データ解析
第八回「検定」

鈴木　大慈
理学部情報科学科
西八号館 W707 号室
s-taiji@is.titech.ac.jp

休講情報

6/24 は休講

今日の講義内容

- 正規性検定
- ２群の比較
  - t-検定
  - Wilcoxon の符号付順位和検定
- 適合度検定
- 独立性検定
- 分散分析

構成

正規性検定

使いドコロ：いろんな検定は変数が正規分布に従うと仮定するけれども，本当に正規分布？
→ 正規性検定

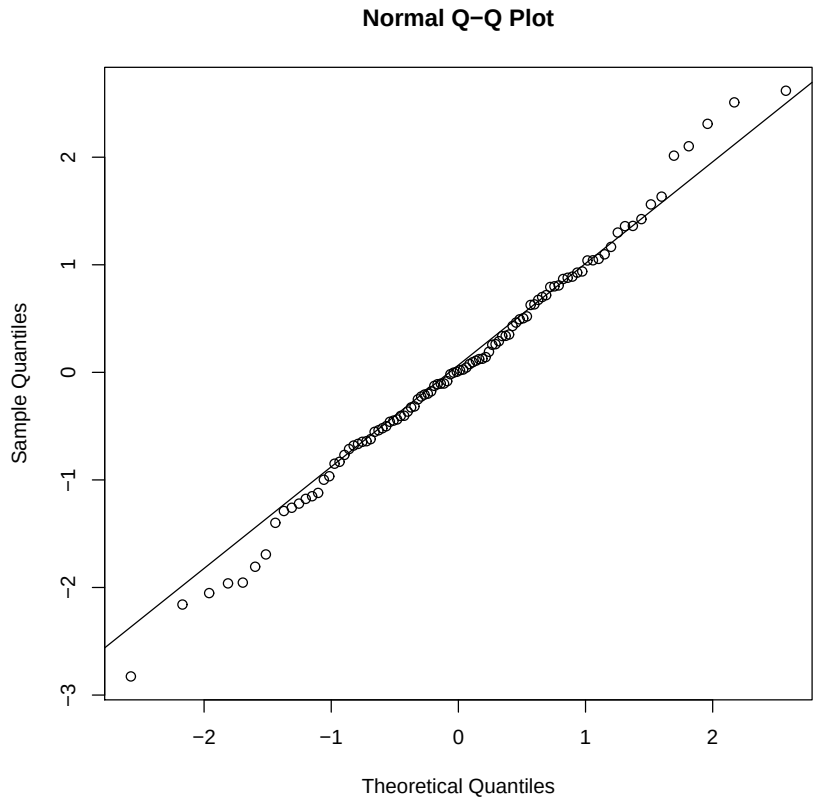
次の 2 つの検定を紹介

- Shapiro-Wilk 検定
- Kolmogorov-Smirnov 検定

※ 正規性検定で棄却されなかったからといって，積極的にその分布が正規分布に従っているとは言いにくい．検定は積極的に棄却はするが，積極的に採択はしない．

正規性検定の前に

Q-Q プロット：標準正規分布における分位点 vs 経験的分位点
(例えば n サンプル中 i 番目のサンプル $x_{(i)}$ は標準正規分布の i/n 分位点と観測値 $x_{(i)}$ を対応させてプロットされる)



対角線から離れていればいるほど正規分布から遠い.
これから紹介する方法はこの離れ具合を検定統計量としている.

Shapiro-Wilk 検定

W = 本当の正規分布からの順序統計量の期待値とサンプルの順序統計量との相関（のようなもの）
値が小さければ正規性が棄却される.

```
> x <- rnorm(100)
> shapiro.test(x)

        Shapiro-Wilk normality test

data:  x
W = 0.9926, p-value = 0.86

> shapiro.test(exp(x))

        Shapiro-Wilk normality test

data:  exp(x)
W = 0.6118, p-value = 7.267e-15
```

Kolmogorov-Smirnov 検定

サンプル： $\{x_i\}_{i=1}^n$.
経験分布関数:

$$F_n(x) = \frac{x \text{ より小さいサンプル } x_i \text{ の数}}{n}.$$

もし，真の分布の分布関数が（連続な）$F(x)$ であれば，$\sup_x |F_n(x) - F(x)| \xrightarrow{p} 0$ となる.

Kolmogorov-Smirnov 検定

サンプル： $\{x_i\}_{i=1}^n$.
経験分布関数:

$$F_n(x) = \frac{x \text{ より小さいサンプル } x_i \text{ の数}}{n}.$$

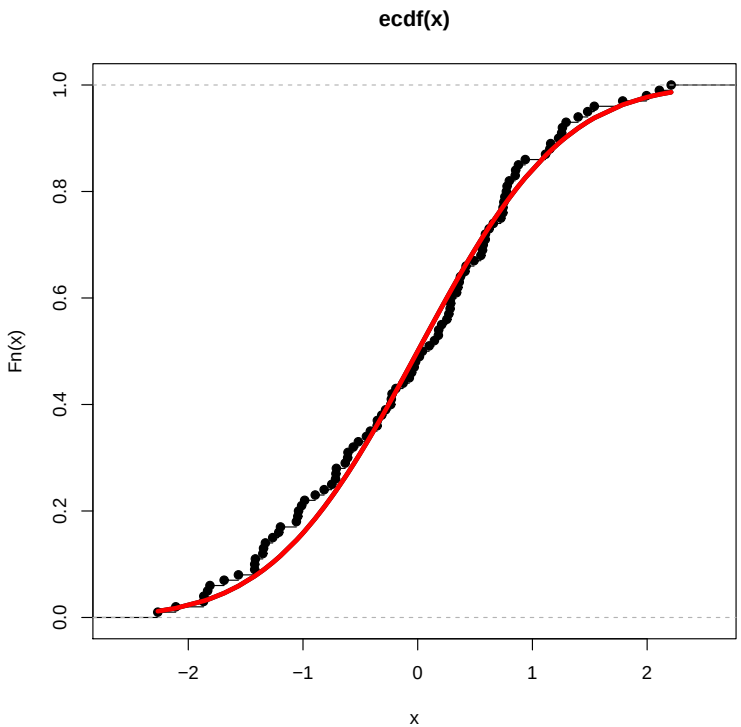
もし，真の分布の分布関数が（連続な）$F(x)$ であれば，$\sup_x |F_n(x) - F(x)| \xrightarrow{p} 0$ となる.

さらに

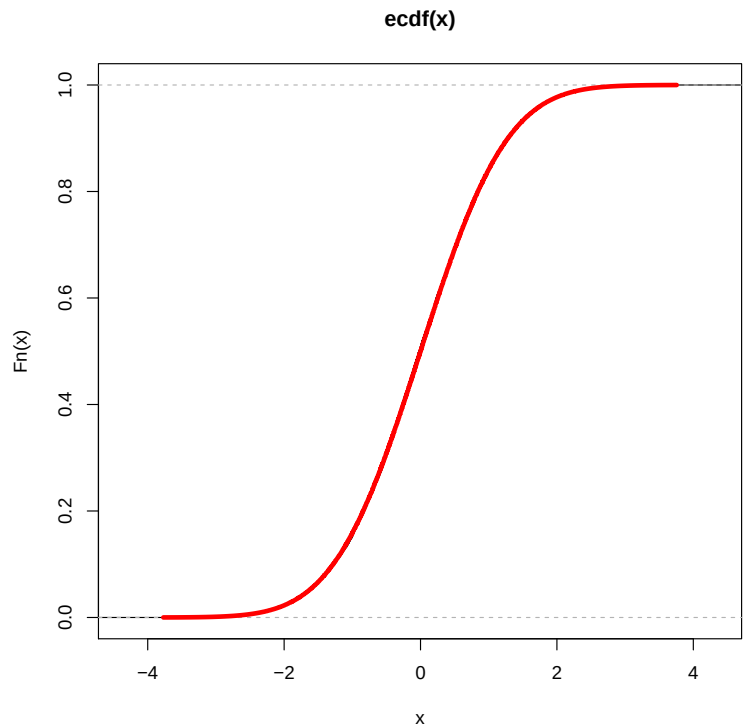
$$P(\sqrt{n} \sup_x |F_n(x) - F(x)| \leq t) \to \frac{\sqrt{2\pi}}{t} \sum_{i=1}^{\infty} e^{-(2i-1)^2\pi^2/(8t^2)}.$$

導出はとっても難しいので省略.
とにかく漸近分布が求まる.

```
> x <- rnorm(100) # rnorm(10000)
> plot(ecdf(x))
> y <- sort(x)
> lines(y,pnorm(y),lwd = 4,col="red")
```



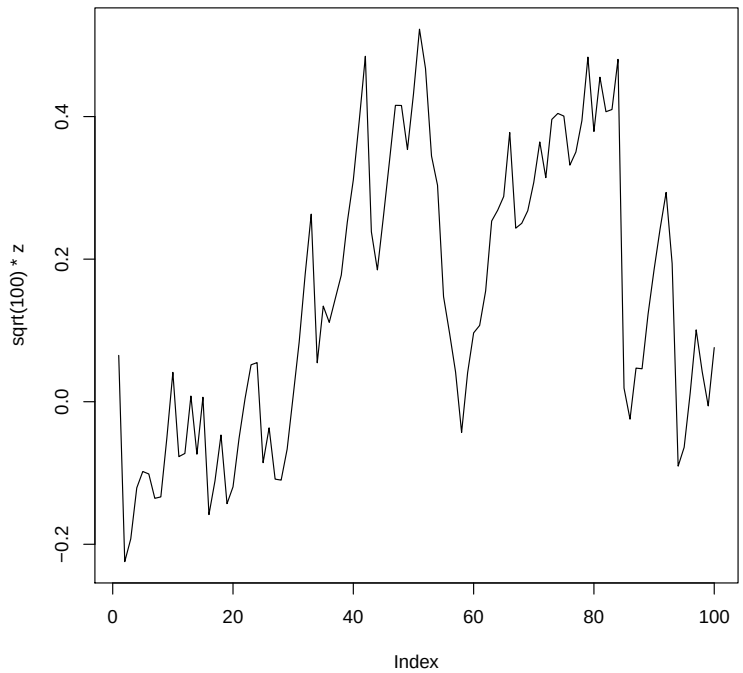
$n = 100$



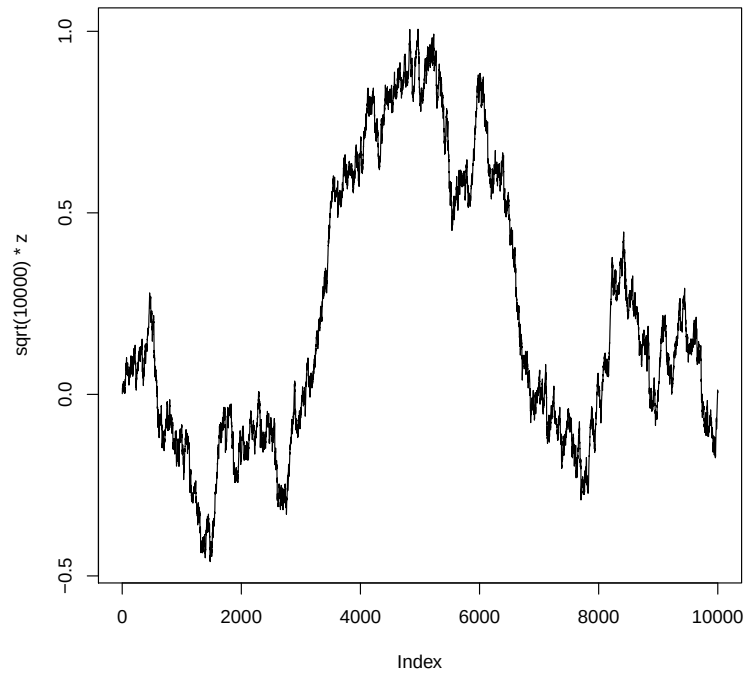
$n = 10000$

$\sqrt{n}(F_n(x) - F(x))$ をプロット

```
> x <- rnorm(100) # rnorm(10000)
> y <- sort(x)
> z <- ecdf(y)(y) - pnorm(y)  #経験分布関数と真の分布関数との差
>  plot(sqrt(100)*z,type='l') #plot(sqrt(10000)*z,type='l')
```



$n = 100$



$n = 10000$

ecdf(x)

ecdf(x)

Kolmogorov-Smirnov 検定を使ってみる

K-S 検定はあらゆる (連続な) 分布関数を帰無仮説にできる.
正規分布の場合は以下のとおり.

```
> x <- rnorm(100)
> ks.test(x, "pnorm", mean=mean(x), sd=sqrt(var(x)))

        One-sample Kolmogorov-Smirnov test

data:  x
D = 0.0678, p-value = 0.7482
alternative hypothesis: two-sided
> y <- exp(x)
> ks.test(y, "pnorm", mean=mean(y), sd=sqrt(var(y)))

        One-sample Kolmogorov-Smirnov test

data:  y
D = 0.2449, p-value = 1.237e-05
alternative hypothesis: two-sided
```

構成

２群の比較

- t-検定 (パラメトリック検定)
  ２つの**正規分布**の平均値が異なるかを検定.
- Wilcoxon の符号付順位和検定 (ノンパラメトリック検定)
  ２つの分布の 中央値 が異なるかを検定.

ちなみに

- **パラメトリック検定**：分布が特定のモデルに含まれていると仮定して検定
- **ノンパラメトリック検定**：パラメトリックモデルの仮定をしない検定

パラメトリックモデルの仮定が正しければパラメトリックの方が検出力が高い.
ノンパラメトリックのほうが仮定が少なくて済む分，保守的.

よくやる使い分け：

- 正規性検定を通過→ t-検定
- 正規性検定で棄却→ Wilcoxon 検定

t-検定

２つの分布が正規分布に従っている時に，その平均値が等しいかどうかを検定.
(正規分布を仮定しているのでパラメトリック検定)
帰無仮説: ２群は平均が等しく分散も等しい正規分布.

$$X_i \sim N(\mu, \sigma^2)\ (i = 1, \ldots, n_1), \tag{1}$$
$$Y_i \sim N(\mu, \sigma^2)\ (i = 1, \ldots, n_2) \tag{2}$$

$V := \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2 + \sum_{i=1}^{n}(Y_i - \bar{Y})^2}{n_1 + n_2}$.　プールされた不偏分散

$$t = \frac{\bar{X} - \bar{Y}}{\sqrt{V(\frac{1}{n_1} + \frac{1}{n_2})}}$$

は自由度 $n_1 + n_2 - 2$ の t-分布に従う.
$|t| \geq t_\alpha$ の時に等平均であることを棄却 (両側検定).

※ ２つの正規分布の分散が異なる場合はウェルチの t 検定を用いる．ここでは省略．等分散性の検定は F 検定を使う.

t-検定を使う

平均が等しい時

```
> x <- rnorm(100)
> y <- rnorm(100)
> t.test(x,y)

        Welch Two Sample t-test

data:  x and y
t = 0.255, df = 195.453, p-value = 0.799
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.2377692  0.3083930
sample estimates:
 mean of x  mean of y
0.04628813 0.01097624
```

R version 3.1.0 では Weltch の t 検定がデフォルト.

t-検定を使う

平均が等しくない時.

```
> x <- rnorm(100)
> y <- rnorm(100) + 1
> t.test(x,y)

        Welch Two Sample t-test

data:  x and y
t = -5.1183, df = 197.983, p-value = 7.273e-07
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.0747345 -0.4769039
sample estimates:
mean of x mean of y
0.1717119 0.9475311
```

t.test(x,y,var.equal=T) とすれば分散が等しい場合. (Student t-検定)

Wilcoxon の順位和検定

２つの分布（正規分布とは限らない）の 中央値 が等しいかどうかを検定.
(特に分布型を仮定していないのでノンパラメトリック検定)

1. 第一群より $X_1, \ldots, X_m$，第二群より $Y_1, \ldots, Y_n$ を得る.
2. ２つの列を一列に並べる: $X_1, \ldots, X_m, Y_1, \ldots, Y_n$.
3. これを小さい順に並べて，Y_i の順番を R_i とする.
4. $W = \sum_{i=i}^{n} R_i$ を計算→ Wilcoxon の順位和検定.

W が大きければ，相対的に Y の分布のほうが大きいことになる.
帰無仮説が正しい時，正規分布で近似できる (Mann-Whitney の U-統計量).

２つの分布が「等しいか」どうかのノンパラメトリック検定は
Kolmogorov-Smirnov 検定などがある.

Wilcoxon の順位和検定を使う

中央値の等しい指数分布

```
> x <- rexp(100)
> y <- rexp(100)
> wilcox.test(x,y)

        Wilcoxon rank sum test with continuity correction

data:  x and y
W = 5136, p-value = 0.7406
alternative hypothesis: true location shift is not equal to 0
```

Wilcoxon の順位和検定を使う

中央値の異なる指数分布

```
> x <- rexp(100)
> y <- rexp(100,rate = 3)
> wilcox.test(x,y)

        Wilcoxon rank sum test with continuity correction

data:  x and y
W = 8103, p-value = 3.439e-14
alternative hypothesis: true location shift is not equal to 0
```

構成

適合度検定

χ^2 検定

すべての目が等しい確率のサイコロの検定:

```
chisq.test(c(8, 12, 10, 9, 5, 6))
```

(帰無仮説：すべての目が出る確率が等しい)

p を指定して，サイコロの眼の出る確率を検定:

```
chisq.test(c(20,8,5,2), p=c(4, 3, 2, 1)/10)
```

(帰無仮説：それぞれの目がでる確率が $4/10, 3/10, 2/10, 1/10$ である)

独立性検定

要因が 2 つある.
要因 1 の水準 i がでる確率$=p_i$, 要因 2 の水準 j がでる確率$=q_j$.

帰無仮説：要因 1 の水準が i かつ要因 2 の水準が j である確率$=p_i \times q_j$.　独立！

	A_1	A_2	
B_1	$n_{1,1}$	$n_{1,2}$	$n_{1,.}$
B_2	$n_{2,1}$	$n_{2,2}$	$n_{2,.}$
	$n_{.,1}$	$n_{.,2}$	$n_{.,.}$

帰無仮説のもと

$$\sum_{i=1}^{r}\sum_{j=1}^{c}\frac{\left(\frac{n_{i,.}n_{.,j}}{n_{.,.}}-n_{i,j}\right)^2}{\frac{n_{i,.}n_{.,j}}{n_{.,.}}}$$

は漸近的に自由度 $(r-1)(c-1)$ の χ^2 分布に従う.

χ^2 独立性検定の使い方

	良品	不良品
A 工場	197	7
B 工場	96	12

```
> x <- matrix(c(197,96,7,12),nrow=2)
> chisq.test(x)

        Pearson's Chi-squared test with Yates' continuity correcti

data:  x
X-squared = 6.0015, df = 1, p-value = 0.01429
```

→独立性は棄却

※ R の chisq.test は Yates の補正がかかっているので，前のページの式とはちょっと異なる．

correct = FALSE を指定すれば補正は切れる．

構成

分散分析

一元分散分析：

$$
\begin{aligned}
A_1 &: Y_{1,1}, \ldots, Y_{1,n_1} \sim N(\mu_1, \sigma^2) \\
A_2 &: Y_{2,1}, \ldots, Y_{2,n_2} \sim N(\mu_2, \sigma^2) \\
&\vdots \\
A_r &: Y_{r,1}, \ldots, Y_{r,n_r} \sim N(\mu_r, \sigma^2)
\end{aligned}
$$

帰無仮説：$\mu_1 = \mu_2 = \cdots = \mu_r$.

$$Y_{ij} = \mu + a_i + \epsilon_{ij}$$

として，$a_i = 0$ $(\forall i)$ かどうかの検定ともみなせる．
→ 線形回帰.

二元分散分析

二元分散分析：

$$Y_{ijk} = \mu + a_i + b_j + \gamma_{ij} + \epsilon_{ijk}$$

帰無仮説：

- $a_i = 0\ (\forall i) \rightarrow$ 要因 A の主効果
- $b_j = 0\ (\forall j) \rightarrow$ 要因 B の主効果
- $\gamma_{ij} = 0\ (\forall i, j) \rightarrow$ 交互作用

分散分析を実行する

```
(fm <- lm(wear ~ material+boy,data=boxshoes))
(av <- anova(fm))
```

これだけで OK.

交互作用も入れたければ
`(fm <- lm(wear ~ (material+boy)^2,data=boxshoes))`
のようにする.

分散分析表の見方

```
Analysis of Variance Table

Response: wear
          Df  Sum Sq Mean Sq F value    Pr(>F)
material   1   0.841  0.8405  11.215  0.008539 **
boy        9 110.491 12.2767 163.811 6.871e-09 ***
Residuals  9   0.675  0.0749
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’
1
```

左から自由度（Degree of freedom）, 平方和 (主効果)， 平均平方和 (平方和を自由度で割ったもの)，F 値, p-値

行は要因を表す．この場合，material と boy という要因がある． Residuals はこの 2 つでは説明できない部分．

p-値の横に*が付いている要因は有意に効果があることを表している．

講義情報ページ

http://www.is.titech.ac.jp/~s-taiji/lecture/dataanalysis/dataanalysis.html