmatrix} \bar{x} \\ \bar{x} \\ \bar{x} \\ \vdots \\ \bar{x} \\ \bar{x} \end{bmatrix} \quad \mathbf{X} - \bar{\mathbf{X}} = \begin{bmatrix} x_1 - \bar{x} \\ x_2 - \bar{x} \\ x_3 - \bar{x} \\ \vdots \\ x_{n-1} - \bar{x} \\ x_n - \bar{x} \end{bmatrix} \quad (\mathbf{X} - \bar{\mathbf{X}})^T = [x_1 - \bar{x}, x_2 - \bar{x}, \dots, x_n - \bar{x}]$$

$$V = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n} = \frac{1}{n} [x_1 - \bar{x}, x_2 - \bar{x}, \dots, x_n - \bar{x}] \begin{bmatrix} x_1 - \bar{x} \\ x_2 - \bar{x} \\ x_3 - \bar{x} \\ \vdots \\ x_{n-1} - \bar{x} \\ x_n - \bar{x} \end{bmatrix} = \frac{1}{n} (\mathbf{X} - \bar{\mathbf{X}})^T (\mathbf{X} - \bar{\mathbf{X}})$$

四分位数 Quartile



箱ひげ図 Boxplot



<https://help.ezbiocloud.net/box-plot/>

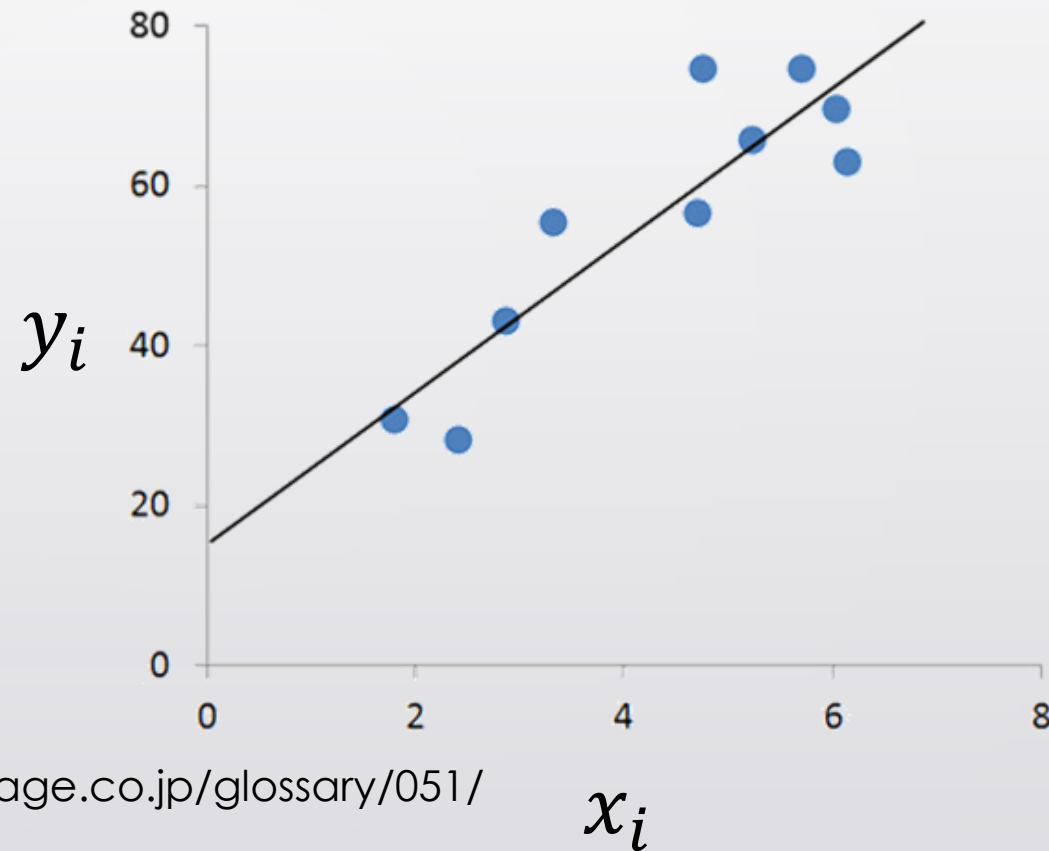
データの可視化: 散布図

Data Visualization: Scatter Plot

(x_i, y_i)

x_i : 家族の人数
Number of
Family Member

y_i : 購入数
Number of purchased
Items



<https://www.intage.co.jp/glossary/051/>

データの可視化: ヒストグラム Data Visualization: Histogram

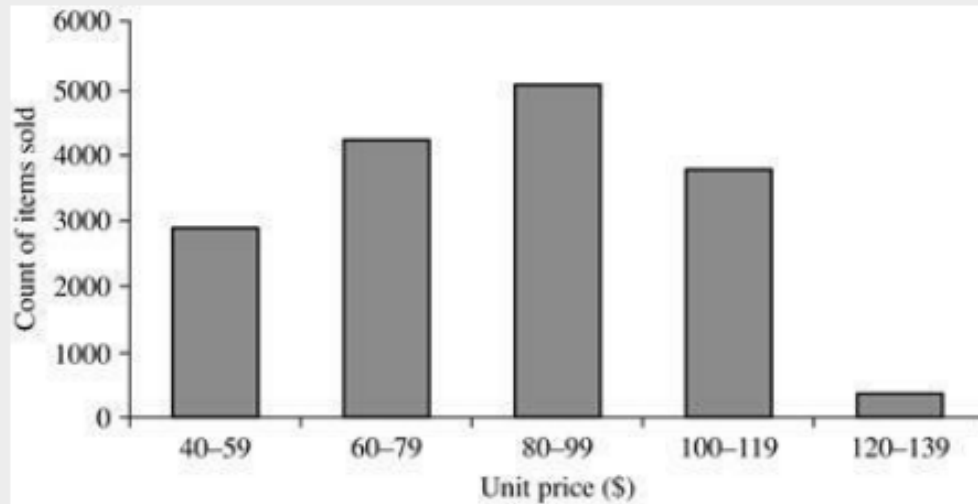


FIGURE 2.6 A histogram for the **Table 2.1** data set.

ヒストグラムを描くことで、

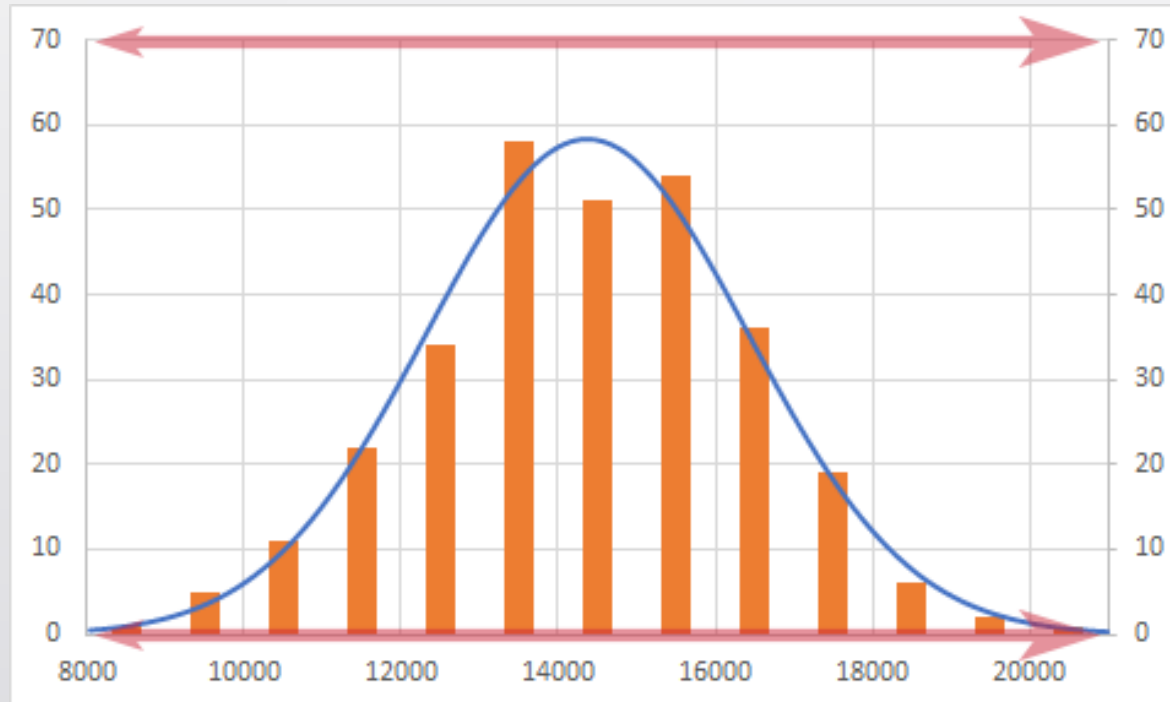
データの分布形状を把握できる

We can grasp data distribution by drawing a histogram

データの分布 Data Distribution

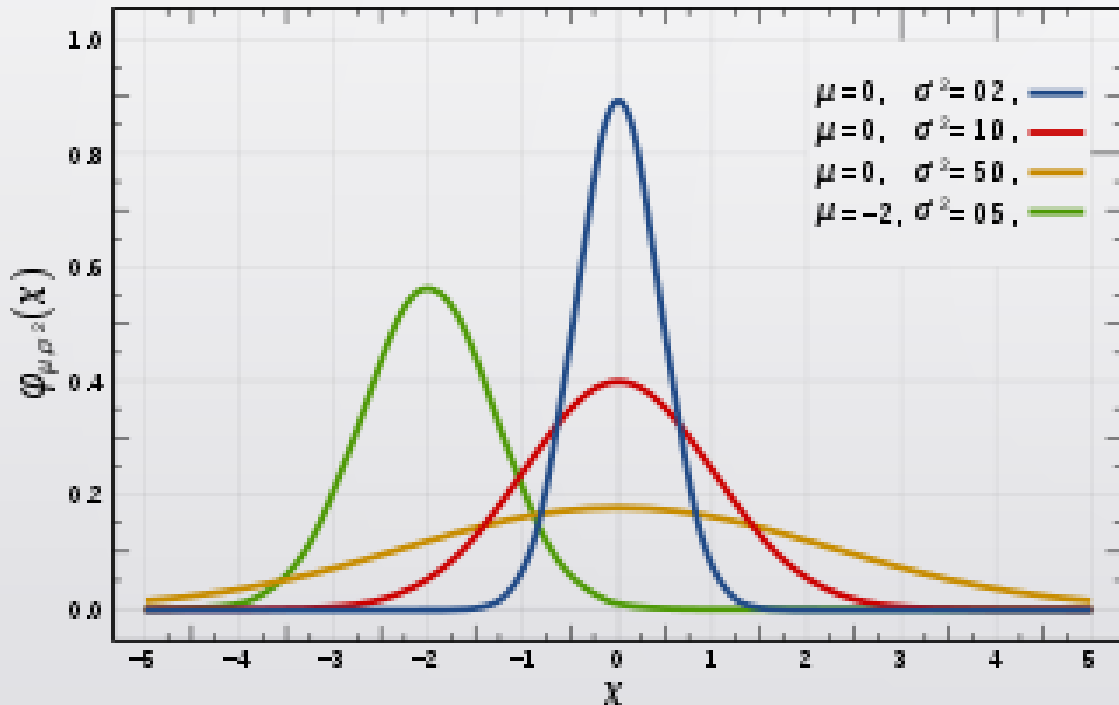
標本数が多ければ、度数分布は曲線で近似できる

With sufficient number of samples, the frequency distribution can be approximated by a curve



<https://bdastyle.net/tools/histogram/page6.html>

正規分布 Normal Distribution



確率密度関数 Probability Density Function

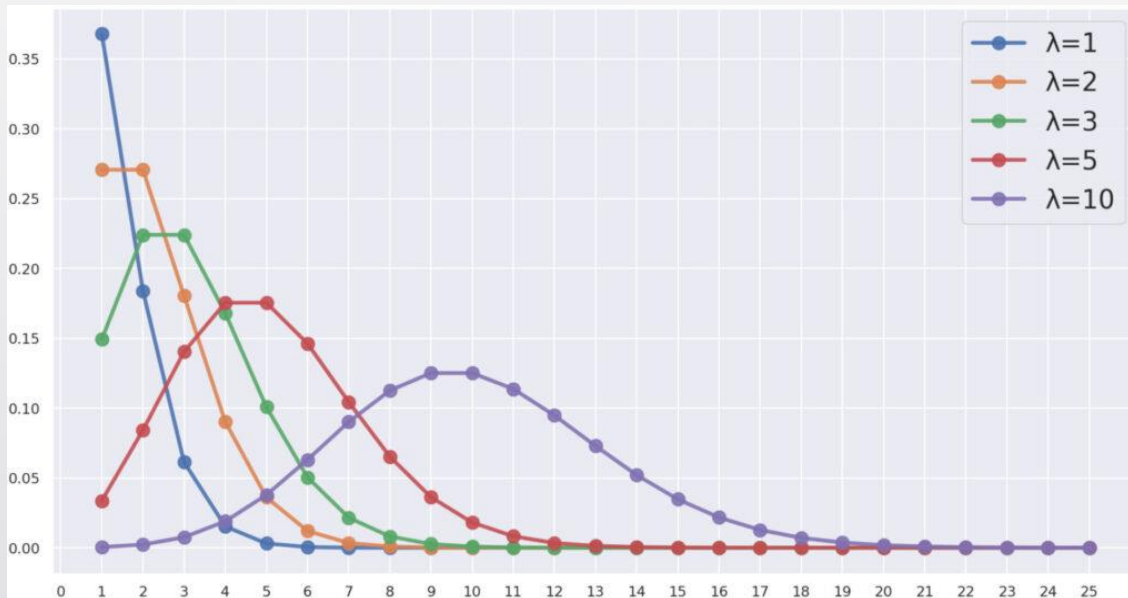
$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \quad (-\infty < x < \infty)$$

$\mu = 0, \sigma = 1$ の時は、標準正規分布

Standard normal distribution when $\mu = 0, \sigma = 1$

<https://ja.wikipedia.org/wiki/%E6%AD%A3%E8%A6%8F%E5%88%86%E5%B8%83>

ポアソン分布 Poisson Distribution



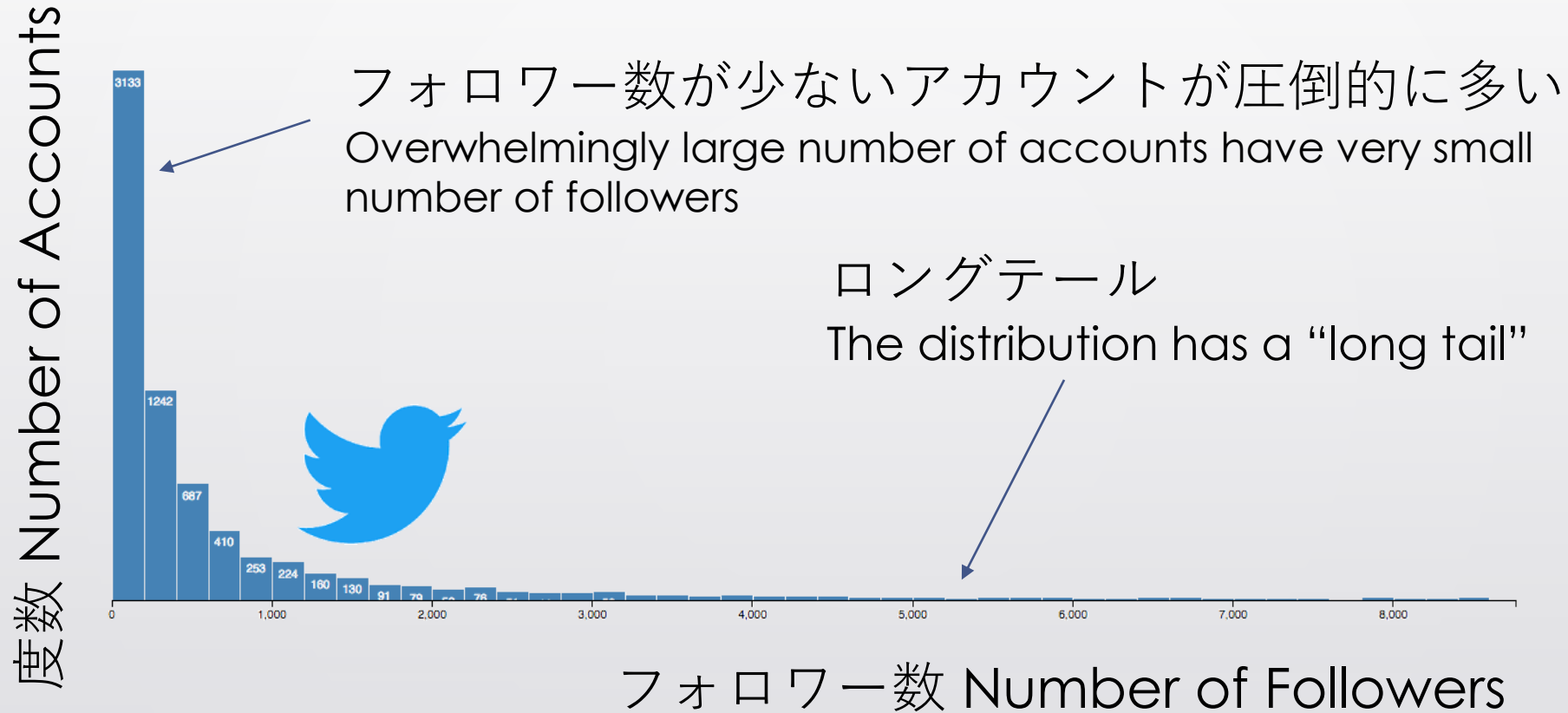
イベント発生回数

<https://mathlandscape.com/poisson-distrib/>

あるイベントが一定時間内に発生する確率を表す

Gives the probability with which a rare event happens within a certain time-window

指数分布 Exponential Distribution



<https://ultrasaurus.com/2015/05/distribution-of-twitter-followers/>

ロングテール戦略 Long Tail Strategy



<https://blogs.ubc.ca/kathzhang/2014/10/05/the-long-tail-theory-with-examples/>

Q-Qプロット Q-Q Plot

QQプロットにより、2つのデータセットが同じ分布に従うかどうかを検証できる

Q-Q plot tells us whether two dataset conforms to identical distribution

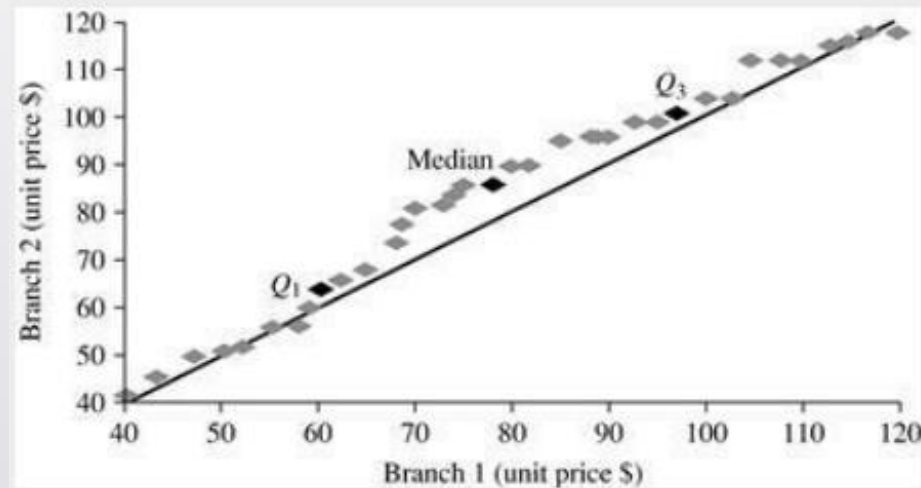
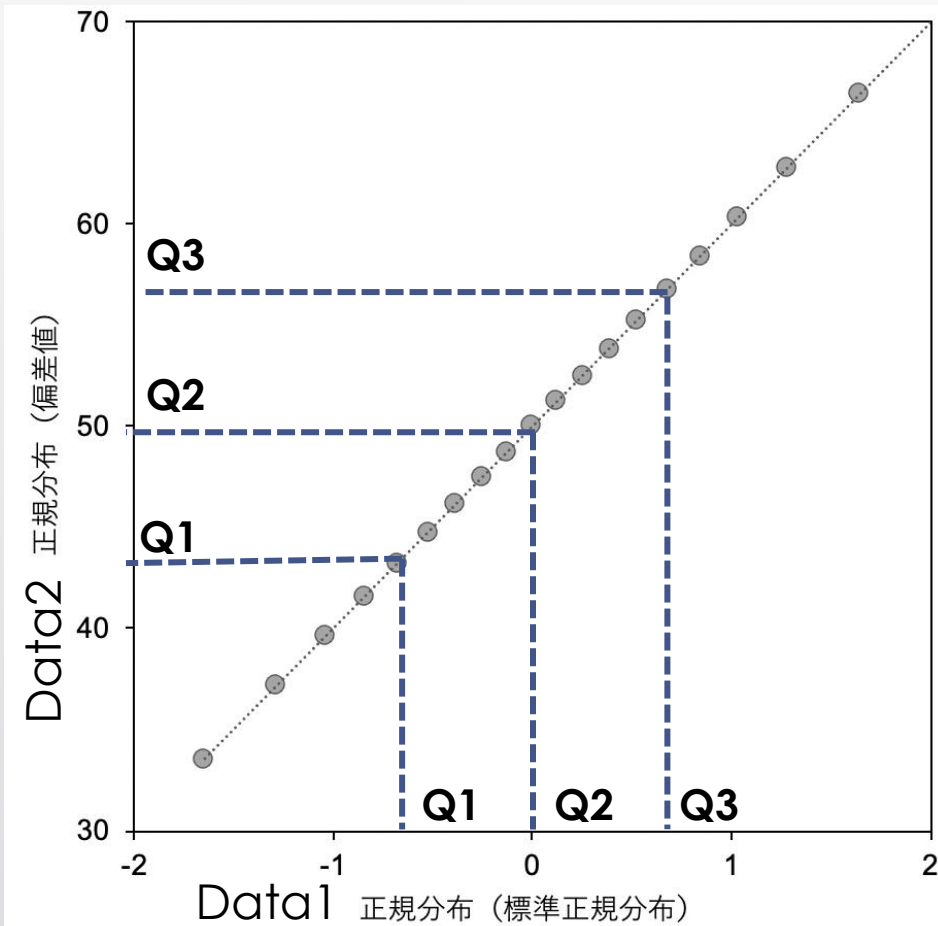


FIGURE 2.5 A q-q plot for unit price data from two *AllElectronics* branches.

Q-Qプロット Q-Q Plot



Data1 は標準正規分布に従う

Data1 comes from standard normal distribution

Data2 は平均50, 標準偏差10の正規分布に従う

Data2 comes from normal distribution with $\mu = 50, \sigma = 10$

データ分布が同じだと、スケールに関係なく、直線上に点がプロットされる

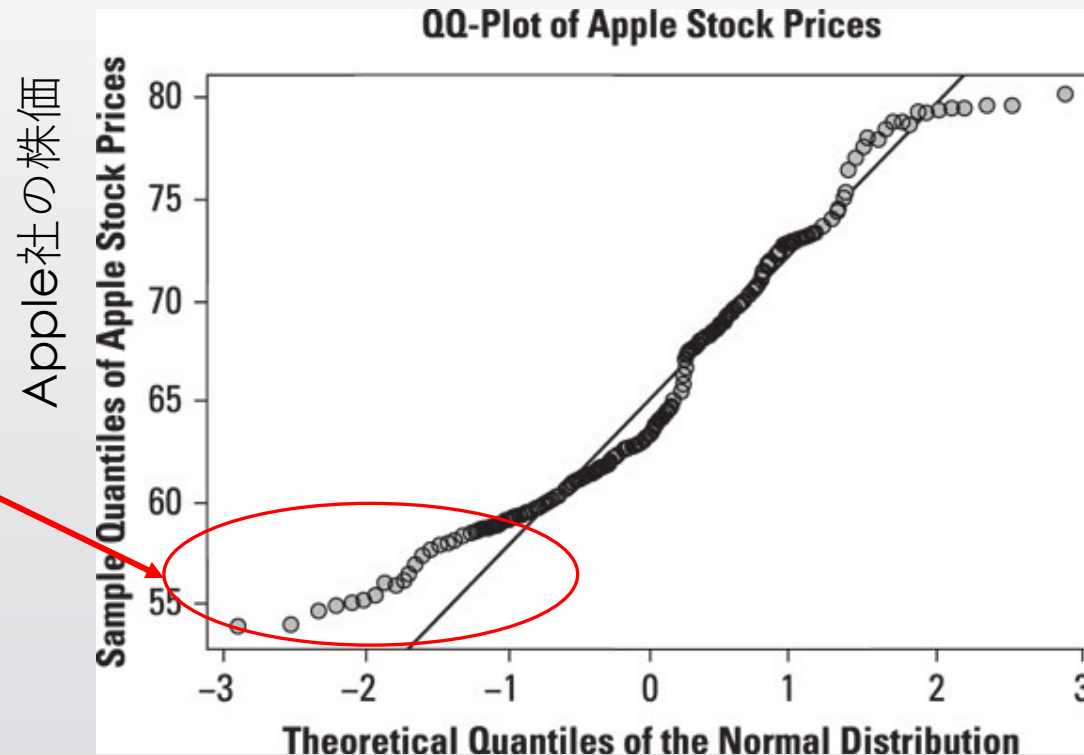
Points of datasets generated under identical distribution are plotted on a straight line irrespective of data scale

<https://best-biostatistics.com/eqr/q-q-plot-eqr.html>

Q-Qプロット Q-Q Plot

QQプロットは、しばしば、ある変数が正規分布に従うかどうかを検証するために使われる

QQ plot is often utilized to check if certain variable conforms to normal distribution



株価の底値は、正規分布で予想される価格よりも高い


Lowest stock price is lower than the price expected based on normal distribution

<https://www.dummies.com/article/technology/information-technology/data-science/big-data/quantile-quantile-qq-plots-graphical-technique-for-statistical-data-141221/>



前处理

Preprocessing



データマイニングの流れ Steps in Data Mining

1. 目標設定 Goal Setting
2. データ収集 Data collection
3. 前処理 Preprocessing
4. 特徴量選択 Feature Selection (必要ないケースもある; can be skipped in some cases)
5. データ分析 Data Analysis ・ モデリング Modeling
6. 性能評価 Performance Evaluation
7. (ディプロイメント Deployment)



データ前処理 Data Preprocessing

アルゴリズムが扱いやすい形式にデータを変換する
Transform dataset into a format easy for an algorithm to handle

欠損の補完 Interpolation of missing value

外れ値・重複除去 Deletion of outliers and redundancy
ノイズ除去 Noise Cleansing

離散化 Discretization

スケーリング Scaling

次元圧縮 Dimension Reduction

欠損値と重複の扱い

Handling of Missing Value and Redundancy

ユーザーID	性別	年齢	年収
sanapon	男	26	411万円
oggi1985	女	33	NaN
murachan	女	39	681万円
sanapon	男	26	411万円
shozan.s	男	23	309万円

欠損値 Missing value

重複 Redundant record

重複がある場合は、特別なケースを除いて、一方を削除する

When there is redundancy, delete either one of the record except for special cases

欠損値の補完 Interpolation of Missing Value

ユーザーID	性別	年齢	年収
sanapon	男	26	411万円
oggi1985	女	33	NaN
murachan	女	39	681万円
sanapon	男	26	411万円
shozan.s	男	23	309万円

削除することが多い

In many cases this row is deleted entirely

対応するデータを探し出す Somehow retrieve the corresponding value

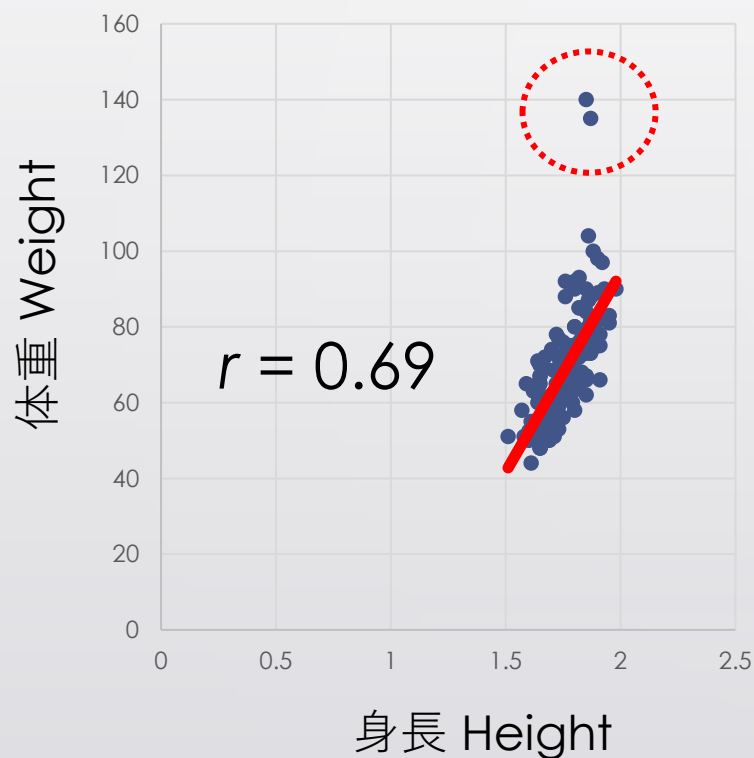
同じクラスの中央値や平均値で代替する Replace NaN with mean or median of the same class

回帰等の方法で補完する Interpolate the missing value by regression etc

外れ値 Outlier

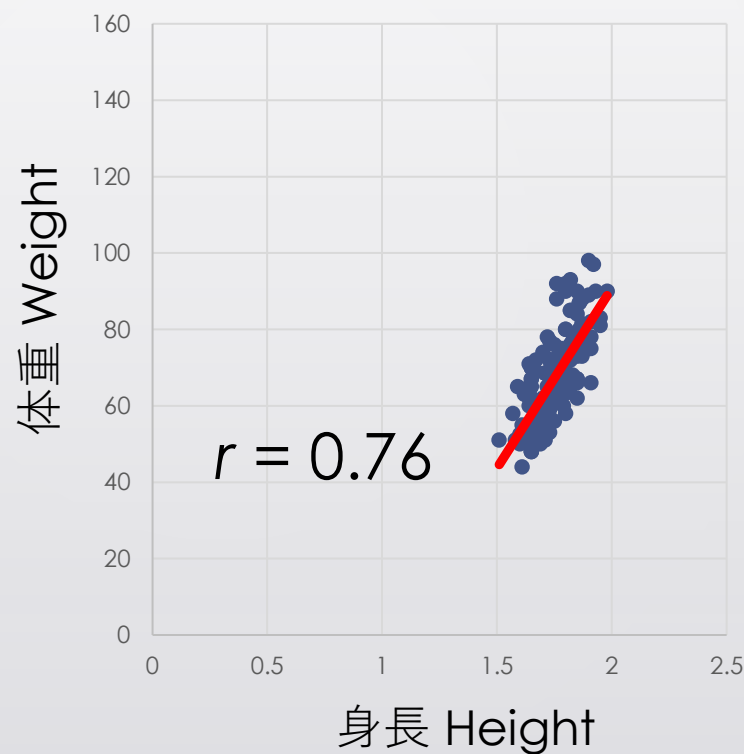
外れ値除去前

Before Outlier Deletion



外れ値除去後

Before Outlier Deletion



結果に影響することがあるので、特にサンプルサイズが小さい場合は、除去することがある

Since outliers potentially influence final results, they are sometimes deleted from the dataset especially when the sample size is relatively small

離散化 Discretization

数量を階級やカテゴリーに分割する

Replace numeric values with classes or categories

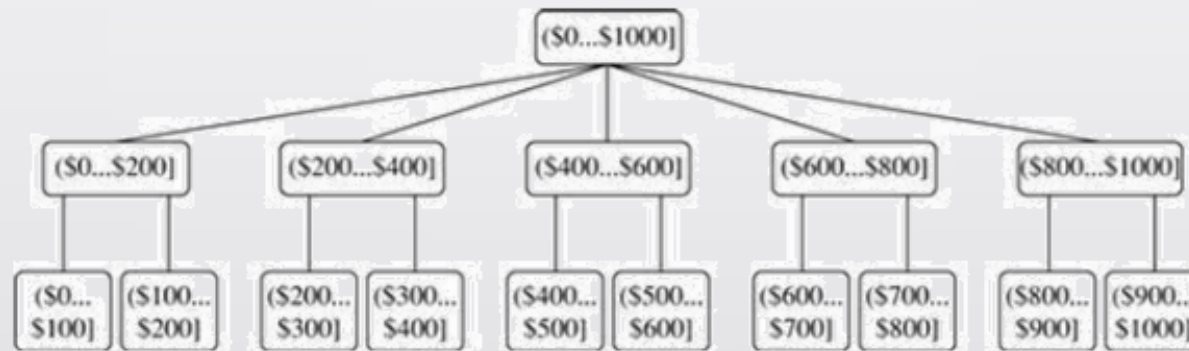


FIGURE 3.12 A concept hierarchy for the attribute *price*, where an interval $(\$X ... \$Y]$ denotes the range from $\$X$ (exclusive) to $\$Y$ (inclusive).

スケーリング Scaling

データの単位やレンジの違いが結果に影響することを防ぐため、
To mitigate potential influences of the difference in unit and range among variables,

データを標準化 /正規化する

Variables are sometimes *standardized/normalized*

Z-スコア化 Z-score normalization

$$x' = \frac{x - \mu}{\sigma}$$

Min-Max normalization

$$v'_i = \frac{v_i - \min_A}{\max_A - \min_A} (\text{new_max}_A - \text{new_min}_A) + \text{new_min}_A.$$

ロバストZスコア Robust Z-Score

Z-score化はデータが正規分布することを前提としている

Z-score transformation rests on the presumption that data conforms to normal distribution

ロバストZスコア化は正規分布の仮定を必要としない

Robust z-scoring does not presume normal distribution

$$\text{robust } z \text{ score} = \frac{x - \text{median}}{NIQR}$$

$$NIQR = \frac{IQR}{1.3489} \quad \text{データが正規分布するとき, } IQR \approx 1.3489\sigma$$



次元削減 Dimension Reduction

出来るだけ多くの情報を残しながら、多次元データを、低次元のデータに変換すること

Transform/compress multidimensional data into data with lower dimensions while retaining as much information as possible

計算・データ処理の高速化 Acceleration of computation and data processing

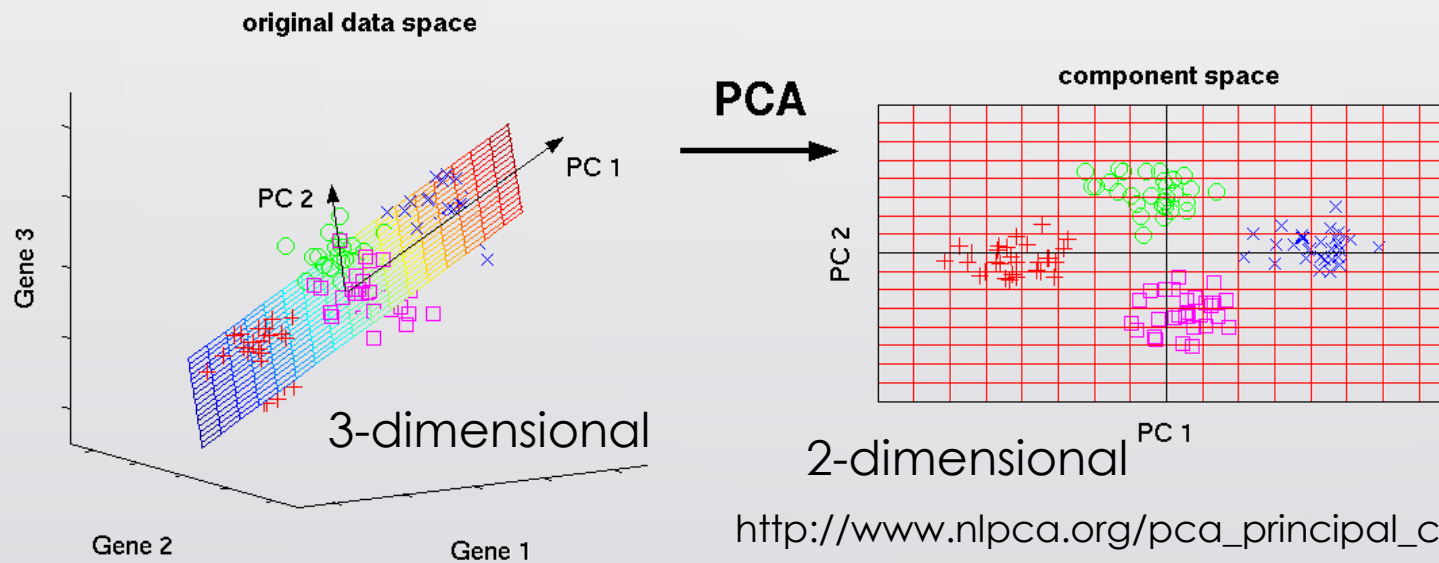
変数の直交化 Orthogonalization of variables

データを解釈しやすくする Enhance interpretability of the data (sometimes...)

主成分分析 Principal Component Analysis (PCA)

変数の線型和により新たな変数(主成分)を合成することで、次元削減を行う手法

Method of dimension reduction with which new composite variables, principal components, are created by linear combination of multiple variables

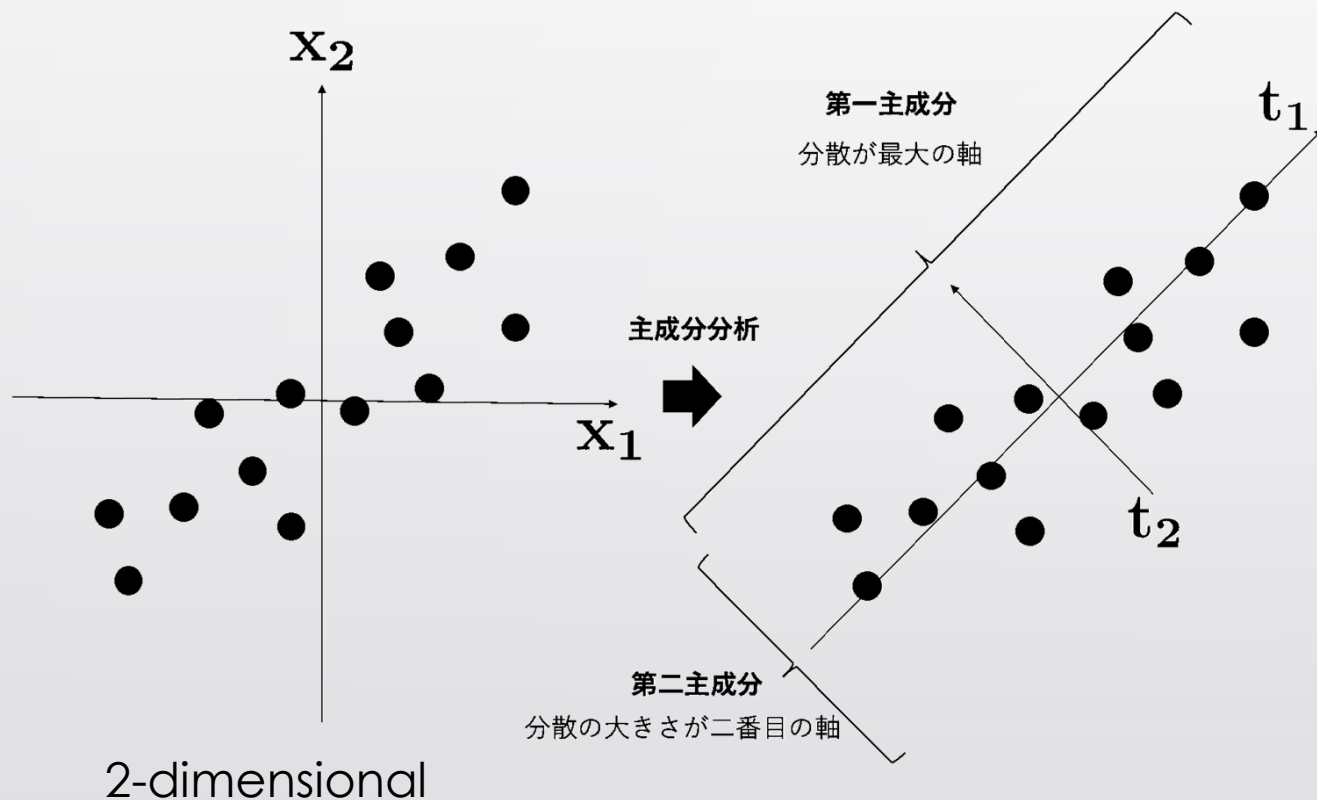


3次元データが、2つの主成分で構成される2次元空間内に表現されている

Three dimensional data are projected onto two-dimensional space defined by two principal components

http://www.nlpc.org/pca_principal_component_analysis.html

主成分 Principal Components



第1主成分軸は、データの分散が最大化される方向を向いている

The first PC axis is oriented in the direction along which variance of projected data is maximized

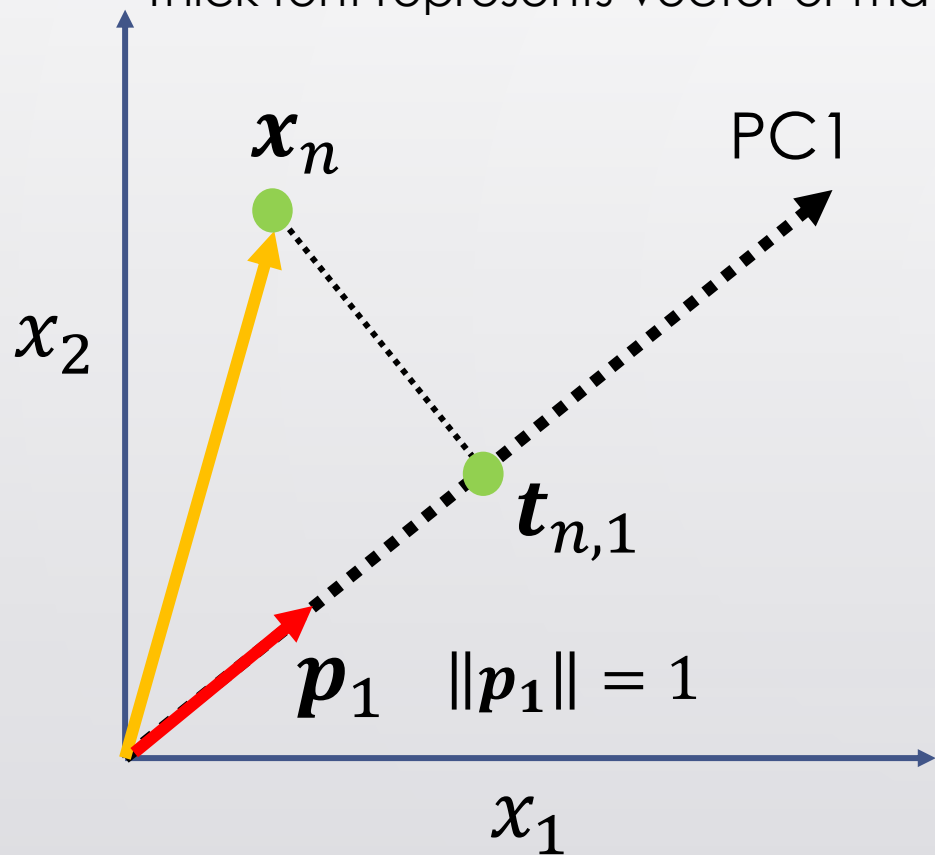
第 j 主成分軸は、データの分散が j 番目の大きさになる方向を向いている

The j -th PC axis is oriented in the direction along which projected data has j -th largest variance

第1主成分の計算 Computation of PC1

太字はベクトルか行列

Thick font represents vector or matrix



変数を中心化しておく Center the variables

観測データ x_n の第1主成分軸方向への射影 $t_{n,1}$ を計算する

Compute the projection $t_{n,1}$ of the observed data x_n onto the first PC axis

$t_{n,1}$ は x_n と p_1 の内積

$t_{n,1}$ is dot(inner) product of x_n and p_1

第1主成分の計算 Computation of PC1

$t_{n,1}$ は x_n と p_1 の内積 $t_{n,1}$ is dot product of x_n and p_1

$x_n = [x_{n,1} \ x_{n,2} \ \dots \ x_{n,M}]$ データは M 次元で N 個の観測値（データ）がある
 $p_1 = [p_{1,1} \ p_{1,2} \ \dots \ p_{1,M}]$ Data is M -dimensional and there are in total of N observations (Data points)

$$t_1 = \begin{bmatrix} t_{1,1} \\ t_{2,1} \\ \vdots \\ t_{N,1} \end{bmatrix} = \begin{bmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,M} \\ x_{2,1} & x_{2,2} & \dots & x_{2,M} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N,1} & x_{N,2} & \dots & x_{N,M} \end{bmatrix} \begin{bmatrix} p_{1,1} \\ p_{2,1} \\ \vdots \\ p_{M,1} \end{bmatrix} = Xp_1$$

第1主成分の計算 Computation of PC1

t_1 の分散 $s_{t_1}^2$ を計算する Compute variance $s_{t_1}^2$ of t_1

$$\text{Variance} = \frac{1}{n} (\mathbf{X} - \bar{\mathbf{X}})^T (\mathbf{X} - \bar{\mathbf{X}})$$

$$s_{t_1}^2 = \frac{1}{N} (\mathbf{t}_1 - \text{mean}(\mathbf{t}_1))^T (\mathbf{t}_1 - \text{mean}(\mathbf{t}_1)) = \frac{1}{N} \mathbf{t}_1^T \mathbf{t}_1 = \frac{1}{N} (\mathbf{X} \mathbf{p}_1)^T (\mathbf{X} \mathbf{p}_1) = \frac{1}{N} \mathbf{p}_1^T \mathbf{X}^T \mathbf{X} \mathbf{p}_1$$

変数は中心化されているので $\text{mean}(t_1) = 0$
 $\text{mean}(t_1) = 0$ since variables are centered

$$\mathbf{t}_1 = \mathbf{X} \mathbf{p}_1$$

$$(\mathbf{AB})^T = \mathbf{B}^T \mathbf{A}^T$$

第1主成分の計算 Computation of PC1

t_1 の分散 $s_{t_1}^2$ を計算する Compute variance $s_{t_1}^2$ of t_1

$$s_{t_1}^2 = \frac{1}{N} \mathbf{p}_1^T \mathbf{X}^T \mathbf{X} \mathbf{p}_1 = \mathbf{p}_1^T \mathbf{V} \mathbf{p}_1 \quad \mathbf{V} = \frac{1}{N} \mathbf{X}^T \mathbf{X}$$

\mathbf{p}_1 が第1主成分軸の方向を向いている時 $s_{t_1}^2$ が最大化されるので...

When \mathbf{p}_1 is oriented in the direction of PC1, $s_{t_1}^2$ is maximized, so...

$$\mathbf{V} \mathbf{p}_1 = \lambda \mathbf{p}_1$$

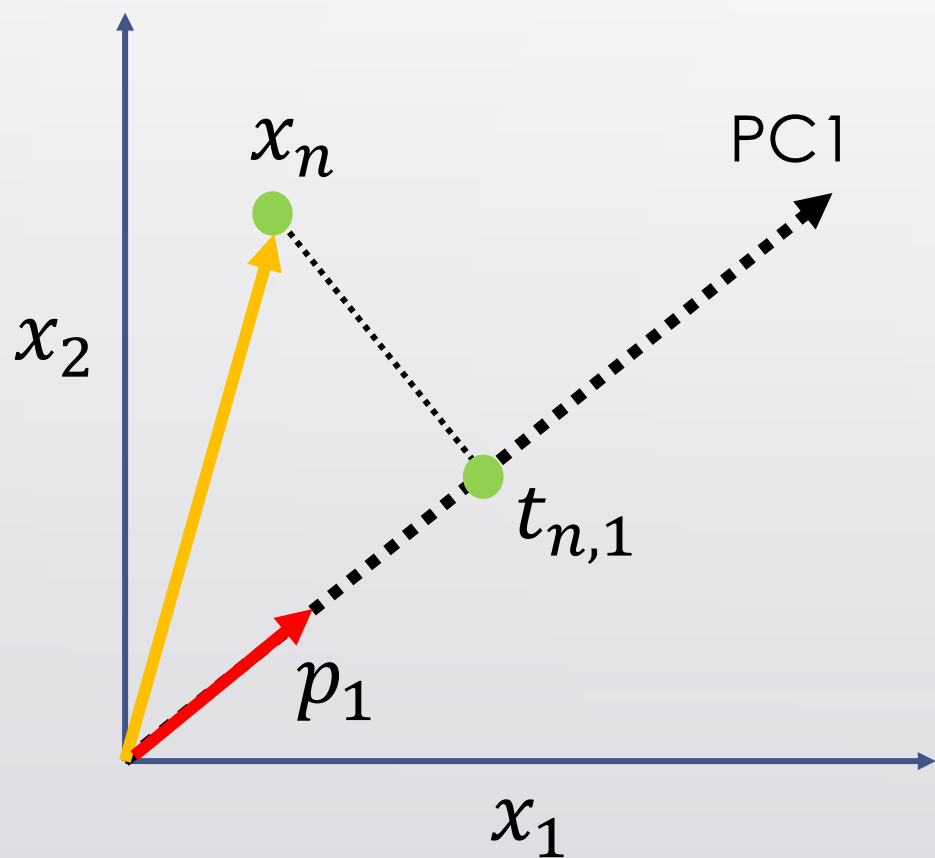
\mathbf{p}_1 は \mathbf{V} の固有ベクトルである

\mathbf{p}_1 is eigenvector of \mathbf{V}

未定乗数法

Lagrange method of multiplier

第1主成分の計算 Computation of PC1

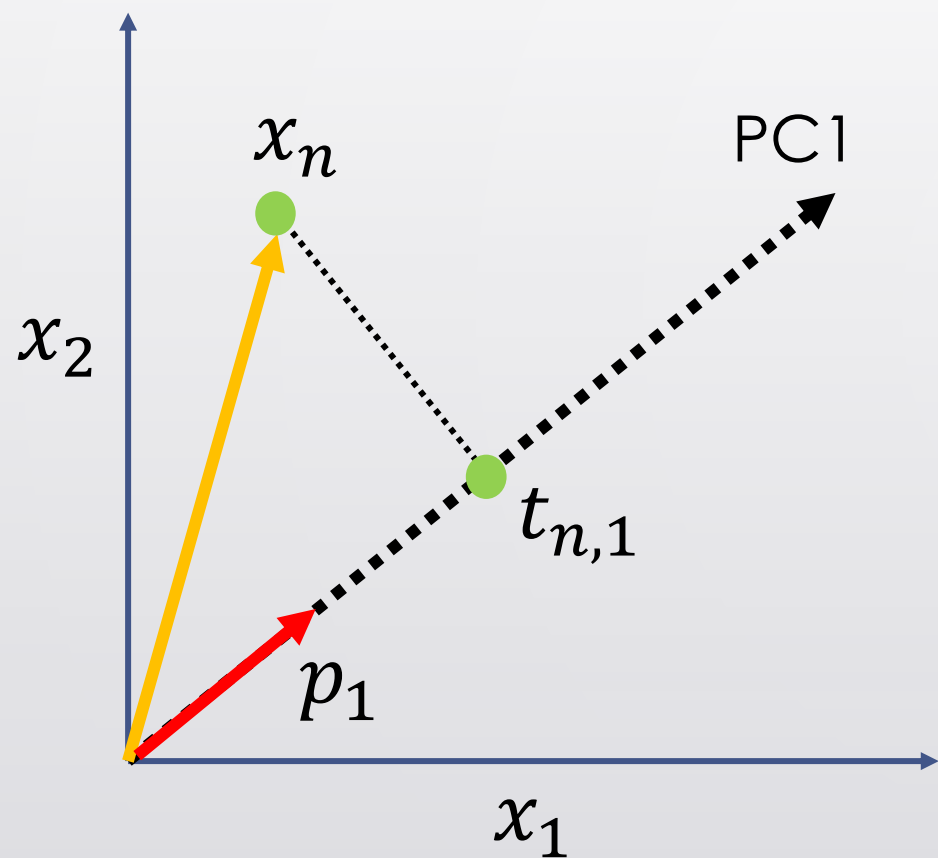


$$Vp_1 = \lambda p_1 \quad V = \frac{1}{N} X^T X$$

p_1 は V の固有ベクトルである p_1 is eigenvector of V

では、固有値 λ は何を表しているのだろうか？
Then, what does eigen value λ represent?

第1主成分の計算 Computation of PC1



では、固有値 λ は何を表しているのだろうか？
Then, what does eigen value λ represent?

$$s_{t_1}^2 = \mathbf{p}_1^T \mathbf{V} \mathbf{p}_1 = \lambda \mathbf{p}_1^T \mathbf{p}_1 = \lambda \|\mathbf{p}_1\|^2 = \lambda$$

$$\mathbf{V} \mathbf{p}_1 = \lambda \mathbf{p}_1$$

$$\|\mathbf{p}_1\| = 1$$

t_1 の分散は λ に一致する Variance of t_1 equals to λ

第1主成分の計算 Computation of PC1

$\mathbf{V}\mathbf{p}_1 = \lambda\mathbf{p}_1$ \mathbf{p}_1 は \mathbf{V} の固有ベクトルである \mathbf{p}_1 is eigenvector of \mathbf{V}

$\mathbf{V} = \frac{1}{N} \mathbf{X}^T \mathbf{X}$ \mathbf{t}_1 の分散は λ に一致する Variance of \mathbf{t}_1 equals to λ

\mathbf{V} とは何か? What is \mathbf{V} ?

$$\mathbf{V} = \frac{1}{N} \mathbf{X}^T \mathbf{X} \quad \mathbf{X} = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,M} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,M} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N,1} & x_{N,2} & \cdots & x_{N,M} \end{bmatrix}$$

分散共分散行列 Variance-Covariance Matrix

\mathbf{V} は \mathbf{X} の分散共分散行列である \mathbf{V} is variance-covariance matrix of \mathbf{X}

$$\mathbf{V} = \frac{1}{N} \mathbf{X}^T \mathbf{X} = \frac{1}{N} \begin{bmatrix} x_{1,1} & x_{2,1} & \dots & x_{N,1} \\ x_{1,2} & x_{2,2} & \dots & x_{N,2} \\ \vdots & \vdots & \ddots & \vdots \\ x_{1,M} & x_{2,M} & \dots & x_{N,M} \end{bmatrix} \begin{bmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,M} \\ x_{2,1} & x_{2,2} & \dots & x_{2,M} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N,1} & x_{N,2} & \dots & x_{N,M} \end{bmatrix}$$

$$= \begin{bmatrix} \sigma_{1,1}^2 & \sigma_{1,2}^2 & \dots & \sigma_{1,M}^2 \\ \sigma_{2,1}^2 & & \ddots & \vdots \\ \vdots & & & \sigma_{M-1,M}^2 \\ \sigma_{M,1}^2 & \sigma_{M,2}^2 & \dots & \sigma_{M,M}^2 \end{bmatrix} \quad \sigma_{i,j}^2 = \frac{1}{N} \sum_{k=1}^N x_{k,i} x_{k,j}$$

主成分分析の手順 Procedure of PCA

1. データ \mathbf{X} の分散共分散行列 \mathbf{V} を計算する

Compute variance-covariance matrix \mathbf{V} of data \mathbf{X}

2. 分散共分散行列 \mathbf{V} を固有値分解する

Eigenvalue decomposition of matrix \mathbf{V}

3. 第 k 主成分の分散は, k 番目に大きな固有値 λ_k

Variance of k -th principal component is k -th largest eigenvalue λ_k of matrix \mathbf{V}

4. 第 k 主成分は, 固有値 λ_k に対応する固有ベクトル \mathbf{p}_k

k -th principal component is eigenvector \mathbf{p}_k that corresponds to eigenvalue λ_k

第k主成分 k-th Principal Component

$$\mathbf{t}_k = \begin{bmatrix} t_{1,k} \\ t_{2,k} \\ \vdots \\ t_{N,k} \end{bmatrix} = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,M} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,M} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N,1} & x_{N,2} & \cdots & x_{N,M} \end{bmatrix} \begin{bmatrix} p_{1,k} \\ p_{2,k} \\ \vdots \\ p_{M,k} \end{bmatrix} = \mathbf{X} \mathbf{p}_k$$

主成分同士は直交している Principal components are orthogonal

$$\mathbf{p}_i^T \mathbf{p}_j = \begin{cases} 1(i = j) \\ 0(i \neq j) \end{cases}$$

第k主成分 k-th Principal Component

$$\mathbf{t}_k = \begin{bmatrix} t_{1,k} \\ t_{2,k} \\ \vdots \\ t_{N,k} \end{bmatrix} = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,M} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,M} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N,1} & x_{N,2} & \cdots & x_{N,M} \end{bmatrix} \begin{bmatrix} p_{1,k} \\ p_{2,k} \\ \vdots \\ p_{M,k} \end{bmatrix} = \mathbf{X} \mathbf{p}_k$$

主成分得点
Factor Score

主成分負荷量
Factor Loading

元データと主成分負荷量の内積で、主成分得点が得られる
Factor score is computed as dot product of observation and factor loading

第k主成分 k-th Principal Component

$$\mathbf{T} = [\mathbf{t}_1 \ \mathbf{t}_2 \ \dots \ \mathbf{t}_M] = \begin{bmatrix} t_{1,1} & t_{1,2} & \dots & t_{1,M} \\ t_{2,1} & t_{2,2} & \dots & t_{2,M} \\ \vdots & \vdots & \ddots & \vdots \\ t_{N,1} & t_{N,2} & \dots & t_{N,M} \end{bmatrix} =$$
$$\begin{bmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,M} \\ x_{2,1} & x_{2,2} & \dots & x_{2,M} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N,1} & x_{N,2} & \dots & x_{N,M} \end{bmatrix} \begin{bmatrix} p_{1,1} & p_{1,2} & \dots & p_{1,M} \\ p_{2,1} & p_{2,2} & \dots & p_{2,M} \\ \vdots & \vdots & \ddots & \vdots \\ p_{M,1} & p_{M,2} & \dots & p_{M,M} \end{bmatrix} = \mathbf{X}[\mathbf{p}_1 \ \mathbf{p}_2 \ \dots \ \mathbf{p}_M] = \mathbf{X}\mathbf{P}$$

固有値が大きい順に固有ベクトルを並べた

Eigenvectors are ordered in a descending order of eigenvalue

次元削減 Dimension Reduction

$$\underset{(N, \mathbf{M})}{\mathbf{T}} = [\mathbf{t}_1 \ \mathbf{t}_2 \ \dots \ \mathbf{t}_M] = \begin{bmatrix} t_{1,1} & t_{1,2} & \dots & t_{1,M} \\ t_{2,1} & t_{2,2} & \dots & t_{2,M} \\ \vdots & \vdots & \ddots & \vdots \\ t_{N,1} & t_{N,2} & \dots & t_{N,M} \end{bmatrix} = \begin{array}{l} \text{第} H+1 \sim M \text{主成分を削除する} \\ \text{Deleting } H+1\text{-th to } M\text{-th PCs} \end{array}$$

$$\underset{(N, M)}{\begin{bmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,M} \\ x_{2,1} & x_{2,2} & \dots & x_{2,M} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N,1} & x_{N,2} & \dots & x_{N,M} \end{bmatrix}} \underset{(M, \mathbf{M})}{\begin{bmatrix} p_{1,1} & p_{1,2} & \dots & p_{1,H} & \cancel{p_{1,H+1}} & \dots & \cancel{p_{1,M}} \\ p_{2,1} & p_{2,2} & \dots & p_{2,H} & \cancel{p_{2,H+1}} & \dots & \cancel{p_{2,M}} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ p_{M,1} & p_{M,2} & \dots & p_{M,H} & \cancel{p_{M,H+1}} & \dots & \cancel{p_{M,M}} \end{bmatrix}}$$

次元削減 Dimension Reduction

$$\begin{matrix} T' = [t_1 & t_2 & \dots & t_H] = \\ (N, H) \end{matrix} \begin{matrix} \begin{bmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,M} \\ x_{2,1} & x_{2,2} & \dots & x_{2,M} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N,1} & x_{N,2} & \dots & x_{N,M} \end{bmatrix} \\ (N, M) \end{matrix} \begin{matrix} \begin{bmatrix} p_{1,1} & p_{1,2} & \dots & p_{1,H} \\ p_{2,1} & p_{2,2} & \dots & p_{2,H} \\ \vdots & \vdots & \ddots & \vdots \\ p_{M,1} & p_{M,2} & \dots & p_{M,H} \end{bmatrix} \\ (M, H) \end{matrix} = X[p_1 \ p_2 \ \dots \ p_H]$$

情報量（分散）が小さい主成分を除くことで、もとのデータが持つ情報を、低い次元で表現できる

Information contained in the original data can be represented in lower number of dimensions by deleting PCs with small amount of information (variance)

次元削減 Dimension Reduction

何番目の主成分まで残すべきか？

How many principal components should we retain in dimension reduction?

3. 第k主成分の分散は, k 番目に大きな固有値 λ_k

第1~H主成分までの情報量の合計は

Sum of amount of information of the first H principal components is

$$\sum_1^H \lambda_k$$

累積寄与率 Cumulative Contribution Ratio

第1~H主成分までの情報量の合計は

Sum of amount of information of the first H principal components is $\sum_1^H \lambda_k$

データ全体の情報量の合計は

Sum of amount of information of the data set is $\sum_1^M \lambda_k$

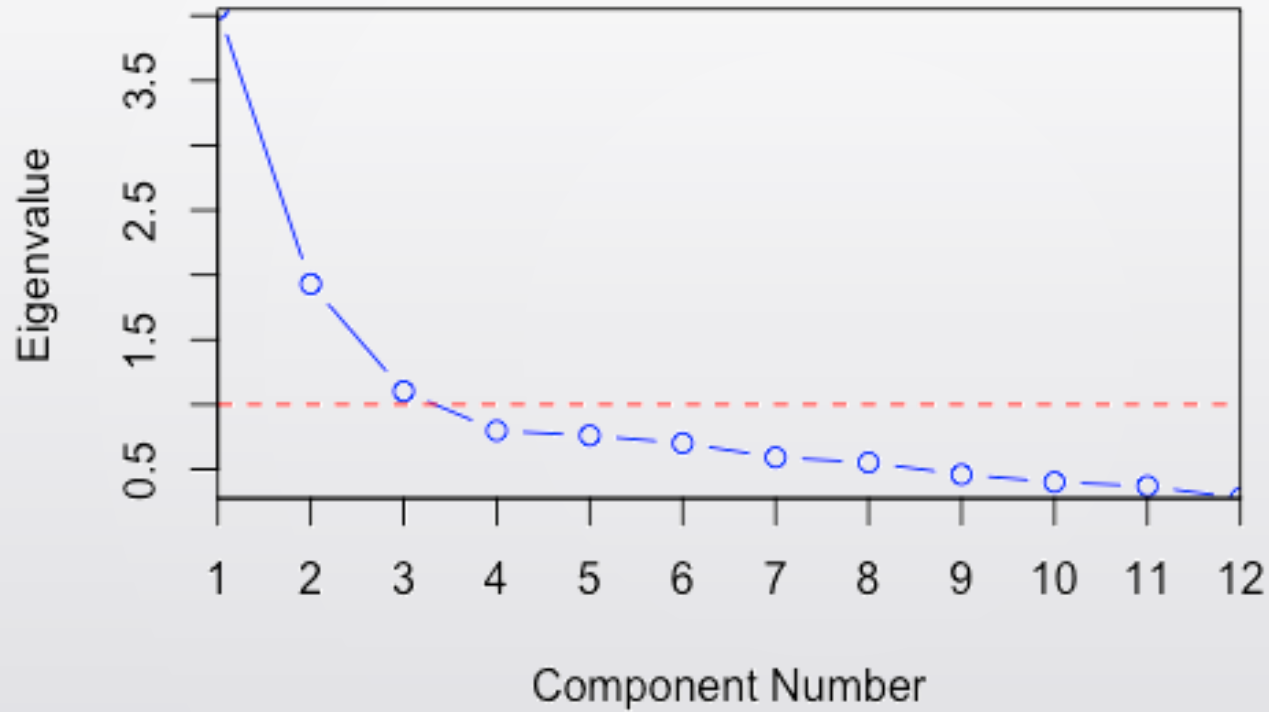
$$\frac{\sum_1^H \lambda_k}{\sum_1^M \lambda_k} \times 100 \geq 80$$

これ以外にも主成分の数を決める基準は色々ある

There are many other customary criteria for determining the number of principal components to be retained

スクリープロット

Scree Plot



https://en.wikipedia.org/wiki/Scree_plot

スクリープロットの肩の位置で、主成分の数を決めることがある

Number of retained PCs is sometimes determined by the location of “shoulder” in scree plot.