

# rd-2. ヒストグラム, 散布 図, 折れ線グラフ, 要約 統計量

データサイエンス演習  
(R システムを使用)

<https://www.kkaneko.jp/de/rd/index.html>

金子邦彦



# 2-1 パッケージの追加インストール

# パッケージの設定 (1/2)

- 次の手順で, 必要なパッケージをインストール
- パッケージをインストールするのにインターネット接続が必要
- `install.packages("ggplot2")` を実行

```
> install.packages("ggplot2")  
Installing package into 'D:/Users/user/Doc  
(as 'lib' is unspecified)  
trying URL 'https://mran.revolutionanalyti  
/contrib/3.2/ggplot2_2.0.0.zip'
```

- `install.packages("dplyr")` を実行

```
> install.packages("dplyr")  
Installing package into 'D:/Users/user/Do  
(as 'lib' is unspecified)  
trying URL 'https://mran.revolutionanalyt  
ws/contrib/3.2/dplyr_0.4.3.zip'
```

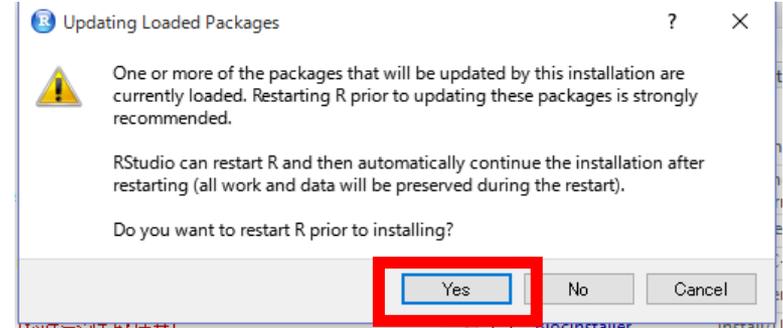
# パッケージの設定 (2/2)



- `install.packages("tidyr")`

を実行

```
> install.packages("tidyr")
Installing package into 'D:/Users/user
(as 'lib' is unspecified)
trying URL 'https://mran.revolutionana
ws/contrib/3.2/tidyr_0.3.1.zip'
Content type 'application/zip' length 1050
downloaded 1050 kb
```



- `install.packages("magrittr")` を実行

```
> install.packages("magrittr")
Error in install.packages : updating loade
Restarting R session...
Microsoft R Open 3.2.3
Default CRAN mirror snapshot taken on 2016
The enhanced R distribution from Microsoft
```

こんな表示が  
でたら **Yes**

- `install.packages("KernSmooth")` を実行

```
> install.packages("KernSmooth")
Installing package into 'D:/Users/user/Doc
(as 'lib' is unspecified)
trying URL 'https://mran.revolutionanalyti
ws/contrib/3.2/KernSmooth_2.23-15.zip'
Content type 'application/zip' length 1050
downloaded 1050 kb
```

※ 「K」と「S」が大文字

# 2-2 R オブジェクトのコンストラクタ

# コンストラクタの例



年次	出生数	死亡数
1985	1432	752
1990	1222	820
1995	1187	922
2000	1191	962
2005	1063	1084
2010	1071	1197

テーブルの例

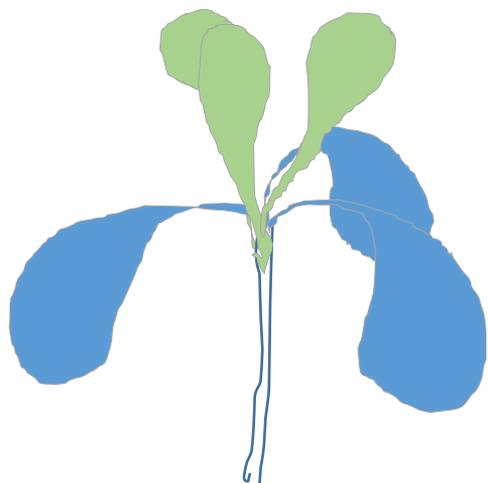
```
x1 <- data.frame( 年次=c(1985, 1990, 1995, 2000, 2005, 2010),  
  出生数=c(1432, 1222, 1187, 1191, 1063, 1071),  
  死亡数=c(752, 820, 922, 962, 1084, 1197) )
```

上記のテーブルを生成するコンストラクタ

```
> x1 <- data.frame( 年次=c(1985, 1990, 1995, 2000, 2005, 2010),  
+                 出生数=c(1432, 1222, 1187, 1191, 1063, 1071),  
+                 死亡数=c(752, 820, 922, 962, 1084, 1197) )  
> |
```

## 2-3 iris データセット

# アヤメ属 (Iris)



- 多年草
- 世界に 150種. 日本に 9種.
- 花被片は 6個
- 外花被片 (がいかひへん) Sepal  
3個 (大型で下に垂れる)
- 内花被片 (ないかひへん) Petal  
3個 (直立する)

# Iris データセット



```
> iris
  Sepal.Length Sepal.Width Petal.Length Petal.Width Species
1           5.1           3.5           1.4           0.2   setosa
2           4.9           3.0           1.4           0.2   setosa
3           4.7           3.2           1.3           0.2   setosa
4           4.6           3.1           1.5           0.2   setosa
5           5.0           3.6           1.4           0.2   setosa
6           5.4           3.9           1.7           0.4   setosa
7           4.6           3.4           1.4           0.3   setosa
8           5.0           3.4           1.5           0.2   setosa
9           4.4           2.9           1.4           0.2   setosa
10          4.9           3.1           1.5           0.1   setosa
11          5.4           3.7           1.5           0.2   setosa
12          4.8           3.4           1.6           0.2   setosa
```

Iris データセットは、  
Rシステムの中に組み込み済み

- 3種のアヤメの外花被辺、内花被片の幅と長さを計測したデータセット

Iris setosa

Iris versicolor

Iris virginica

- データ数は  $50 \times 3$
- 作成者：Ronald Fisher
- 作成年：1936

## 2-4 ヒストグラム<sup>1</sup>の例

# iris の 4属性それぞれのヒストグラム

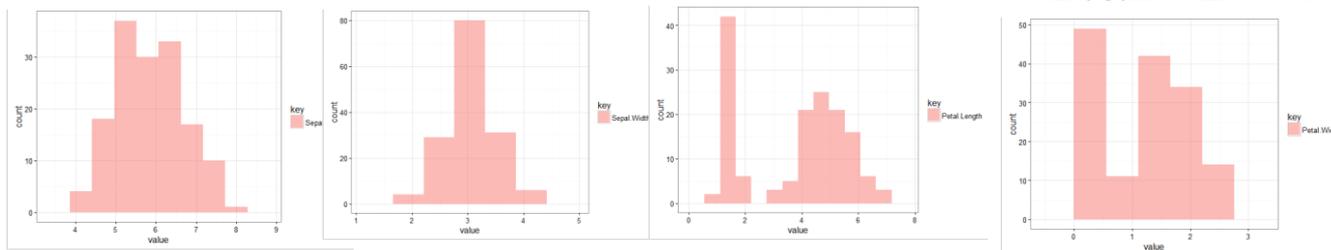


属性： Sepal.Length, Sepal.Width, Petal.Length, Petal.Width

```
> iris
```

	Sepal.Length	Sepal.width	Petal.Length	Petal.width	Species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
5	5.0	3.6	1.4	0.2	setosa
6	5.4	3.9	1.7	0.4	setosa
7	4.6	3.4	1.4	0.3	setosa
8	5.0	3.4	1.5	0.2	setosa
9	4.4	2.9	1.4	0.2	setosa
10	4.9	3.1	1.5	0.1	setosa
11	5.4	3.7	1.5	0.2	setosa
12	4.8	3.4	1.6	0.2	setosa

各属性のヒストグラム



# 複数ヒストグラムの重ね合わせ表示



```
library(dplyr)
```

```
d2 <- tbl_df( iris )
```

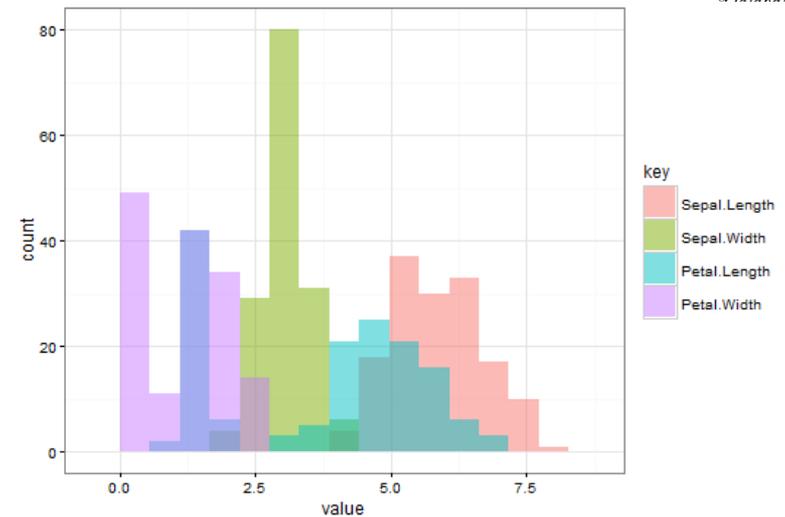
```
library(tidyr)
```

```
library(magrittr)
```

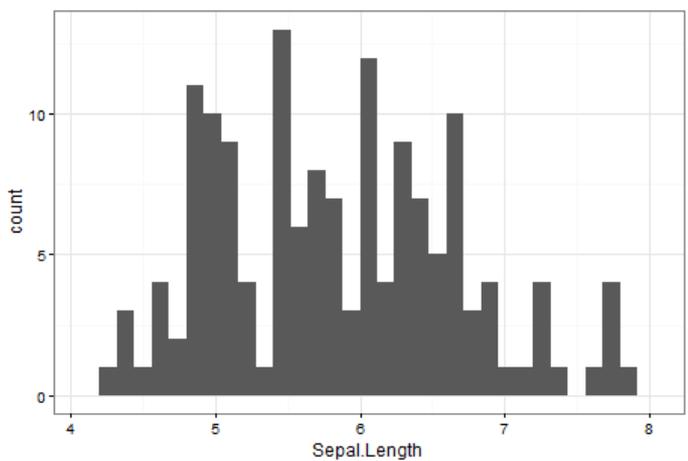
```
library(KernSmooth)
```

```
library(ggplot2)
```

```
d2 %>% select( Sepal.Length, Sepal.Width, Petal.Length,  
Petal.Width ) %>% gather() %>% ggplot( aes(x=value, fill=key) ) +  
  geom_histogram( binwidth=dpih( use_series(d2, Sepal.Length) ),  
alpha=0.5, position="identity" ) +  
  theme_bw()
```

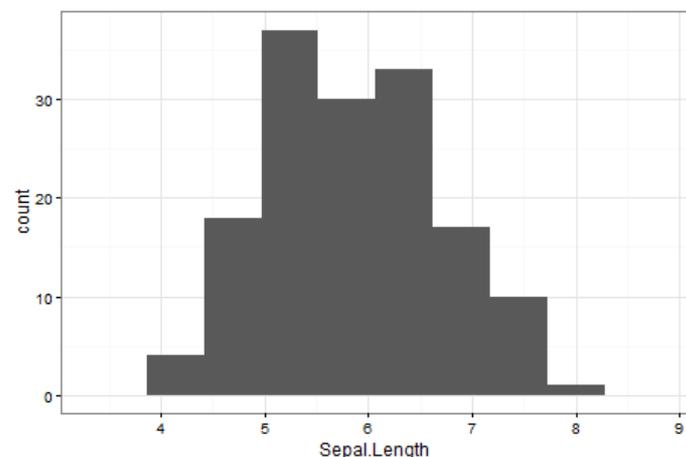


# ヒストグラムでの区間幅の調整



区間幅 = 0.1

```
library(ggplot2)
ggplot(iris, aes(x = Sepal.Length)) +
  geom_histogram(binwidth=0.1) +
  theme_bw()
```



区間幅を、**dphi** 関数を用いて調整

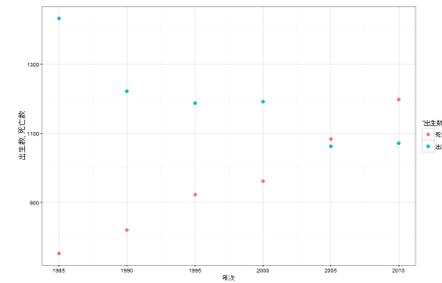
```
library(magrittr)
library(KernSmooth)
library(ggplot2)
ggplot(iris, aes(x = Sepal.Length)) +
  geom_histogram(
    binwidth=dphi( iris$Sepal.Length ) ) +
  theme_bw()
```

## 2-5 散布図, 折れ線グラフ

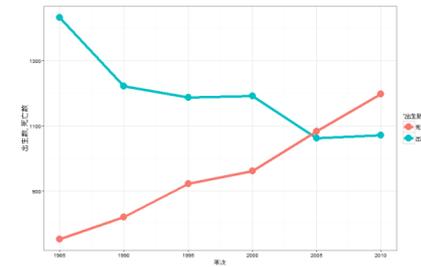
# 散布図、折れ線グラフのバリエーション



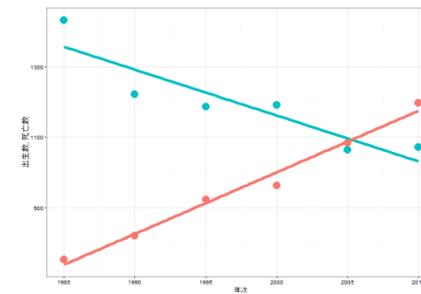
年次	出生数 (千人)	死亡数 (千人)
1985	1432	752
1990	1222	820
1995	1187	922
2000	1191	962
2005	1063	1084
2010	1071	1197



散布図



散布図  
+ 折れ線



散布図  
+ 線形近似

出生数、死亡数の推移

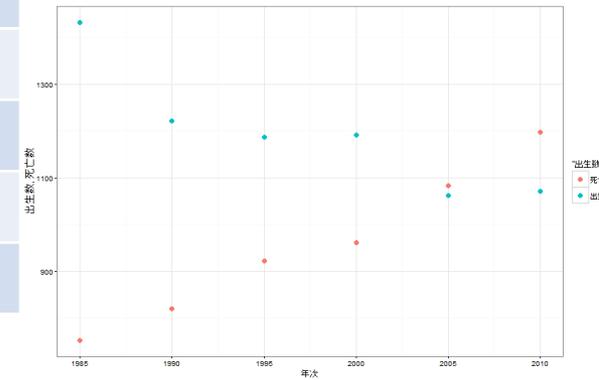
出典：総務省「第63回 日本統計年鑑 平成26年」

# 散布図



年次	出生数	死亡数
1985	1432	752
1990	1222	820
1995	1187	922
2000	1191	962
2005	1063	1084
2010	1071	1197

x 軸 (フィールド名)	年次
y 軸 (フィールド名)	出生数, 死亡数
点の大きさ (数値)	3
x 軸の名前 (文字列)	年次
y 軸の名前 (文字列)	出生数, 死亡数



```
x1 <- data.frame( 年次=c(1985, 1990, 1995, 2000, 2005, 2010),  
  出生数=c(1432, 1222, 1187, 1191, 1063, 1071),  
  死亡数=c(752, 820, 922, 962, 1084, 1197) )
```

```
library(ggplot2)
```

```
ggplot(x1, aes(x=年次)) +
```

```
  geom_point( aes(y=出生数, colour="出生数"), size=3 ) +
```

```
  geom_point( aes(y=死亡数, colour="死亡数"), size=3 ) +
```

```
  labs(x="年次", y="出生数, 死亡数") +
```

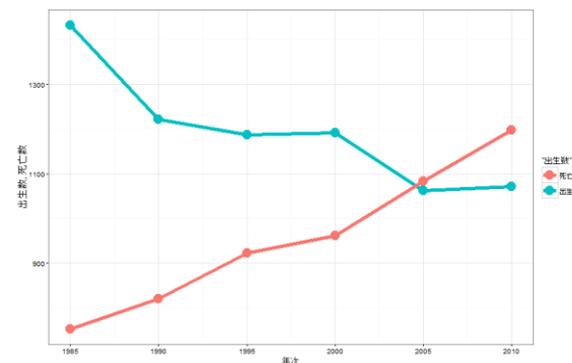
```
  theme_bw()
```

# 散布図 + 折れ線



年次	出生数	死亡数
1985	1432	752
1990	1222	820
1995	1187	922
2000	1191	962
2005	1063	1084
2010	1071	1197

x 軸 (フィールド名)	年次
y 軸 (フィールド名)	出生数, 死亡数
点の大きさ (数値)	3
x 軸の名前 (文字列)	年次
y 軸の名前 (文字列)	出生数, 死亡数



```
x1 <- data.frame( 年次=c(1985, 1990, 1995, 2000, 2005, 2010),  
  出生数=c(1432, 1222, 1187, 1191, 1063, 1071),  
  死亡数=c(752, 820, 922, 962, 1084, 1197) )
```

```
library(ggplot2)
```

```
ggplot(x1, aes(x=年次)) +
```

```
  geom_point( aes(y=出生数, colour="出生数"), size=6 ) +
```

```
  geom_point( aes(y=死亡数, colour="死亡数"), size=6 ) +
```

```
  geom_line( aes(y=出生数, colour="出生数"), size=2 ) +
```

```
  geom_line( aes(y=死亡数, colour="死亡数"), size=2 ) +
```

```
  labs(x="年次", y="出生数, 死亡数") +
```

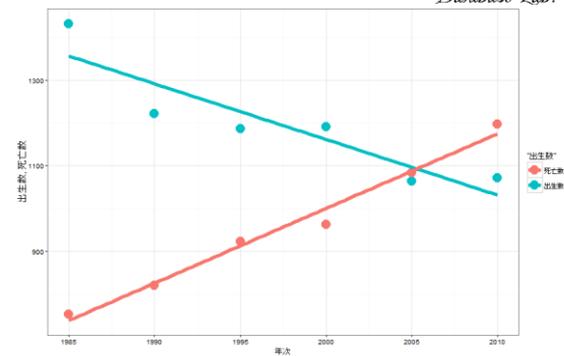
```
  theme_bw()
```

# 散布図 + 線形近似



年次	出生数	死亡数
1985	1432	752
1990	1222	820
1995	1187	922
2000	1191	962
2005	1063	1084
2010	1071	1197

x 軸 (フィールド名)	年次
y 軸 (フィールド名)	出生数, 死亡数
点の大きさ (数値)	3
x 軸の名前 (文字列)	年次
y 軸の名前 (文字列)	出生数, 死亡数



```
x1 <- data.frame( 年次=c(1985, 1990, 1995, 2000, 2005, 2010),  
                 出生数=c(1432, 1222, 1187, 1191, 1063, 1071),  
                 死亡数=c(752, 820, 922, 962, 1084, 1197) )
```

```
library(ggplot2)
```

```
ggplot(x1, aes(x=年次)) +
```

```
  geom_point( aes(y=出生数, colour="出生数"), size=6 ) +
```

```
  geom_point( aes(y=死亡数, colour="死亡数"), size=6 ) +
```

```
  stat_smooth( method="lm", se=FALSE, aes(y=出生数, colour="出生数"),  
              size=2 ) +
```

```
  stat_smooth( method="lm", se=FALSE, aes(y=死亡数, colour="死亡数"),  
              size=2 ) +
```

```
  labs(x="年次", y="出生数, 死亡数") +
```

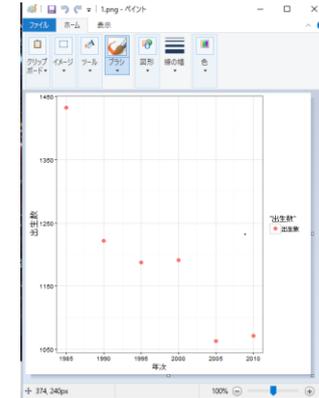
```
  theme_bw()
```

## 2-6 グラフのファイルへの保存

# png ファイルの作成



ファイル f:/1.png に保存



```
x1 <- data.frame( 年次=c(1985, 1990, 1995, 2000, 2005, 2010),  
                  出生数=c(1432, 1222, 1187, 1191, 1063, 1071),  
                  死亡数=c(752, 820, 922, 962, 1084, 1197) )
```

```
library(ggplot2)
```

```
png("f:/1.png")
```

```
ggplot(x1, aes(x=年次)) +
```

```
  geom_point( aes(y=出生数, colour="出生数"), size=3 ) +
```

```
  labs(x="年次", y="出生数") +
```

```
  theme_bw()
```

```
dev.off()
```

# 2-7 要約統計量, 頻度, ヒストグラム

# ここで行うこと

各フィールドの頻度 (数え上げ)

種類ごとの数え上げ

各フィールドの要約統計量の算出

平均 (mean)、標準偏差 (sd)、分散 (var)

中央値 (median)、四分位点 (quantile)、

最大値 (max)、最小値 (min)

科目	受講者	得点
国語	A	90
国語	B	80
算数	A	95
算数	B	90
理科	A	80

元データ



科目	受講者	得点
Length:5	Length:5	Min. :80
Class :character	Class :character	1st Qu.:80
Mode :character	Mode :character	Median :90
		Mean :87
		3rd Qu.:90
		Max. :95

要約統計量の例

# ここでの説明で使用するデータ



## 成績データ

科目	受講者	得点
国語	A	90
国語	B	80
算数	A	95
算数	B	90
理科	A	80

```
d1 <- data.frame(  
  科目=c("国語", "国語", "算数", "算数", "理科"),  
  受講者=c("A", "B", "A", "B", "A"),  
  得点=c(90, 80, 95, 90, 80) )
```

## コンストラクタ

```
> iris  
  Sepal.Length Sepal.Width Petal.Length Petal.Width Species  
1         5.1         3.5         1.4         0.2   setosa  
2         4.9         3.0         1.4         0.2   setosa  
3         4.7         3.2         1.3         0.2   setosa  
4         4.6         3.1         1.5         0.2   setosa  
5         5.0         3.6         1.4         0.2   setosa  
6         5.4         3.9         1.7         0.4   setosa  
7         4.6         3.4         1.4         0.3   setosa  
8         5.0         3.4         1.5         0.2   setosa  
9         4.4         2.9         1.4         0.2   setosa  
10        4.9         3.1         1.5         0.1   setosa  
11        5.4         3.7         1.5         0.2   setosa  
12        4.8         3.4         1.6         0.2   setosa
```

## iris データセット

※ **iris データセット**は  
R システムに組み込み済み

# 要約統計量 (summary を使用) ①



成績

科目	受講者	得点
国語	A	90
国語	B	80
算数	A	95
算数	B	90
理科	A	80



```
> summary(d1)
  科目  受講者      得点
国語:2  A:3   Min.   :80
算数:2  B:2   1st Qu.:80
理科:1                      Median :90
                                   Mean  :87
                                   3rd Qu.:90
                                   Max.  :95
```

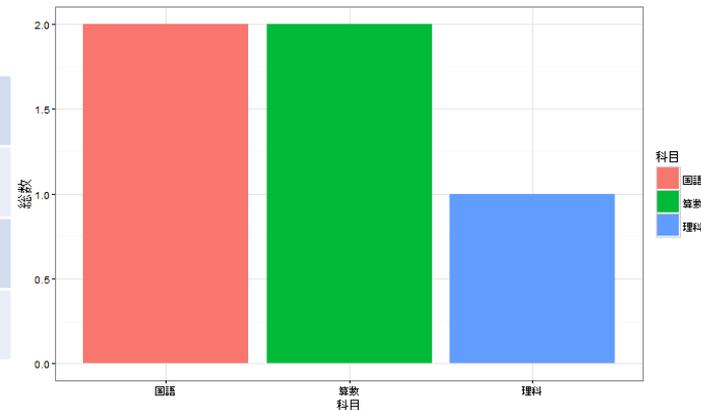
- ◆ 数値属性に対しては  
最小、最大、平均、  
中央値、四分位点

```
d1 <- data.frame(
  科目=c("国語", "国語", "算数", "算数", "理科"),
  受講者=c("A", "B", "A", "B", "A"),
  得点=c(90, 80, 95, 90, 80) )
summary(d1)
```

# 頻度のグラフ化 ①



科目	受講者	得点	集約を行うテーブルの変数名	d1
国語	A	90	集約したいフィールド名	科目
国語	B	80	x 軸の名前 (文字列)	科目
算数	A	95	y 軸の名前 (文字列)	総数
算数	B	90		
理科	A	80		



```
d1 <- data.frame(  
  科目=c("国語", "国語", "算数", "算数", "理科"),  
  受講者=c("A", "B", "A", "B", "A"),  
  得点=c(90, 80, 95, 90, 80) )
```

```
library(ggplot2)  
ggplot(d1, aes( x=科目, fill=科目 )) +  
  geom_bar(stat="count") +  
  labs(x="科目", y="総数") +  
  theme_bw()
```

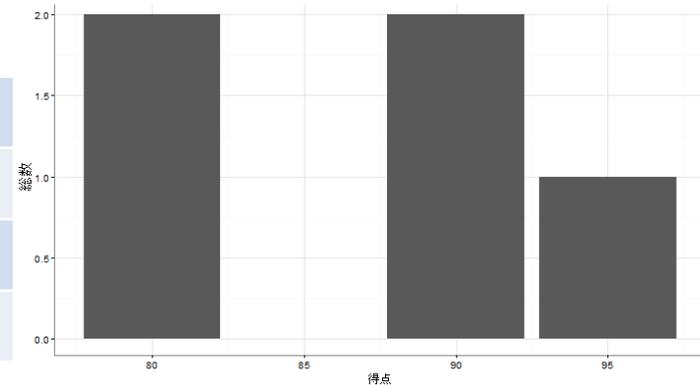
# 頻度のグラフ化 ②



Database Lab.

科目	受講者	得点
国語	A	90
国語	B	80
算数	A	95
算数	B	90
理科	A	80

集約を行うテーブルの変数名	d1
集約したいフィールド名	得点
x 軸の名前 (文字列)	得点
y 軸の名前 (文字列)	総数



```
d1 <- data.frame(  
  科目=c("国語", "国語", "算数", "算数", "理科"),  
  受講者=c("A", "B", "A", "B", "A"),  
  得点=c(90, 80, 95, 90, 80) )
```

```
library(ggplot2)  
ggplot(d1, aes( x=得点 )) +  
  geom_bar(stat="count") +  
  labs(x="得点", y="総数") +  
  theme_bw()
```

# 要約統計量 (summary を使用) ②



```
> iris
  Sepal.Length Sepal.Width Petal.Length Petal.Width Species
1           5.1           3.5           1.4           0.2  setosa
2           4.9           3.0           1.4           0.2  setosa
3           4.7           3.2           1.3           0.2  setosa
4           4.6           3.1           1.5           0.2  setosa
5           5.0           3.6           1.4           0.2  setosa
6           5.4           3.9           1.7           0.4  setosa
7           4.6           3.4           1.4           0.3  setosa
8           5.0           3.4           1.5           0.2  setosa
9           4.4           2.9           1.4           0.2  setosa
10          4.9           3.1           1.5           0.1  setosa
11          5.4           3.7           1.5           0.2  setosa
12          4.8           3.4           1.6           0.2  setosa
```



```
 Sepal.Length      Sepal.Width      Petal.Length
Min.      :4.300      Min.      :2.000      Min.      :1.000
1st Qu.   :5.100      1st Qu.   :2.800      1st Qu.   :1.600
Median    :5.800      Median    :3.000      Median    :4.350
Mean      :5.843      Mean      :3.057      Mean      :3.758
3rd Qu.   :6.400      3rd Qu.   :3.300      3rd Qu.   :5.100
Max.      :7.900      Max.      :4.400      Max.      :6.900
 Petal.Width      Species
Min.      :0.100      setosa      :50
1st Qu.   :0.300      versicolor:50
Median    :1.300      virginica  :50
Mean      :1.199
3rd Qu.   :1.800
Max.      :2.500
```

iris データセット

summary(iris)