st-1. 統計処理の概要

（統計処理演習，スライド）

https://www.kkaneko.jp/cc/st/index.html

金子邦彦

アウトライン

- 1 記述統計量
- 2 ヒストグラム
- 3 クロス集計表
- 4 検定

コマンドを使う理由

- 記録が残り，再現が容易

- なぜプログラミング?

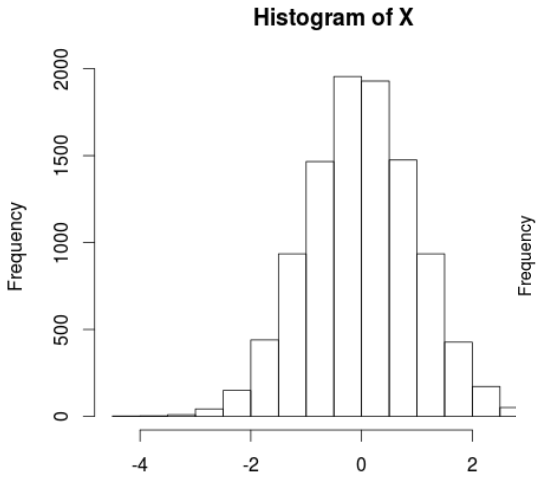
統計処理手順のコマンドをファイルに記録．再実行や確認が容易．※ コマンドは難しくは無い

統計処理の例

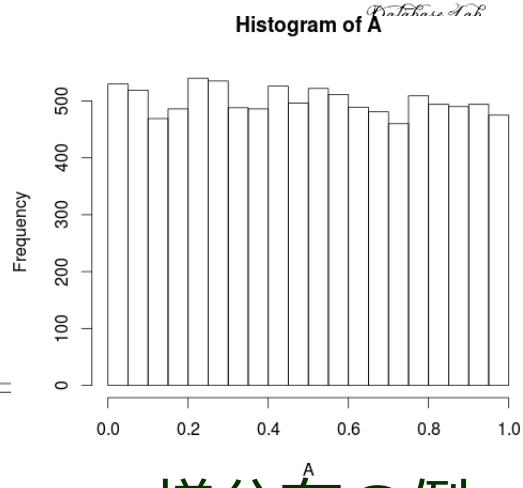
		SPSS 言語	R システム
記述統計量	describe, skew, kurt	descriptive	summary, sd, skewness, kurtosis
頻度表	hist	frequencies	table
クロス集計表	crosstab	crosstabs	table
集約	aggregate	aggregate	aggregate, table, as.data.frame, list
Welch の t 検定	ttest_ind (equal_var=False)	t-test	t.test
one-way ANOVA 検定	f_oneway	oneway	oneway.test
Wilcox rank sum 検定	mannwhitneyu (use_continuity=True)	npar tests	wilcox.test (exact=F or T, correct=T)
Shapiro-Wilk 検定	shapiro,	examine	shapiro.test,
Kolmogorov-Smirnov 検定	kstest		ks.test

記述統計量の例

- 平均
- 中央値
- 四分位点
- 最小値・最大値
- 分散（分布の広がり）
- 歪度（左右非対称の度合い）
- 尖度（分布の峰の鋭さ）



正規分布の例
（標本数: 10000）



一様分布の例
（標本数: 10000）

```
> summary(X)
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
-4.076000 -0.659100 -0.000611  0.003233  0.682600  3.405000
> sd(X)
[1] 0.9958542
> skewness(X)
[1] -0.001900942
> kurtosis(X)
[1] -0.003416198
```

Rプログラムの例

平均	0.00323	0.492
中央値	-0.000611	0.495
第一・四分位点	-0.659	0.246
第三・四分位点	0.683	0.745
最小値	-4.08	0.000175
最大値	3.41	1.000
分散	0.996	0.288
歪度	-0.001901	0.01991
尖度	**-0.00342**	**-1.193**

頻度表の例

c	b	b	a	c	c	c	b	a	c

■ 元データの例

a	b	c
2	3	5

■ 頻度表の例

```
> R
 [1] "c" "b" "b" "a" "c" "c" "c" "b" "a" "c"
> table(R)
R
a b c
2 3 5
```

元データの確認

「table(R)」で頻度表を作成

Rプログラムの例

クロス集計表の例

c	b	b	a	c	c	c	b	a	c
e	d	e	e	d	d	e	e	d	e

■ 元データの例

	d	e
a	1	1
b	1	2
c	2	3

■ 頻度表の例

```
> R
 [1] "c" "b" "b" "a" "c" "c" "c" "b" "a" "c"
> S
 [1] "e" "d" "e" "e" "d" "d" "e" "e" "d" "e"
> table(R, S)
   S
R   d e
  a 1 1
  b 1 2
  c 2 3
```

元データの確認

「table(R, S)」で頻度表を作成

Rプログラムの例

集計の例

■ 元データの例

NAME	PRODUCT	NUM
A	apple	10
A	orange	20
B	apple	5
B	orange	6
B	banana	50

■ 頻度表の例
(件数のカウント)

A	2
B	3

apple	2
banana	1
orange	2

```
> as.data.frame(table(T$NAME))
  Var1 Freq
1    A    2
2    B    3
> as.data.frame(table(T$PRODUCT))
    Var1 Freq
1  apple    2
2 banana    1
3 orange    2
```

Rプログラムの例

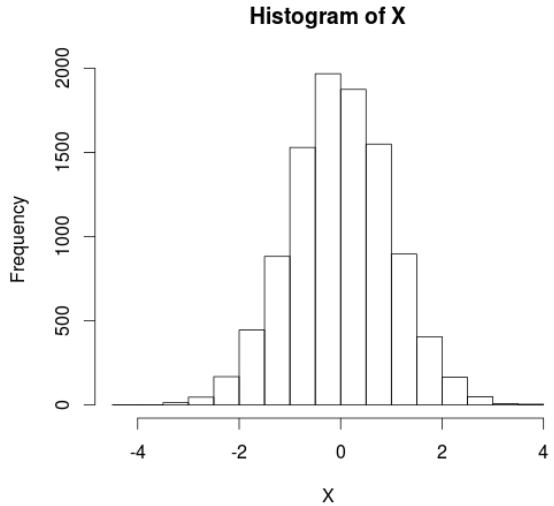
■ 頻度表の例
(合計値)

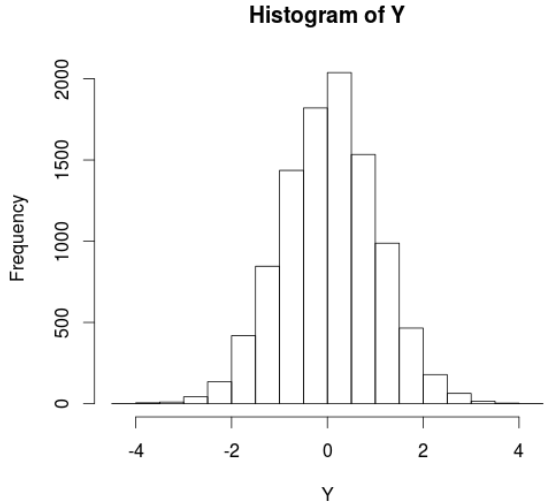
A	30
B	61

```
> aggregate(T$NUM, list(T$NAME), sum)
  Group.1  x
1       A 30
2       B 61
> aggregate(T$NUM, list(T$NAME), summary)
  Group.1 x.Min. x.1st Qu. x.Median x.Mean x.3rd Qu. x.Max.
1       A  10.00     12.50    15.00  15.00     17.50  20.00
2       B   5.00      5.50     6.00  20.33     28.00  50.00
```

Rプログラムの例

t 検定の例 (1/2)





■ 分布 X （標本数: 10000）　■ 分布 Y （標本数: 10000）

※ 平均0, 標準偏差1 になるように合成　※ 平均0.05, 標準偏差1 になるように合成

帰無仮説：「分布X, Y の母集団の平均が等しい」

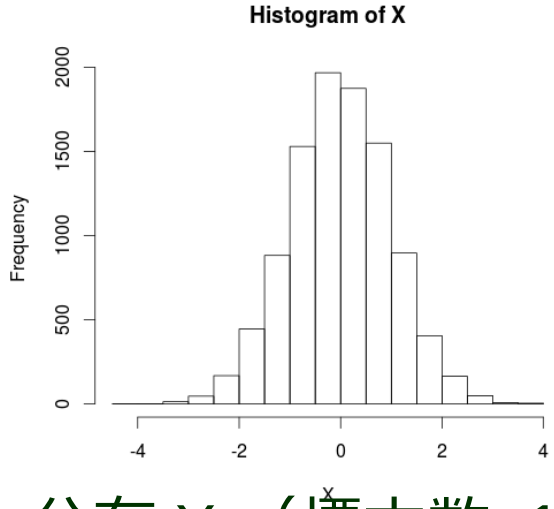
⇒ t 検定により，「帰無仮説が成り立つ確率は 0.00048 % (=p値)」

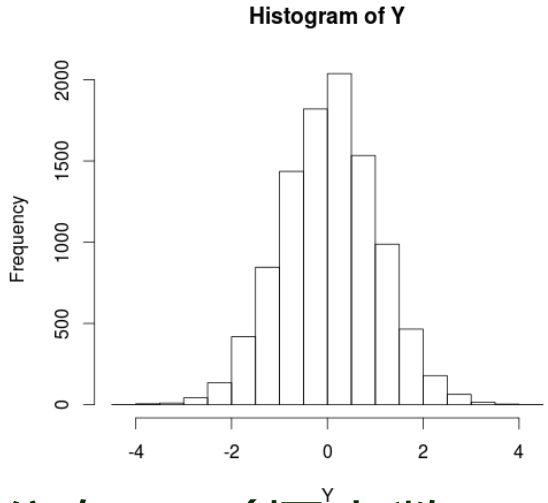
```
> t.test(X, Y, var.equal=F)$p.value
[1] 4.833646e-06
```

「t.test(X, Y, var.equal=F)$p.value」で t 検定

Rプログラムの例

t 検定の例 (2/2)





■ 分布 X （標本数: 10000）

※ 平均0, 標準偏差1 になるように合成

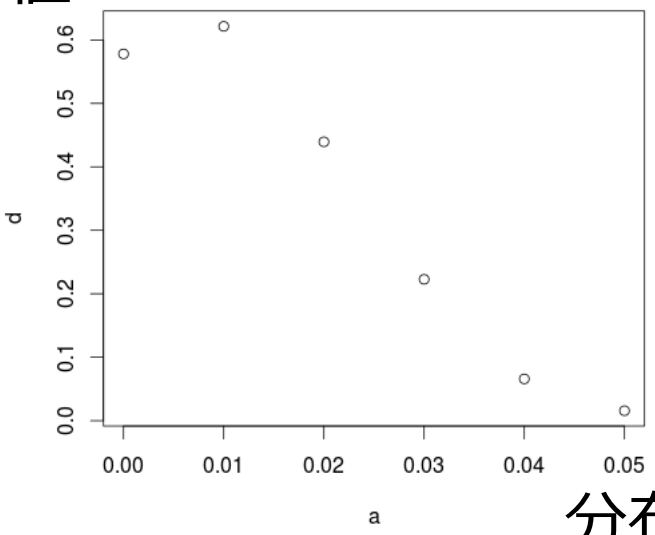
■ 分布 Y （標本数: 10000）

※ 平均を {0, 0.01, 0.02, 0.03, 0.04, 0.05}, 標準偏差1 になるように合成

```
> X = rnorm(10000)
> a = c(0, 0.01, 0.02, 0.03, 0.04, 0.05)
> d = numeric(6)
> c = 1
> for (i in a) {
+     p = numeric(100)
+     for (j in 1:100) {
+         Y <- rnorm(10000, mean=i)
+         b <- t.test(X, Y, var.equal=F)
+         p[j] = b$p.value
+     }
+     print(i)
+     d[c] = mean(p)
+     c = c + 1
+ }
```

Rプログラムの例

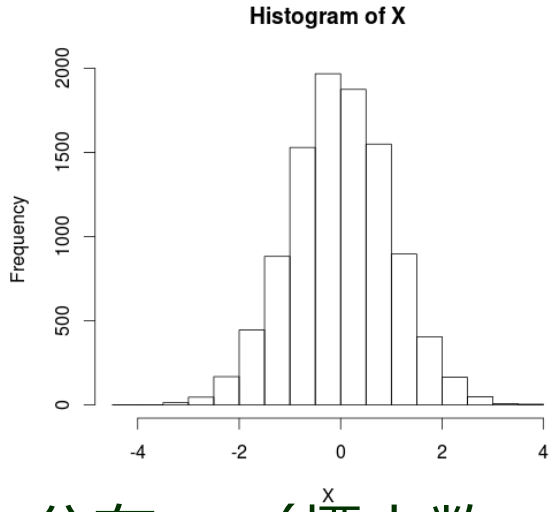
p値

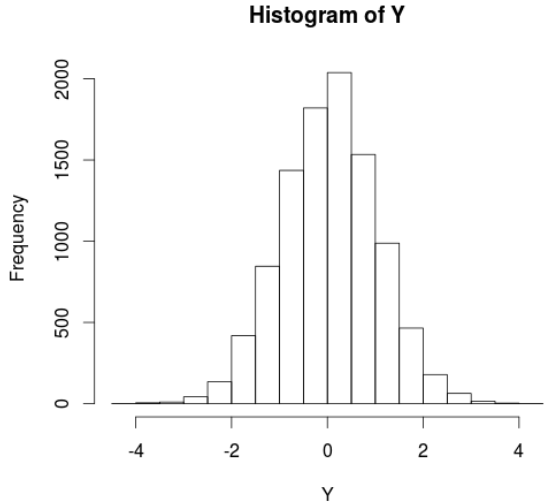


標本が10000個あれば平均 0.05 の差の識別ができる

分布Yの平均値

ノンパラメトリック検定の例 (1/2)





- 分布 X （標本数: 10000）
  ※ 平均0, 標準偏差1 になるように合成
- 分布 Y （標本数: 10000）
  ※ 平均0.05, 標準偏差1 になるように合成

帰無仮説：「分布X, Y の母集団の平均が等しい」
⇒ t 検定により，「帰無仮説が成り立つ確率は 0.00054 % (=p値)」

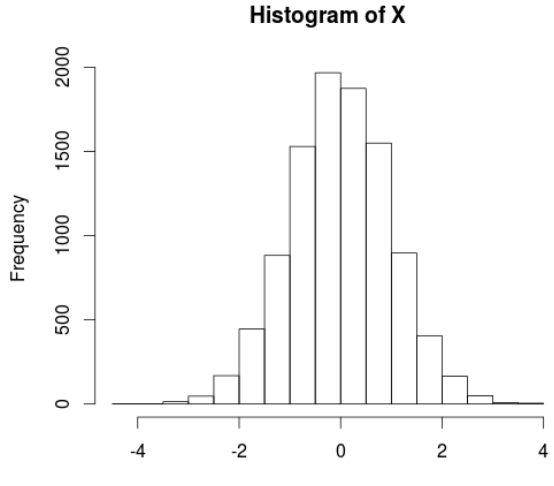
```
> wilcox.test(X, Y, correct=F)

 Wilcoxon rank sum test

data:  X and Y
W = 48142366, p-value = 5.361e-06
alternative hypothesis: true location shift is not equal to 0
```

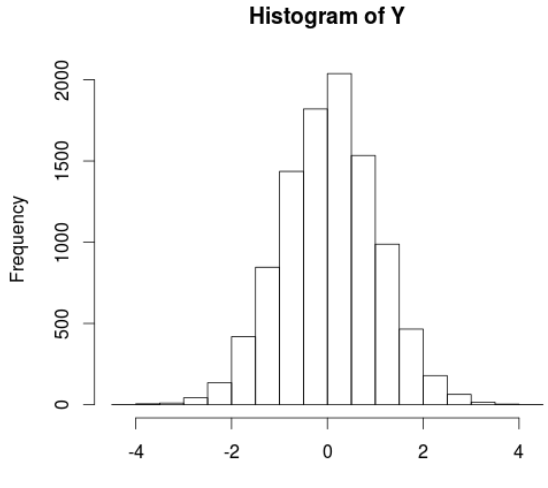
Rプログラムの例

「wilcox.test(X, Y, correct=F)」で検定

ノンパラメトリック
⇒ 「母集団が正規分布だと仮定しない」

ノンパラメトリック検定の例 (2/2)



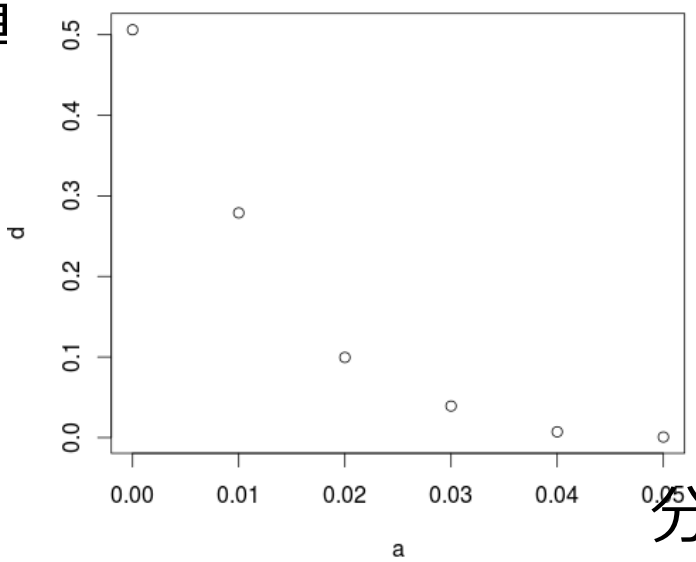


- 分布 X （標本数: 10000）
  - ※ 平均0, 標準偏差1 になるように合成
- 分布 Y （標本数: 10000）
  - ※ 平均を {0, 0.01, 0.02, 0.03, 0.04, 0.05}, 標準偏差1 になるように合成

```
> X = rnorm(10000)
> a = c(0, 0.01, 0.02, 0.03, 0.04, 0.05)
> d = numeric(6)
> c = 1
> for (i in a) {
+     p = numeric(100)
+     for (j in 1:100) {
+         Y <- rnorm(10000, mean=i)
+         b <- wilcox.test(X, Y, correct=F)
+         p[j] = b$p.value
+     }
+     print(i)
+     d[c] = mean(p)
+     c = c + 1
+ }
```

Rプログラムの例

p値

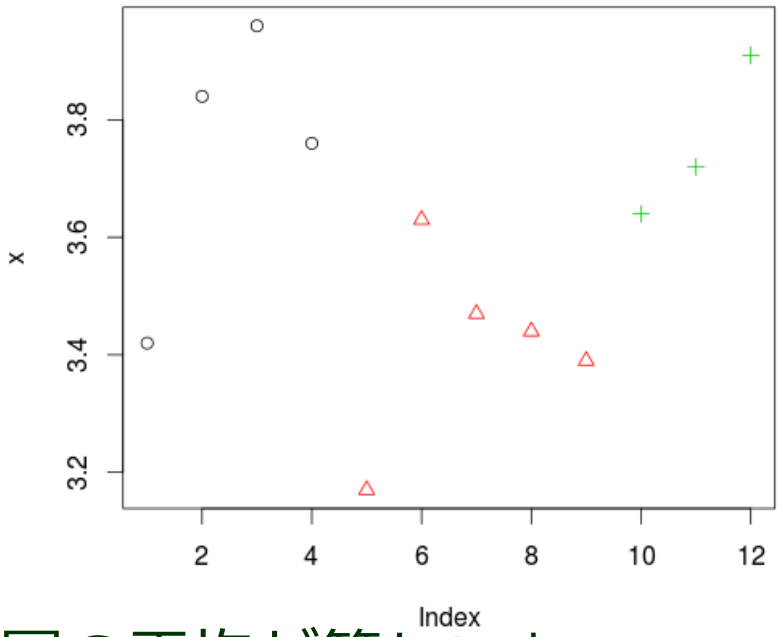


分布Yの平均値

一次元分散配置に関する検定の例

■ 元データの例

種別	観測値
A	{3.42, 3.84, 3.96, 3.76}
B	{3.17, 3.63, 3.47, 3.44, 3.39}
C	{3.64, 3.72, 3.91}



帰無仮説：「分布A, B, C の母集団の平均が等しい」

⇒ 検定により，「帰無仮説が成り立つ確率は 5.9 % (=p値)」

```
> x <- c(3.42, 3.84, 3.96, 3.76, 3.17, 3.63, 3.47, 3.44, 3.39, 3.64, 3.72, 3.91)
> g <- c(1,1,1,1,2,2,2,2,2,3,3,3)
> oneway.test( x ~ g, var.equal=F )$p.value
[1] 0.05907792
```

Rプログラムの例

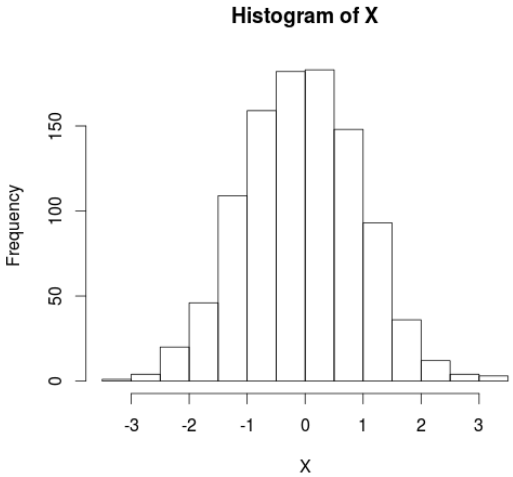
元データの確認

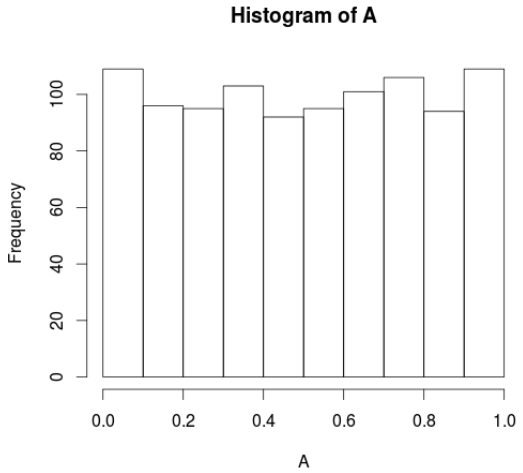
「oneway.test(x~g, var.equal=F)$p.value」で検定 (A, B, C の等分散を仮定しない)

Wilcoxon の順位和検定の例

データの引用元: http://aoki2.si.gunma-u.ac.jp/lecture/Average/oneway-ANOVA.htm

正規分布かの検定の例





■ 正規分布 X （標本数: 10000）■ 一様分布 A （標本数: 10000）

帰無仮説：「母集団の分布が正規分布である」

75 % (=p値)

0.000000000000022 % (=p値)

```
> X = rnorm(1000)
> shapiro.test(X)

 Shapiro-Wilk normality test

data:  X
W = 0.9988, p-value = 0.7577
```

Rプログラムの例

Shapiro-Wilk 検定

```
> A = runif(1000)
> shapiro.test(A)

 Shapiro-Wilk normality test

data:  A
W = 0.9536, p-value < 2.2e-16
```

Rプログラムの例