



統計モデリング入門: エンジニアならおさえておきたい 5 つの仮説検定テクニック



吉野 紘和, PhD

Application Engineering

$$p(\mathbf{x}) = \frac{\exp\left\{\frac{1}{2}(\mathbf{x} - \mu)^{\top} \Sigma^{-1}(\mathbf{x} - \mu)\right\}}{(2/\pi)^{n/2}}$$

$$\Sigma = \text{E} \left[(\mathbf{x} - \mu) (\mathbf{x} - \mu)^{\top} \right]$$



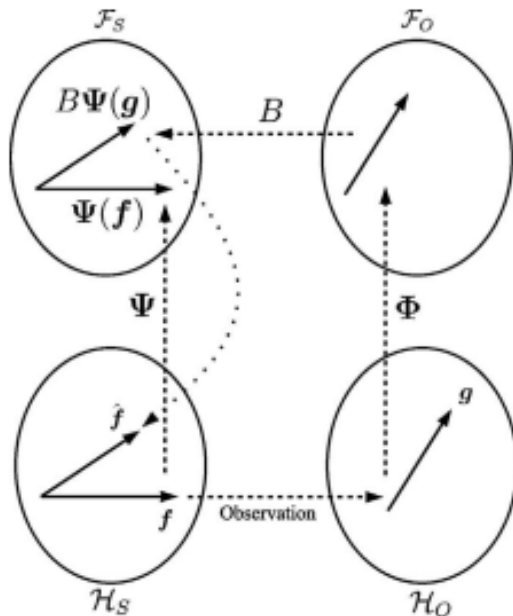
Abstract

- 本セミナーでは確率変数・確率分布からスタートして仮説検定とは何かを理解し、現場でよくある状況を取り上げ、仮説検定を適用していきます。また、仮説検定の応用の1つとして機械学習における特徴選択についても解説します。
- 以下の内容に触れる予定です:
 - 記述統計と可視化
 - 確率分布
 - 仮説検定
 - 分散分析
 - 仮説検定を利用した特徴選択

はじめての t 検定 @ パターン認識

分類問題を解く新手法!

$$B_{\text{aKWF}} = \sum_{i=1}^L \sum_{k=1}^L w_{ik} \Psi(f_i) \Phi(h_k)^T.$$



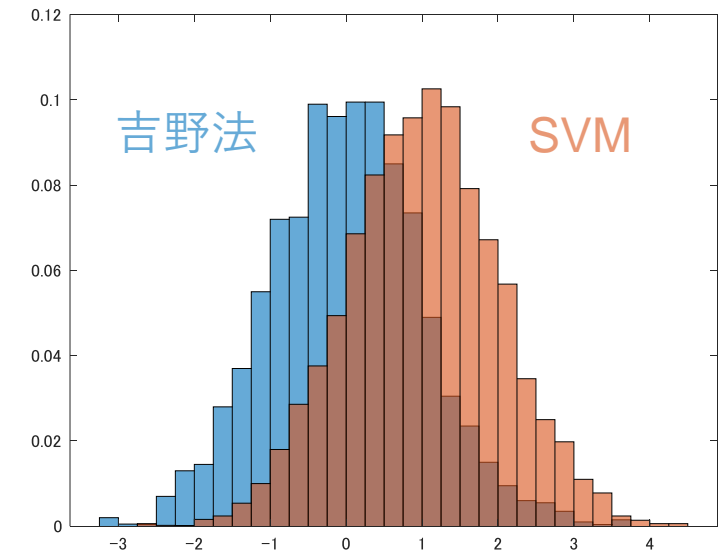
誤認識率

Data Set #	SVM	吉野法
1	0.032	<u>0.031</u>
2	0.040	0.044
3	0.051	<u>0.047</u>
4	0.027	<u>0.026</u>
...
99	0.030	0.035
100	0.037	<u>0.036</u>

何となく
良さそう

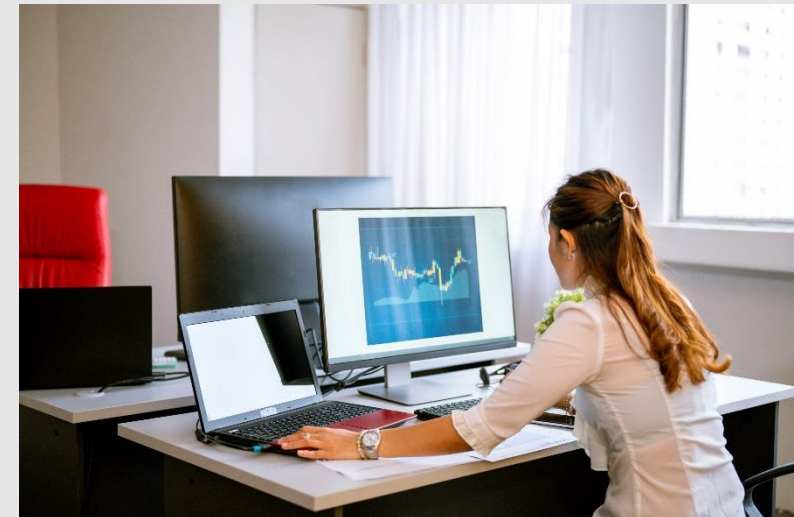
T 検定

優位に差が有る!



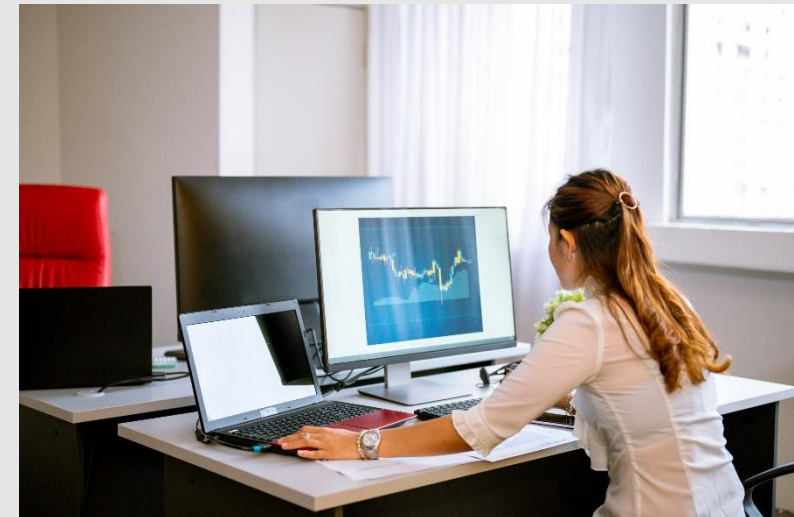
Agenda

- 統計学の違い
- 確率変数とは
- 確率分布
- 区間推定
 - 母平均区間推定
 - 母分散区間推定
- 仮説検定
 - 母平均仮説検定
 - T検定 & ウェルチの検定
 - 一元配置分散分析 (One-way ANOVA)
- 周囲に展開する方法
 - Live Editor
 - アプリ化



Agenda

- 統計学の違い
- 確率変数とは
- 確率分布
- 区間推定
 - 母平均区間推定
 - 母分散区間推定
- 仮説検定
 - 母平均仮説検定
 - T 検定 & ウェルチの検定
 - 一元配置分散分析 (One-way ANOVA)
- 周囲に展開する方法
 - Live Editor
 - アプリ化



統計学の違い



記述統計学

データの特徴を分かりやすく表現

- \bar{x} の平均、 s^2 の分散、グラフや表
- 古典統計学 (推計統計学との対比において)

推計統計学

有限の標本から母集団を推測する

- 「手元のサンプル (標本) は偶然今回得られただけ」というスタンス
- 頻度主義

ベイズ統計学

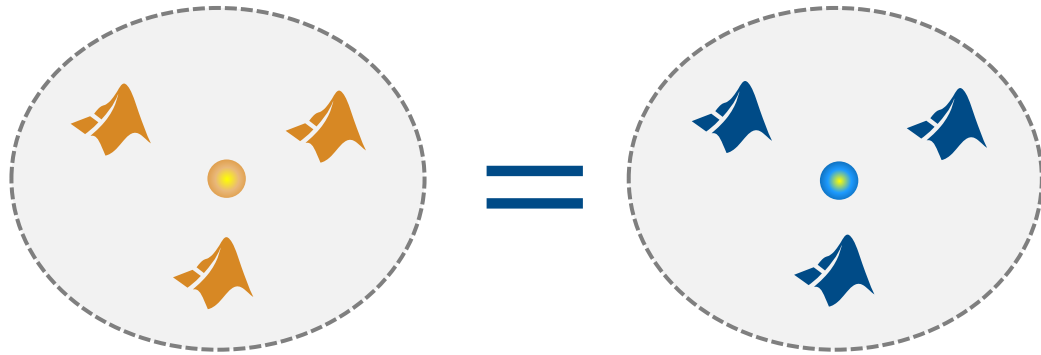
事前知識をデータによって裏付ける

- 主観的な部分があり、頻度主義と相容れない部分がある



記述統計学

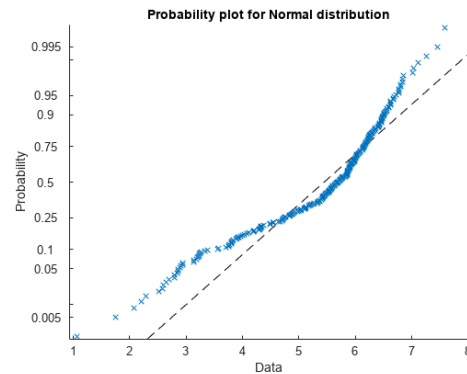
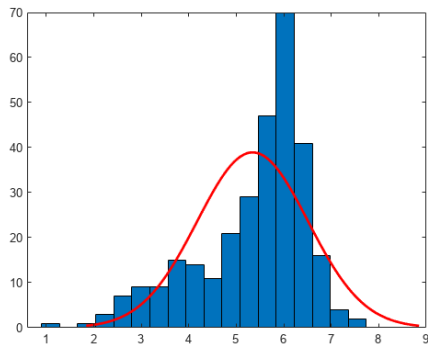
統計量に必要なデータを把握している



標本
(偶然) 観測されたデータ

母集団

パッと分かりやすく



例えば ..

我が家の家族の身長

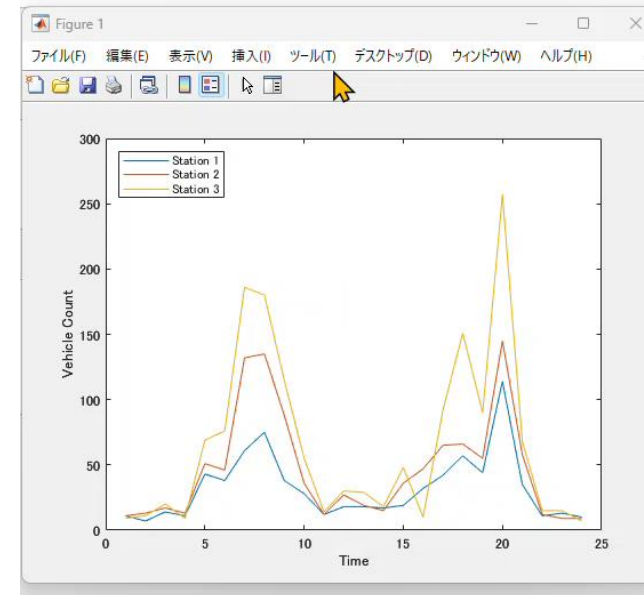
- 平均は 156.5 cm
- 分散は 400 cm^2

全データを把握

~~人類の身長~~

- 平均は 161.5 cm
- 分散は 378 cm^2

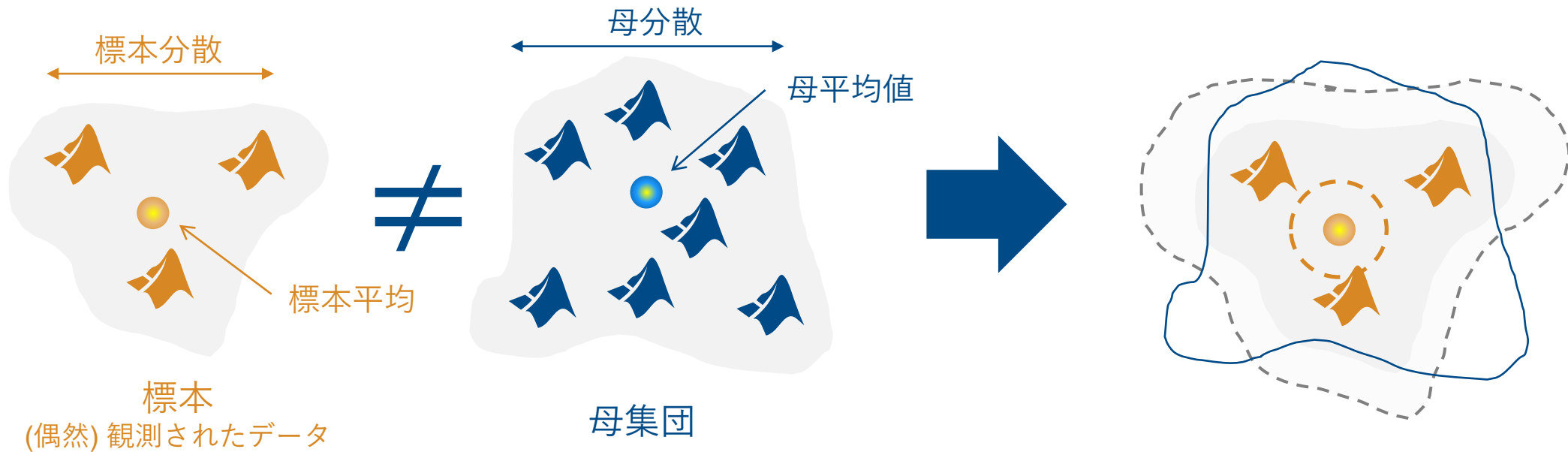
一部から全体を推計



推計統計学

統計量に必要なデータを全て把握していない

一部から全体を推計



“現実的な制約により少数の標本しか得られない”
状況下で母集団の規則性を推計する

記述統計学

便利な統計量の可視化ツール: グループ別に計算

記述統計学 - 可視化ツール
過去の台風のデータ

```
1 load cycloneTracks
2 cycloneTracks
```

グループ別に計算

`newTable` = ID でグループ化された `cycloneTracks` のカウントを計算

▼ 計算対象とするグループとデータを選択する

グループ化: `cycloneTracks` ID 固有の値によるグループ化

計算対象: 指定された変数 Pressure Latitude

▼ グループに対する計算を選択する

グループ別に統計値を計算 グループ別に交換 グループ別にフィルター処理

グループごとの計算: カウント

▶ 結果の表示

3	701	"KONG-REY"	2007-03-31 00:00:00	2
4	701	"KONG-REY"	2007-03-31 06:00:00	2
5	701	"KONG-REY"	2007-03-31 12:00:00	2
6	701	"KONG-REY"	2007-03-31 18:00:00	2
7	701	"KONG-REY"	2007-04-01 00:00:00	3
8	701	"KONG-REY"	2007-04-01 06:00:00	3
9	701	"KONG-REY"	2007-04-01 12:00:00	3

`newTable` = 263x2 table

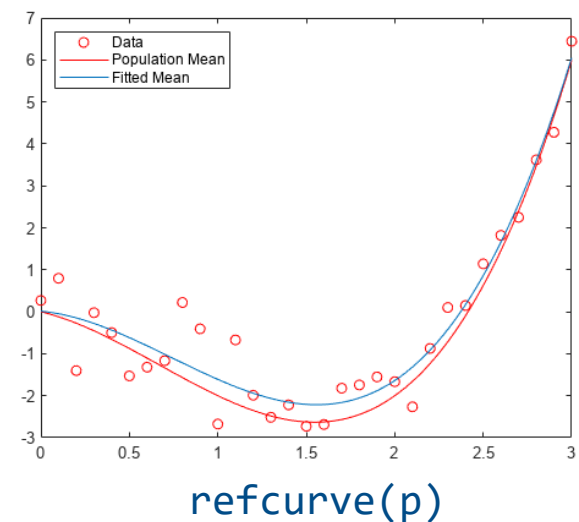
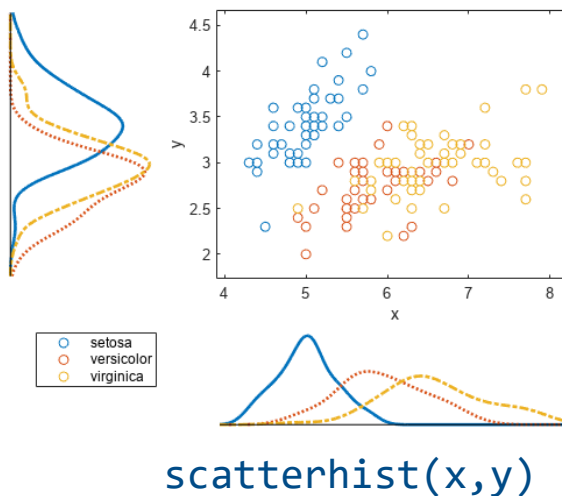
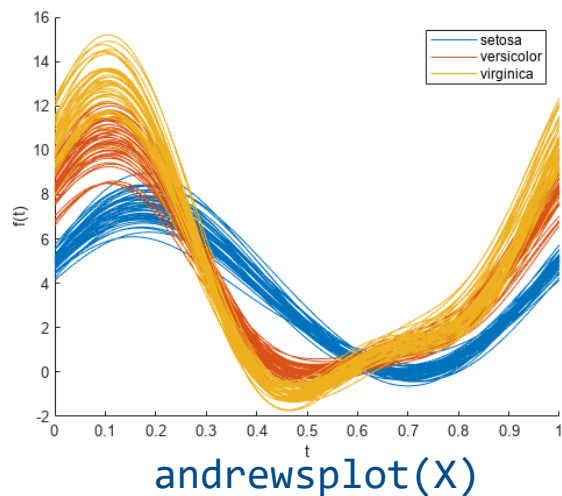
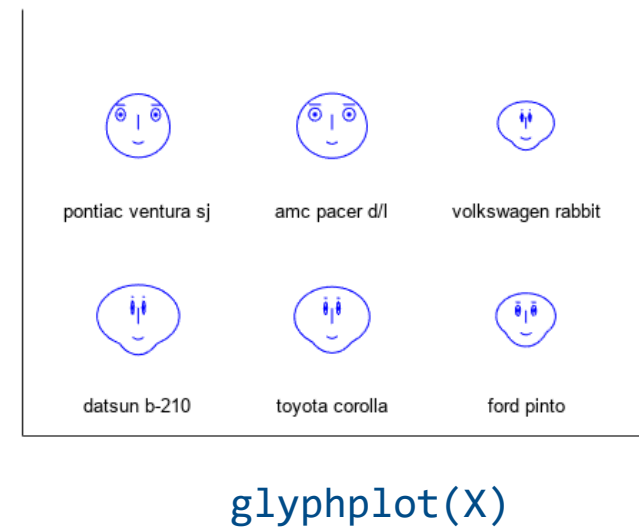
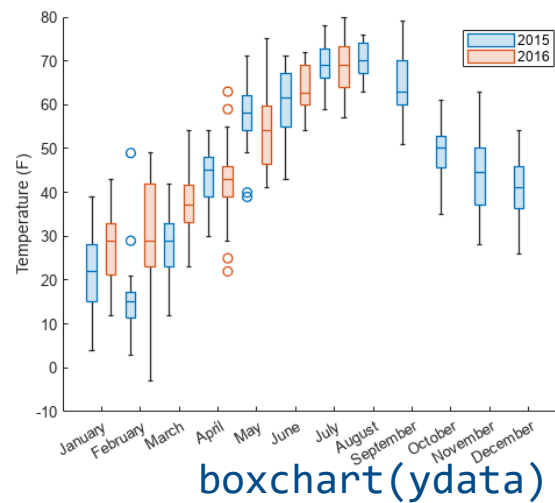
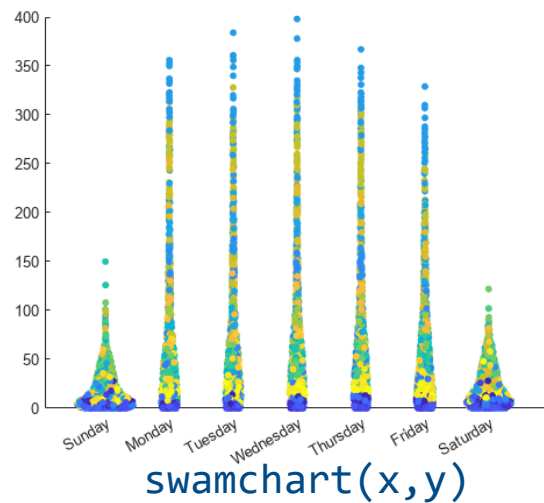
	ID	GroupCount
1	701	30
2	702	43
3	703	13
4	704	81
5	705	48
6	706	46
7	707	12
8	708	54
9	709	65

記述統計学

便利な統計量の可視化ツール

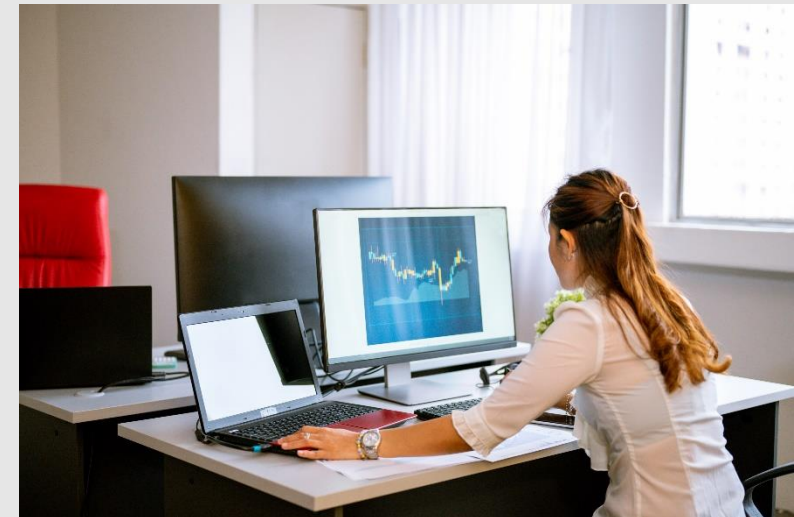
データ分布プロット

統計量の可視化



Agenda

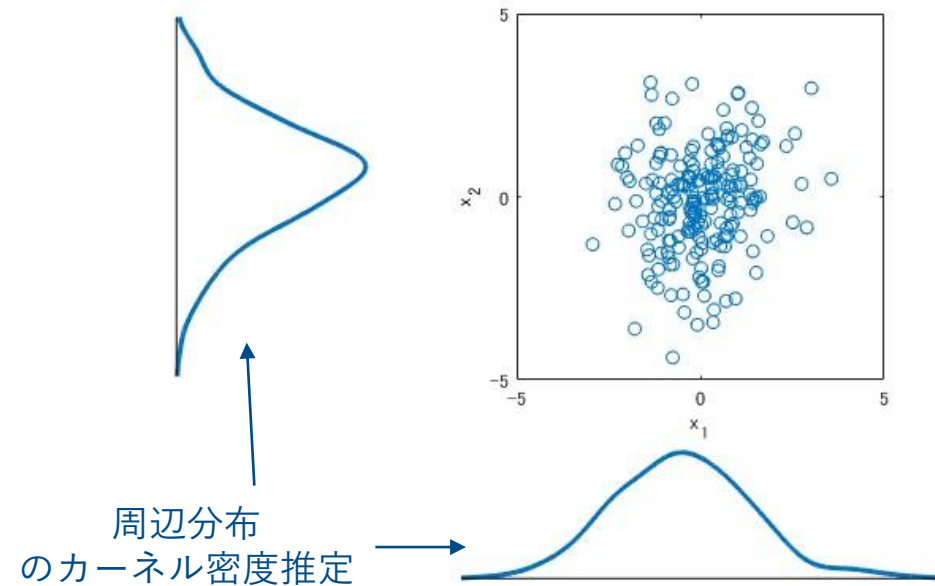
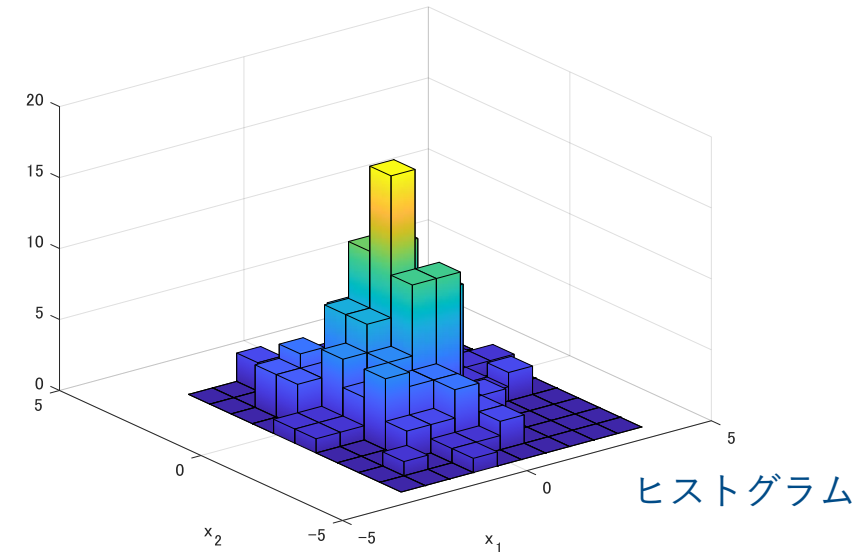
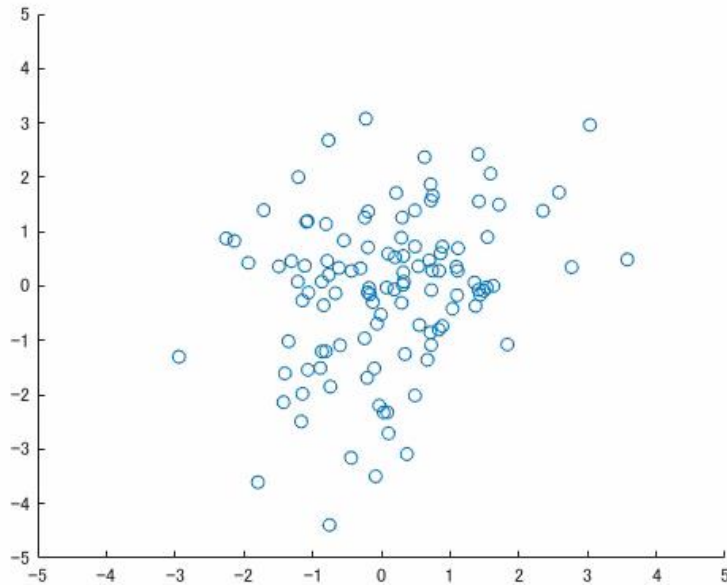
- 統計学の違い
- 確率変数とは
- 確率分布
- 区間推定
 - 母平均区間推定
 - 母分散区間推定
- 仮説検定
 - 母平均仮説検定
 - T 検定 & ウェルチの検定
 - 一元配置分散分析 (One-way ANOVA)
- 周囲に展開する方法
 - Live Editor
 - アプリ化



確率変数とは

- 確率的にある値を取る変数のこと
 - ある値の”取り易さ”がある

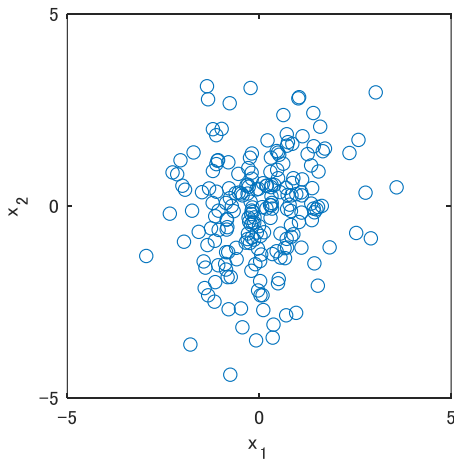
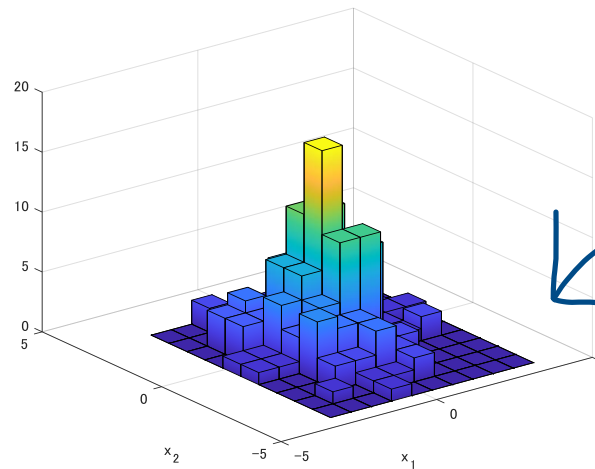
(例) ダーツが刺さる位置: $\mathbf{R} = [X, Y]^T$



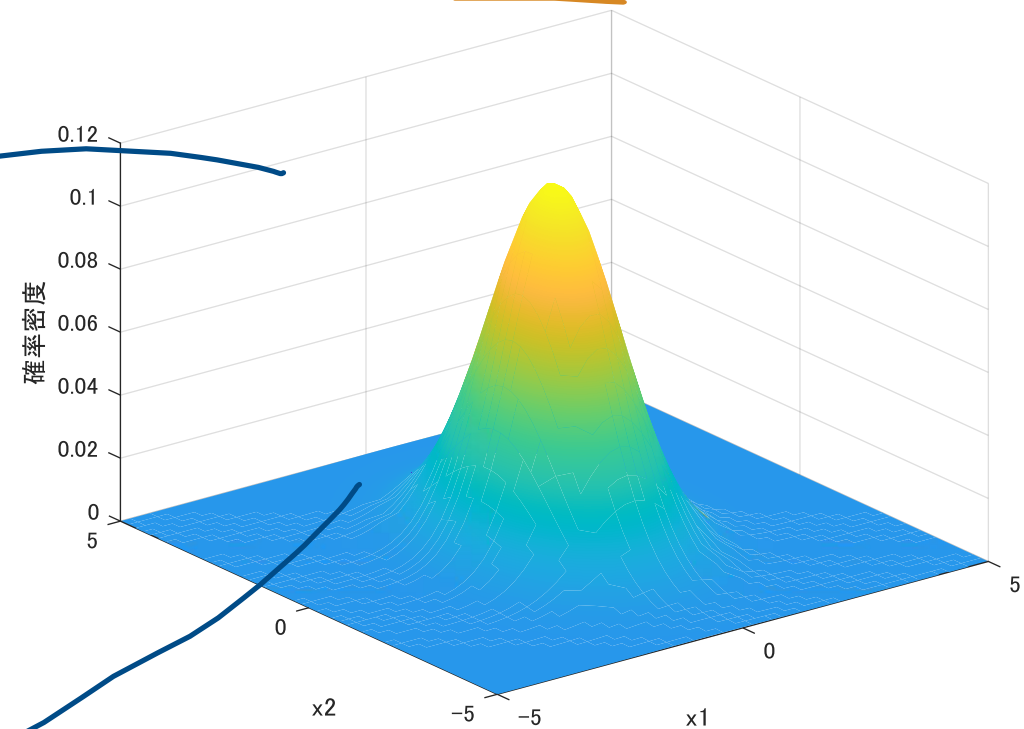
確率変数とは (Cont.)

サンプル (標本) の考え方

(偶然) 観測されたサンプル



元となる確率分布が存在する

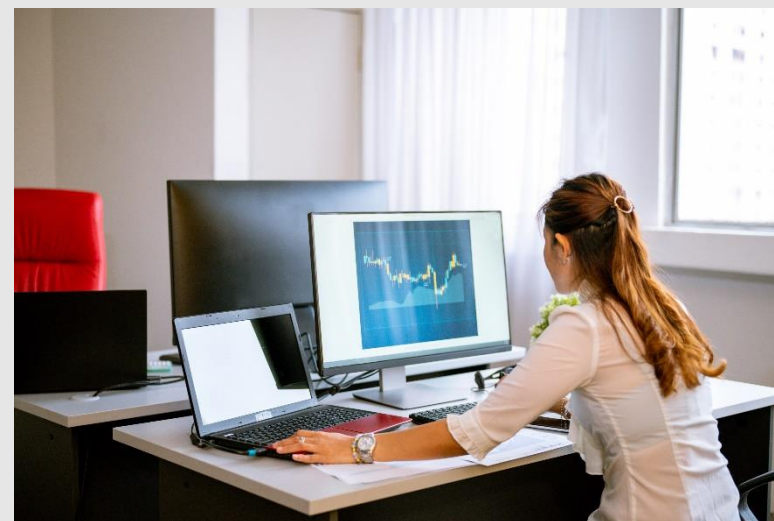


$$y = \text{mvnpdf}(X, \mu, \text{Sigma})$$

多変量正規分布の確率密度分布

Agenda

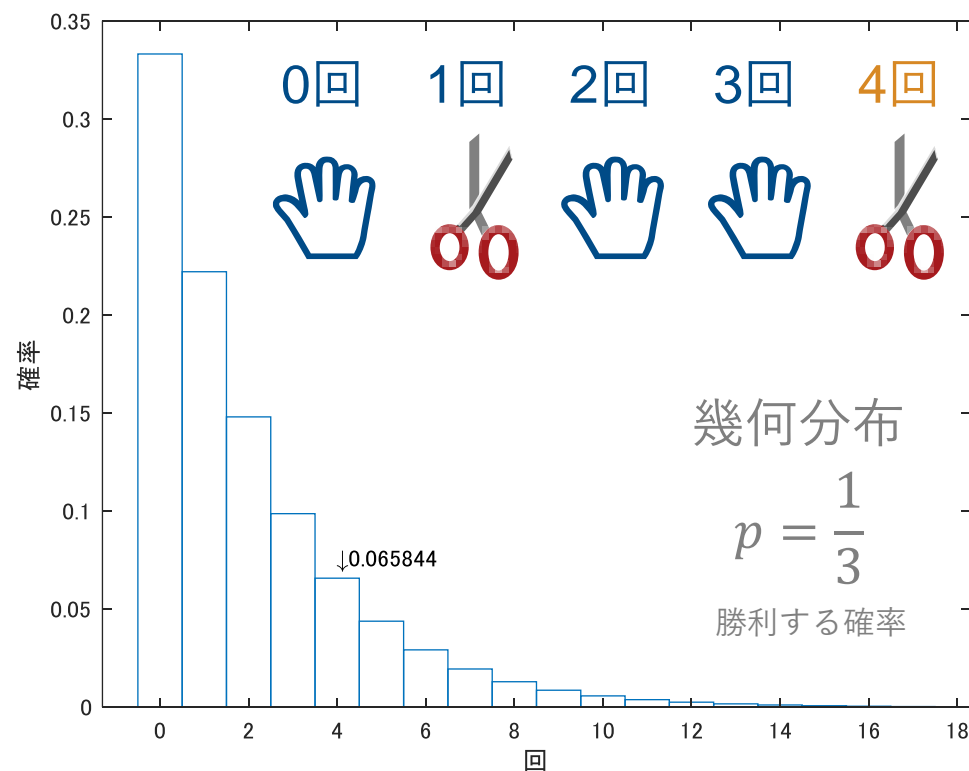
- 統計学の違い
- 確率変数とは
- 確率分布
- 区間推定
 - 母平均区間推定
 - 母分散区間推定
- 仮説検定
 - 母平均仮説検定
 - T検定 & ウェルチの検定
 - 一元配置分散分析 (One-way ANOVA)
- 周囲に展開する方法
 - Live Editor
 - アプリ化



確率分布

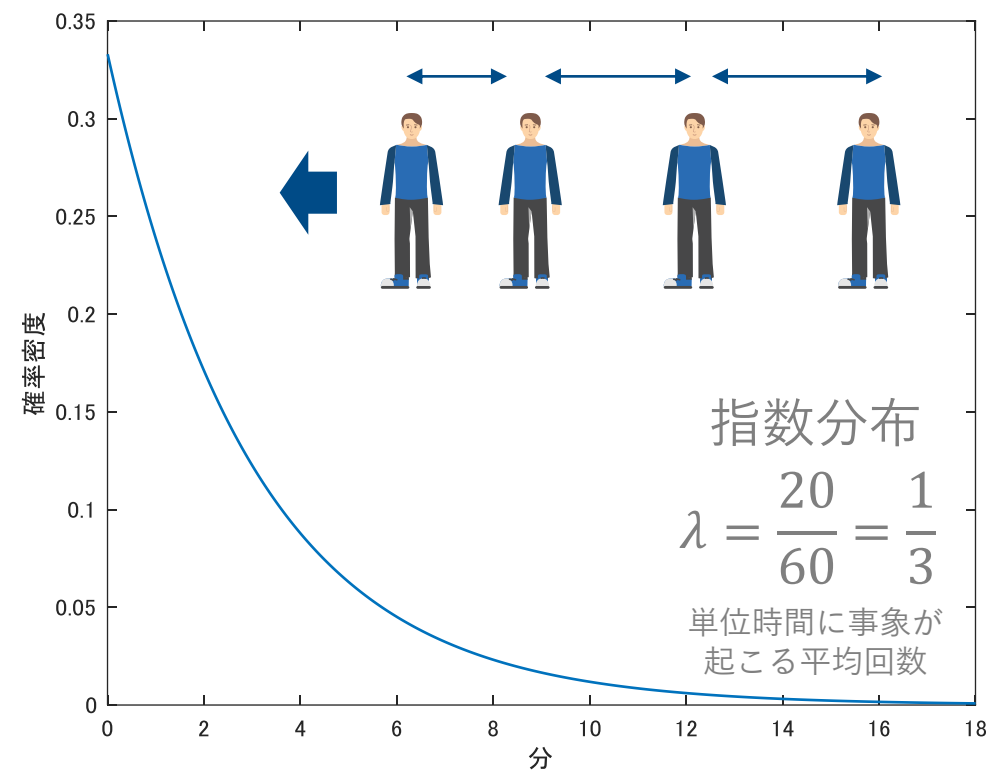
2種類の確率分布

じゃんけんですべて勝つまで X 回失敗



確率変数 X が離散

60分に20人が来る銀行窓口の
お客が来る時間間隔 X (分)

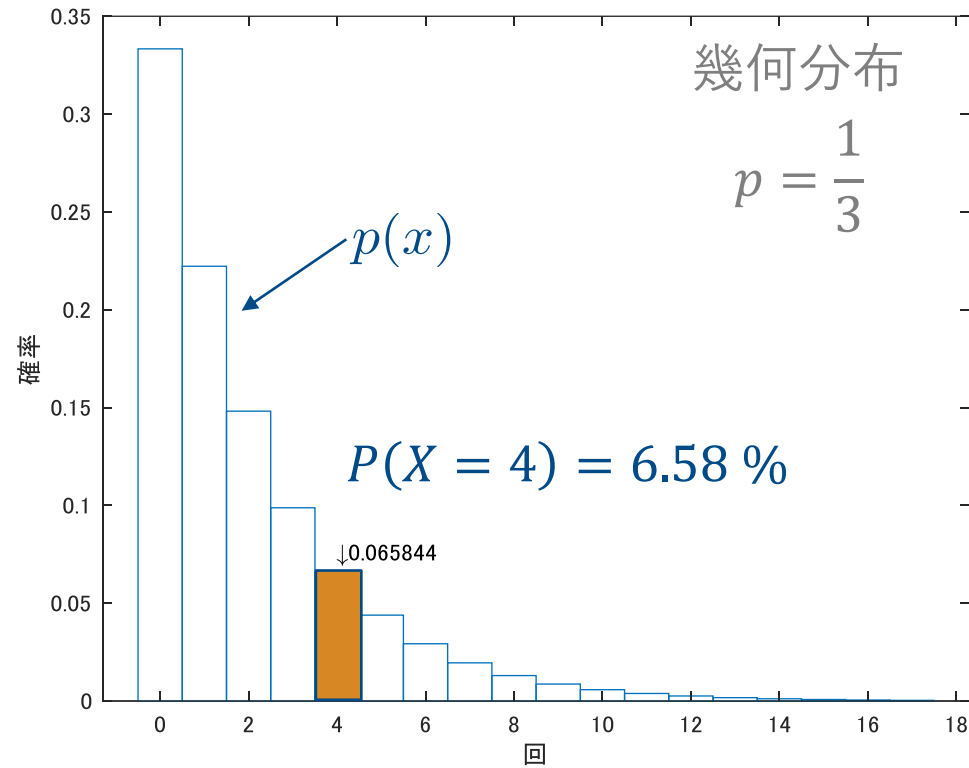


確率変数 X が連続

確率分布 (Cont.)

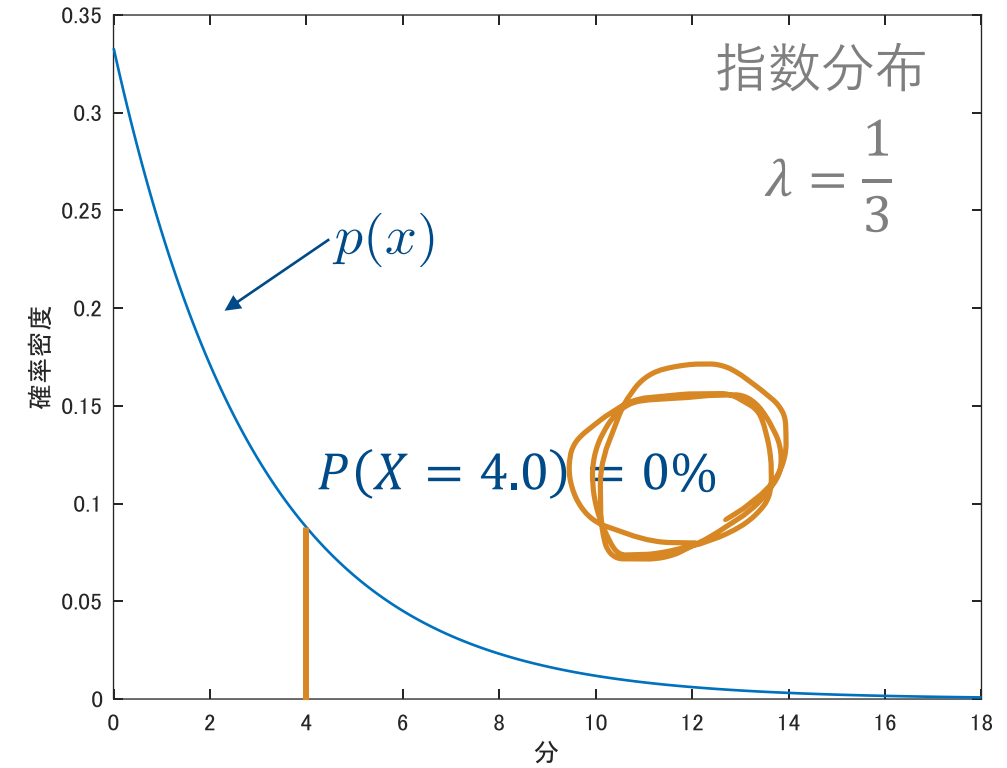
2種類の確率分布

確率 p で起こる事象が初めて観測される
までの試行回数 x の確率



確率質量関数 $p(x)$

単位時間内に平均 λ 回起こる事象の
発生間隔 x の確率



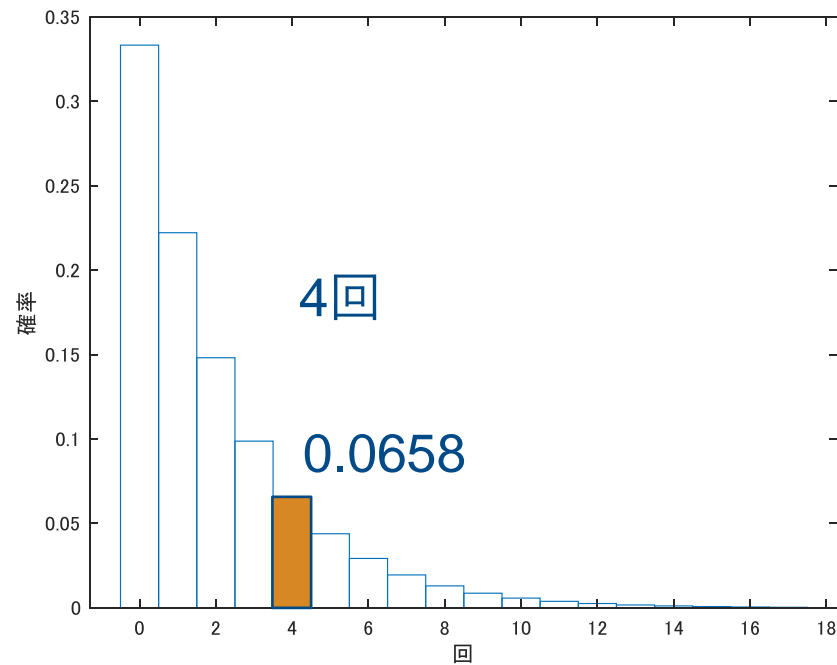
確率密度関数 $p(x)$

確率分布の利用法

確率の計算方法

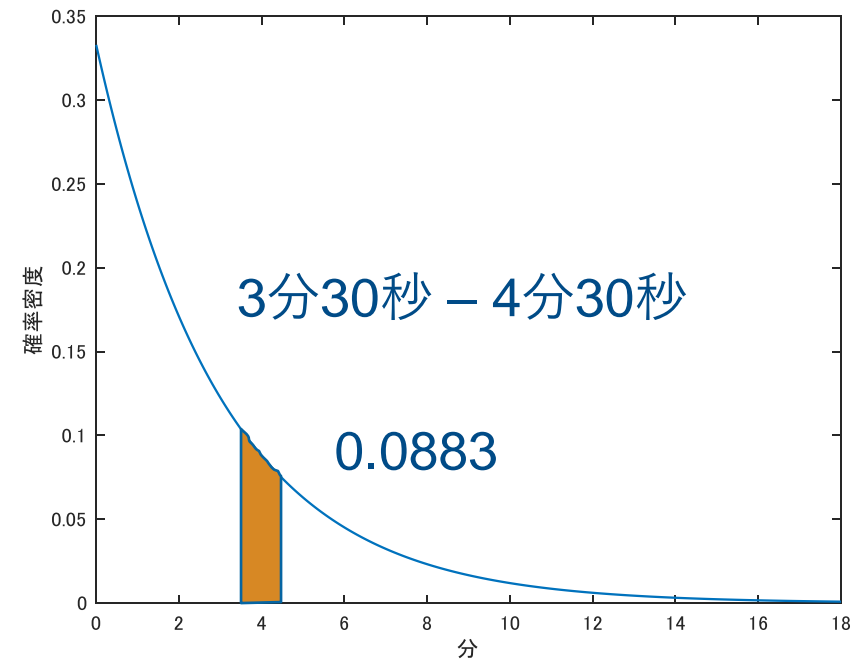
確率質量関数

$$P(X = x) = p(x)$$



確率密度関数

$$P(x_1 \leq X \leq x_2) = \int_{x_1}^{x_2} p(x) dx$$



確率分布の利用法 (Cont.)

期待値・分散の計算方法

確率質量関数

$$E_p[X] = \sum x \cdot p(x) \quad \text{← 期待値}$$

$$V_p[X] = \sum (x - E[X])^2 \cdot p(x) \quad \text{← 分散}$$

確率密度関数

$$E_p[X] = \int x \cdot p(x) dx \quad \text{← 期待値}$$

$$V_p[X] = \int (x - E[X])^2 \cdot p(x) dx \quad \text{← 分散}$$

平均・分散と言えばコレだろ!...

$$\bar{\mu} = \frac{1}{n} \sum_{i=1}^n x_i \quad \sigma^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{\mu})^2$$



確率分布が一様分布の下での計算

確率分布の利用法 (Cont.)

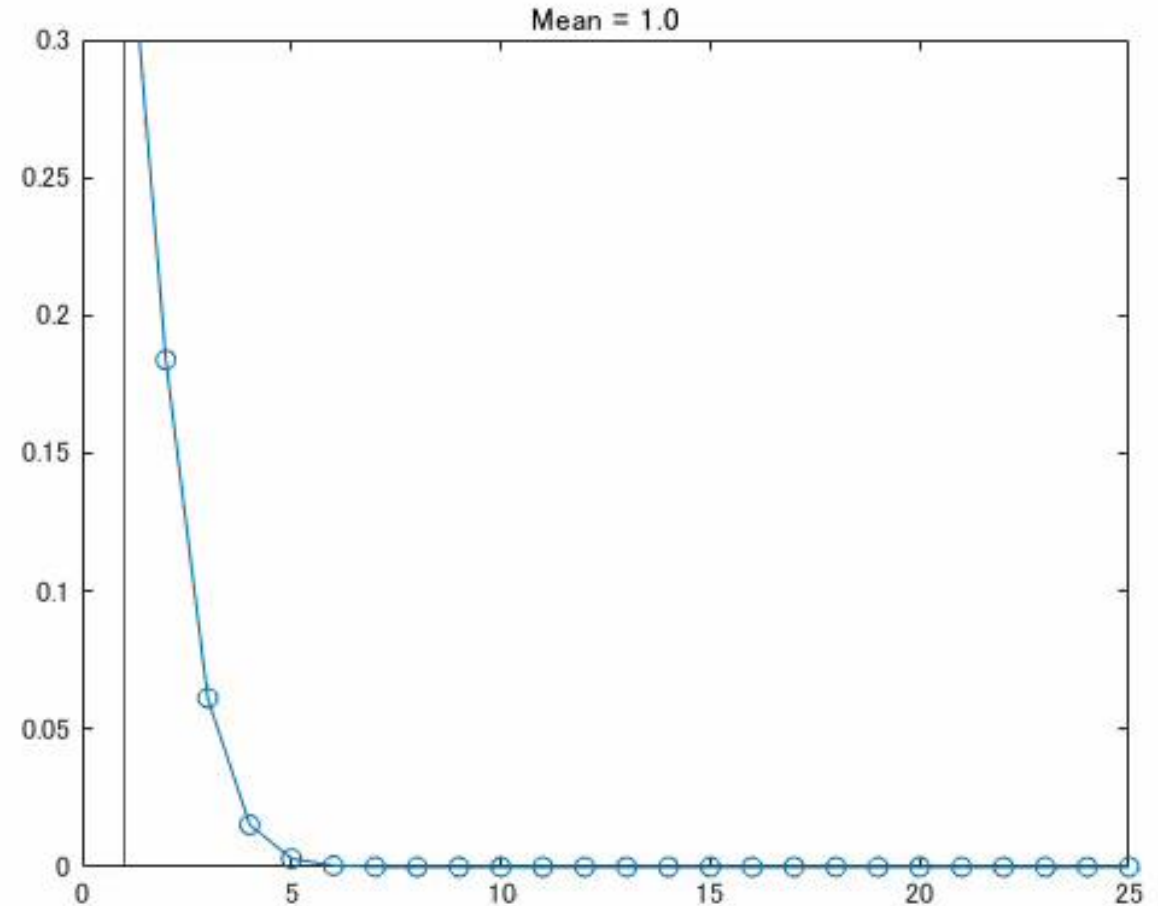
”重みを付ける” イメージ

$$E_p[X] = \sum x \cdot p(x)$$

- ポアソン分布
- 単位時間内に稀な事象が起こる確率
 - (例) 1日当たりの交通事故の件数

$$p(x) = \frac{\lambda^x}{x!} e^{-\lambda}$$

様々な λ を持つポアソン分布



確率分布の利用法 (Cont.)

正規分布の期待値・分散とその性質

平均 μ , 分散 σ^2 の1次元正規分布

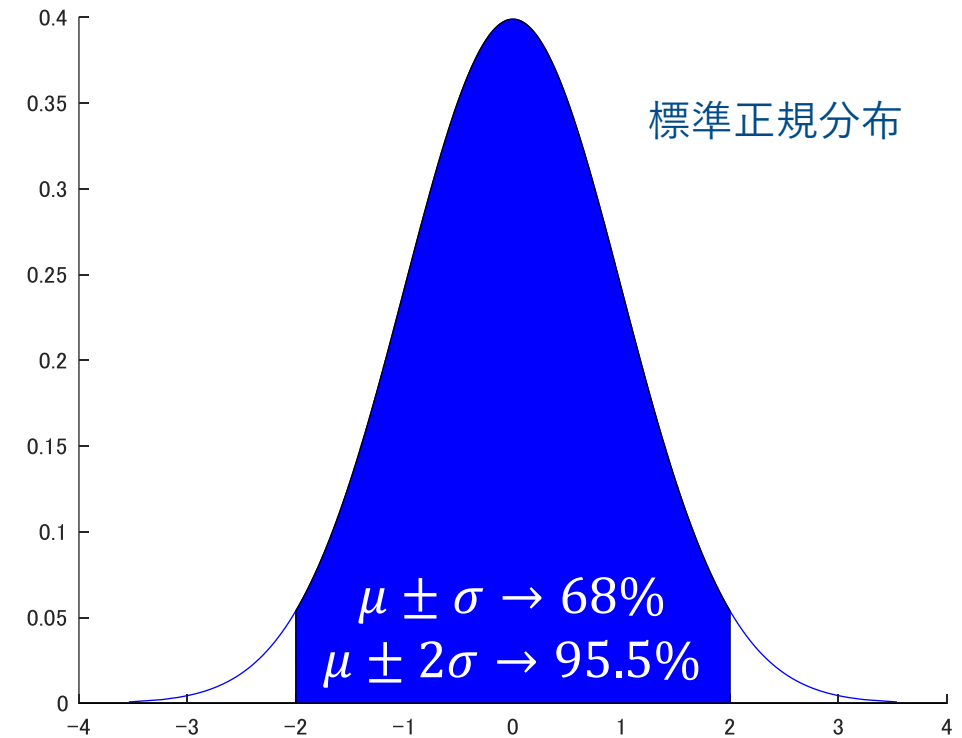
$$X \sim N(\mu, \sigma^2)$$

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

期待値、分散の定義に従って計算すると

$$E[x] = \int_{-\infty}^{\infty} x \cdot p(x) dx = \underline{\underline{\mu}}$$

$$V[x] = E[(x - E[x])^2] = \int_{-\infty}^{\infty} (x - \mu)^2 \cdot p(x) dx = \underline{\underline{\sigma^2}}$$



確率分布の利用法

実践してイメージを持つ

幾何分布

$$X \sim \text{Geo}(p)$$
$$p(x) = p \cdot (1 - p)^{x-1}$$

成功確率 p である独立なベルヌーイ試行を繰り返すとき、初めて成功する間での試行回数 X が従う確率分布

指数分布

$$X \sim \text{Ex}(\lambda)$$
$$p(x) = \lambda e^{-\lambda x} \quad (x \geq 0)$$

ある期間に平均して λ 回生じる現象が、次に生じるまでの期間 X が従う確率分布

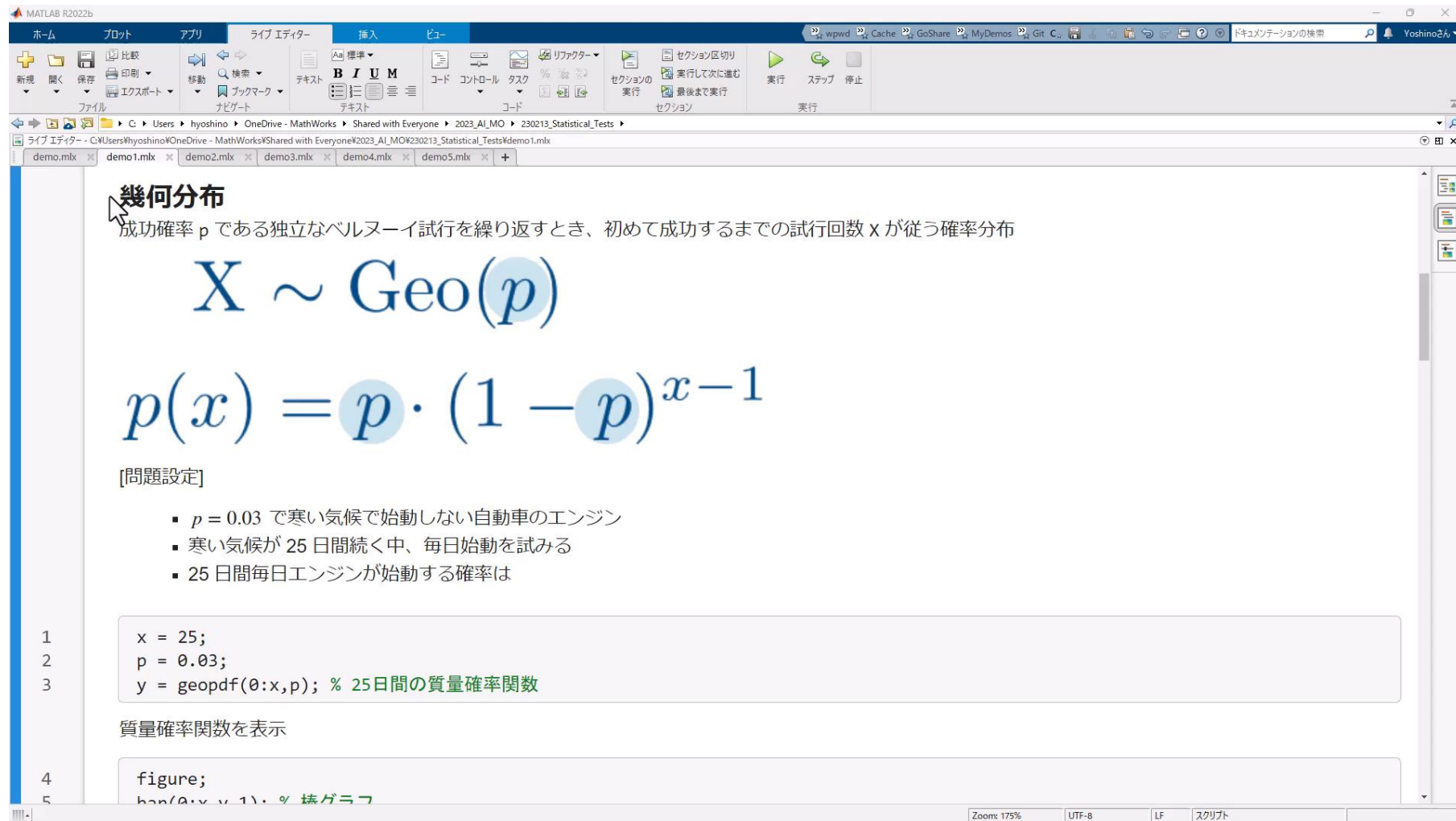
正規分布

$$X \sim N(\mu, \sigma^2)$$
$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left\{ -\frac{(x-\mu)^2}{2\sigma^2} \right\}$$

自然に生じる誤差や個体差は正規分布に従うことが多い

確率分布の利用法

実践してイメージを持つ



The image shows the MATLAB R2022b interface. The main window displays a script titled 'demo1.mlx' with the following content:

幾何分布
成功確率 p である独立なベルヌーイ試行を繰り返すとき、初めて成功するまでの試行回数 x が従う確率分布

$$X \sim \text{Geo}(p)$$
$$p(x) = p \cdot (1 - p)^{x-1}$$

[問題設定]

- $p = 0.03$ で寒い気候で始動しない自動車のエンジン
- 寒い気候が 25 日間続く中、毎日始動を試みる
- 25 日間毎日エンジンが始動する確率は

```
1 x = 25;  
2 p = 0.03;  
3 y = geopdf(0:x,p); % 25日間の質量確率関数
```

質量確率関数を表示

```
4 figure;  
5 bar(0:x-1,y); % 棒グラフ
```

The interface includes a toolbar with various editing and execution tools, and a command window at the bottom.

確率分布の利用法

確率密度分布の選択方法

1. データが離散か連続かを判断
2. データの分布形がどのような形をしているかを観察
3. データの分布形に近い確率密度関数を候補とする
4. 候補となる確率密度関数のパラメータを推定
5. 推定したパラメータを用いて、データと確率密度関数の適合度を検定する
6. 適合度が高い確率密度関数を選択

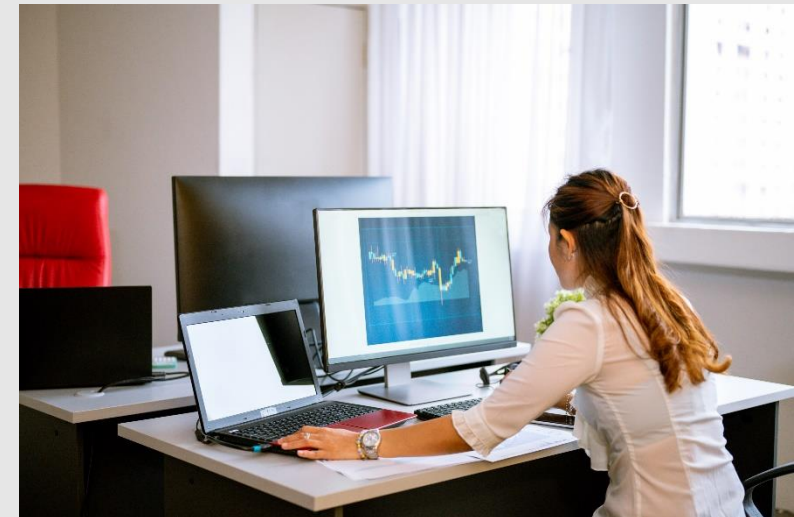
カイ二乗検定適合度検定

E.g.,

- データが連続、分布形が釣り鐘型 >> 正規分布、t 分布 ...
- データが連続、分布形が左右非対称 >> 指数分布、ガンマ分布 ...
- データが離散、成功・失敗のベルヌーイ試行 >> 二項分布、ポアソン分布 ...

Agenda

- 統計学の違い
- 確率変数とは
- 確率分布
- 区間推定
 - 母平均区間推定
 - 母分散区間推定
- 仮説検定
 - 母平均仮説検定
 - T検定 & ウェルチの検定
 - 一元配置分散分析 (One-way ANOVA)
- 周囲に展開する方法
 - Live Editor
 - アプリ化



区間推定

点推定と区間推定



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



標本を元に得られた推定値をどれだけ信じられる?
 $\{x_1, x_2, x_3, \dots, x_n\}$

点推定

偶然、今回得られた

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n X_i$$



統計量 (確率変数) を使い、"このくらい" を表現
 $\{X_1, X_2, X_3 \dots X_n\}$

区間推定

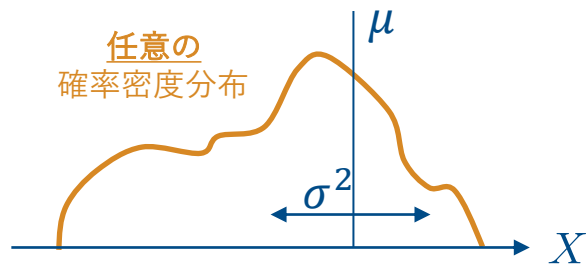
区間推定

中心極限定理 (Central Limit Theorem)

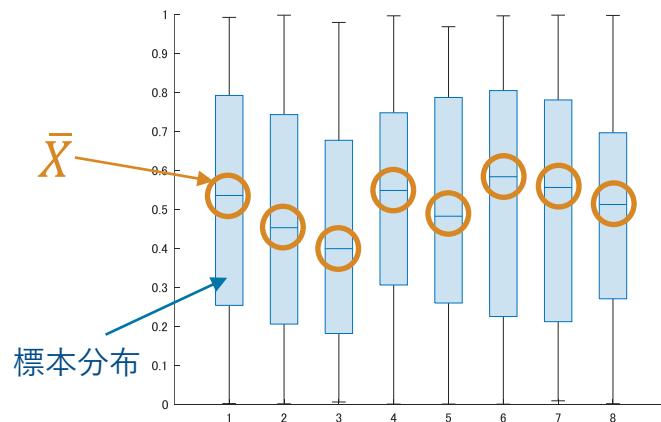
n 大!

Central Limit Theorem

独立に同一分布に従う確率変数の標本平均は正規分布に分布収束する

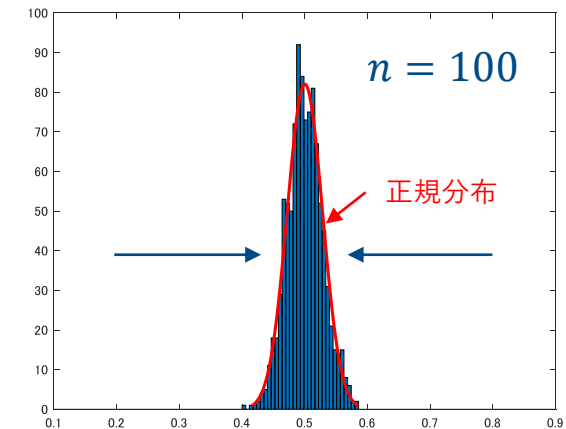
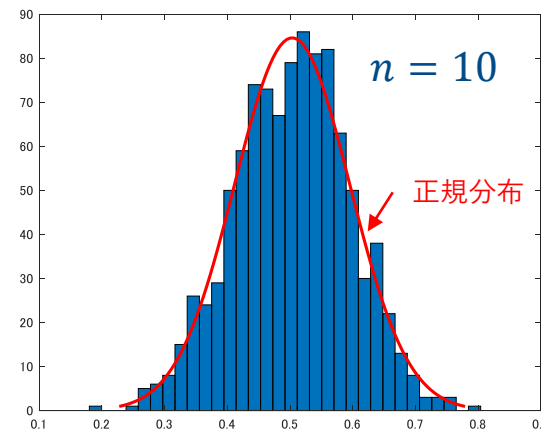


$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i = \frac{1}{n} (X_1 + \cdots + X_n) \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$



$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

のヒストグラム



区間推定

母平均区間推定

$\bar{X} \sim N(\mu, \sigma^2/n)$: 正規分布の線形変換は正規分布

$$\frac{\bar{X} - \mu}{\sqrt{\sigma^2/n}} \sim N(0, 1) \xrightarrow[\text{標本分散で置き換え}]{\sigma^2 \text{ unknown}} \frac{\bar{X} - \mu}{\sqrt{s^2/n}} \sim t(n-1)$$

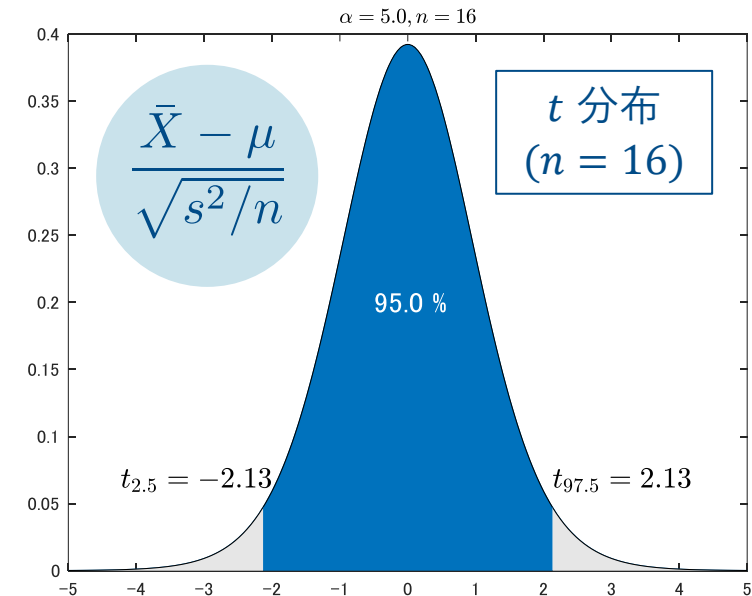
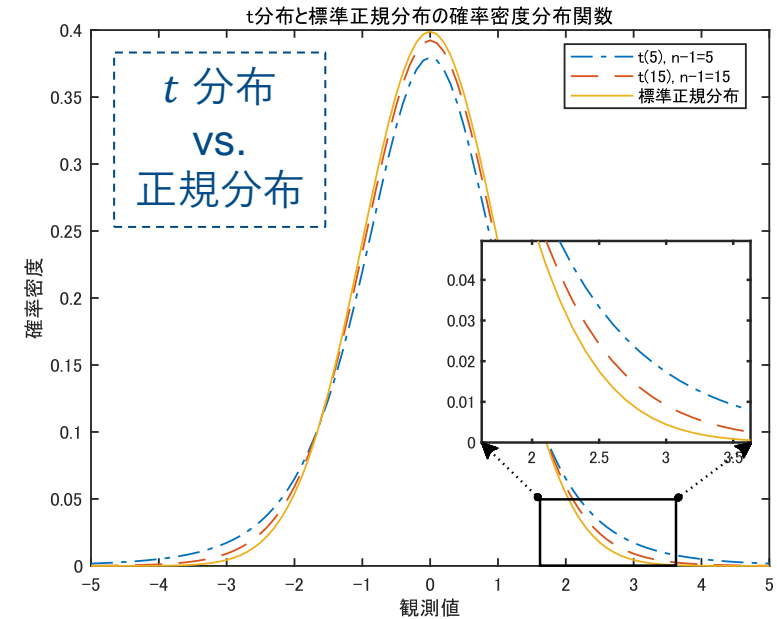
標準正規分布 t 分布 (自由度 $n-1$)

$$P\left(-t_{\alpha/2} \leq \frac{\bar{X} - \mu}{\sqrt{s^2/n}} \leq t_{\alpha/2}\right) = 0.95$$

※ 信頼区間 $\sim \alpha$

母平均区間推定 (但し、これが成立するのは 95%)

$$\Rightarrow \bar{X} - t_{\alpha/2} \cdot \sqrt{\frac{s^2}{n}} \leq \mu \leq \bar{X} + t_{\alpha/2} \cdot \sqrt{\frac{s^2}{n}}$$



区間推定 (Cont.)

母平均区間推定

※元の分布は正規分布だと仮定

ある LED 電球から 12 個を無作為抽出し、その耐久時間を測定:
 {1330 1060 1200 1190 1500 1430 1150 1200 1490 1020 1410 1390} 平均値を見積もれ

$$\bar{x} = \frac{1}{n} \sum_i x_i = \frac{1}{12} (1330 + 1060 + \cdots + 1390) = 1280.3$$

$$s^2 = \frac{1}{n-1} \sum_i (x_i - \bar{x})^2 \longrightarrow E[s^2] = \sigma^2$$

不偏推定量
(unbiased estimator)

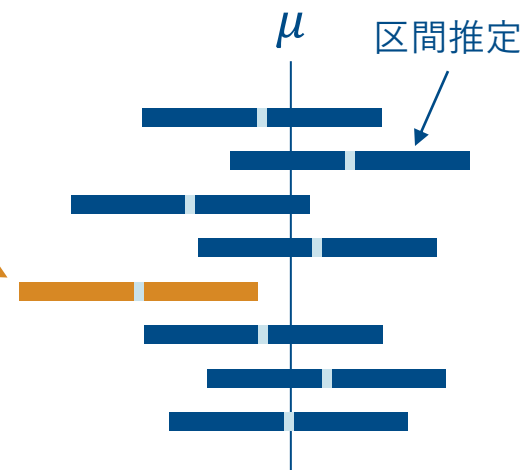
$$= \frac{1}{11} \{ (1330 - 1280.3)^2 + (1060 - 1280.3)^2 + \cdots \} = 27299$$

母平均区間推定の公式

$$\bar{X} - t_{\alpha/2} \cdot \sqrt{\frac{s^2}{n}} \leq \mu \leq \bar{X} + t_{\alpha/2} \cdot \sqrt{\frac{s^2}{n}}$$

1175.9 1385.8

α の割合で
外れる



区間推定 (Cont.)

母分散区間推定 – 参考

標本分散と母分散の比がカイ二乗分布に従う性質に注目

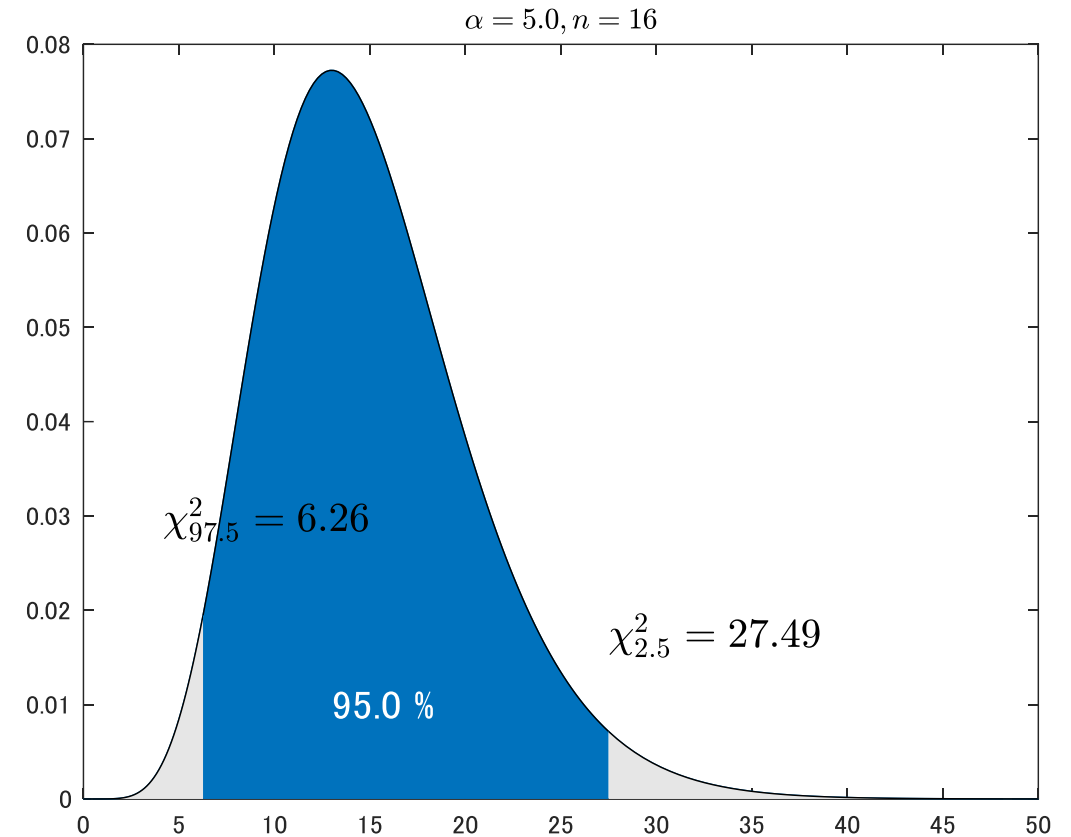
$$\frac{(n-1)s^2}{\sigma^2} \sim \chi^2(n-1) \quad \text{※ 自由度 } (n-1)$$

母分散区間推定の公式

$$\frac{(n-1)s^2}{\chi_{\alpha/2}^2(n-1)} \leq \sigma^2 \leq \frac{(n-1)s^2}{\chi_{1-\alpha/2}^2(n-1)}$$

$\chi_{\alpha/2}^2(n-1)$: 自由度 $n-1$ のカイ二乗分布の上側 $\alpha/2$ 点

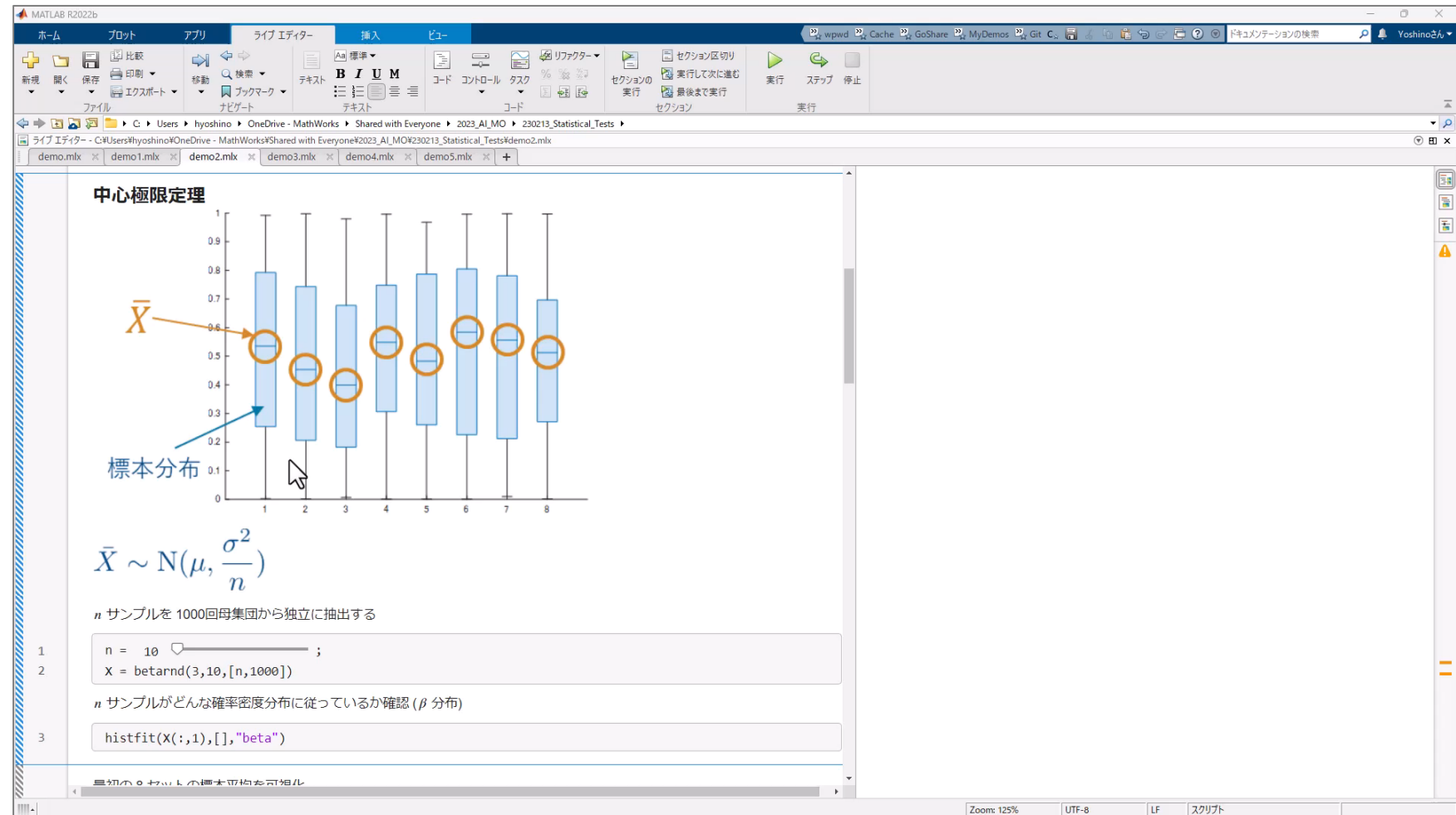
MATLAB で実行 `[h,p,ci,stats] = vartest(x,v)`



自由度 15 のカイ二乗分 $\chi^2(15)$

区間推定

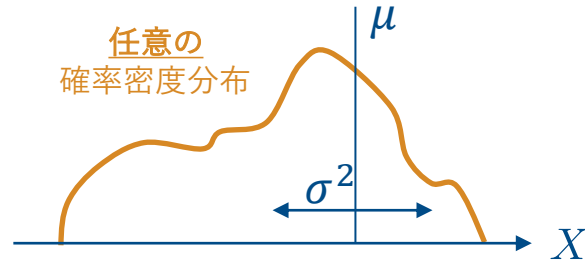
- 中心極限定理
- 母平均区間推定
- 母分散区間推定



区間推定 (Cont.)

母平均区間推定 – まとめ (厳密性無視)

標本は大きめ！



$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

良く分からない分布からの
標本平均が正規分布に分布収束する

(標本平均 – 母平均) / 標準誤差 の比!

$$\frac{\bar{X} - \mu}{\sqrt{\sigma^2/n}} \sim N(0, 1) \xrightarrow[\text{標本分散で置き換え}]{\sigma^2 \text{ unknown}} \frac{\bar{X} - \mu}{\sqrt{s^2/n}} \sim t(n-1)$$

標準正規分布 t 分布 (自由度 $n-1$)

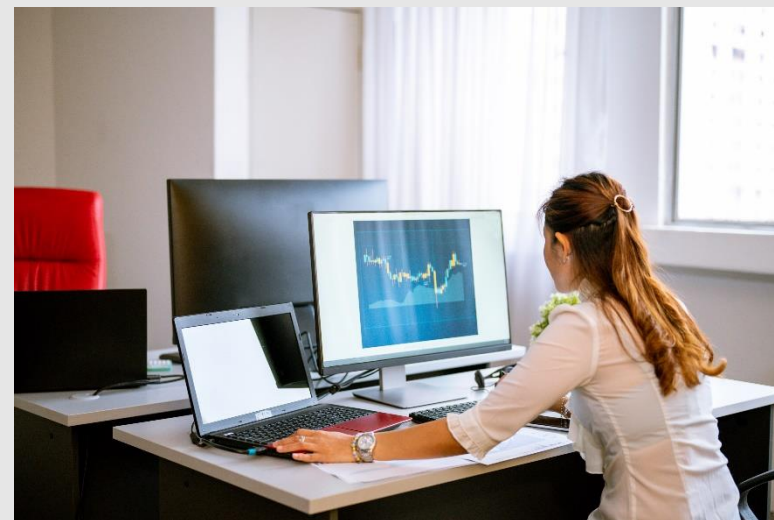
母分散を知らないので
標本分散で代用 → t 分布に従う

$$\bar{X} - t_{\alpha/2} \cdot \sqrt{\frac{s^2}{n}} \leq \mu \leq \bar{X} + t_{\alpha/2} \cdot \sqrt{\frac{s^2}{n}}$$

t 分布の性質を利用して
母平均を評価 (区間推定)
“信頼区間 **% でこれです”

Agenda

- 統計学の違い
- 確率変数とは
- 確率分布
- 区間推定
 - 母平均区間推定
 - 母分散区間推定
- 仮説検定
 - 母平均仮説検定
 - T検定 & ウェルチの検定
 - 一元配置分散分析 (One-way ANOVA)
- 周囲に展開する方法
 - Live Editor
 - アプリ化



仮説検定とは

第一種の過誤

- 帰無仮説 H_0 (e.g., $\mu = 0$) に対して、それが正しいか否かを統計学的に検証
- 主張したいのは 対立仮説 H_1 (e.g., $\mu \neq 0$)

判断 \ 真実	H_0 が正しい	H_1 が正しい
H_0 を棄却 / H_1 を採択	Type I error ・ 第一種の過誤 α (有意水準) でコントロール	✓
H_0 を棄却しない	✓	Type II error ・ 第二種の過誤 β でコントロール

仮説検定

統計量と有意水準のイメージ

統計量 (確率変数で作られる)

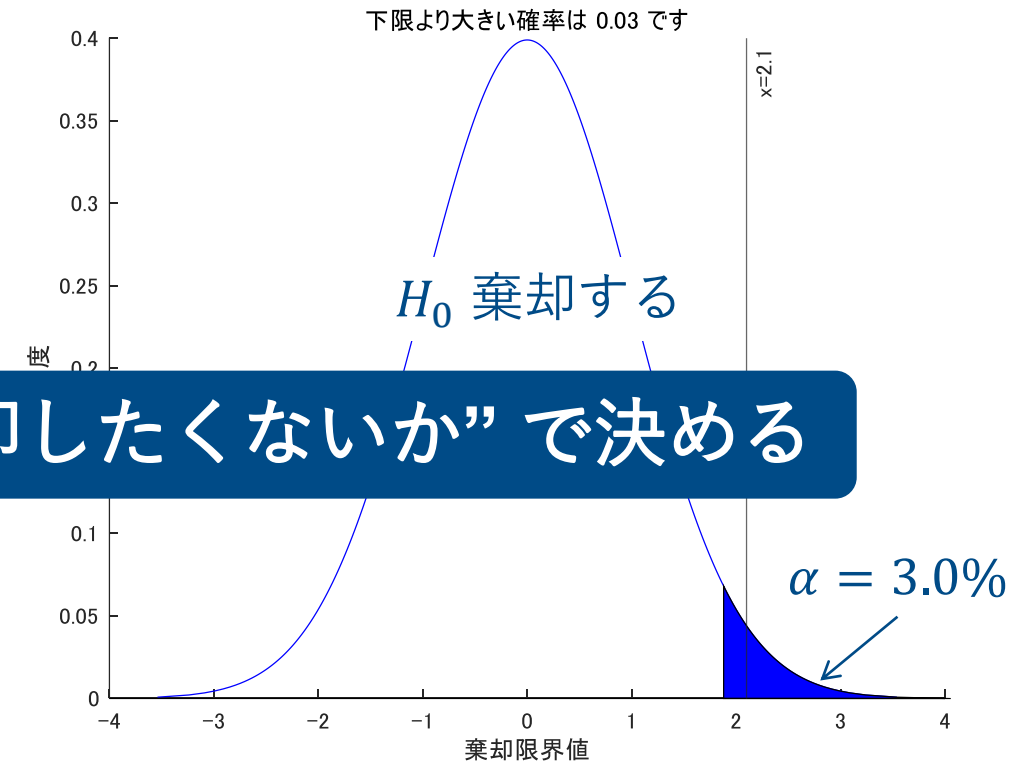
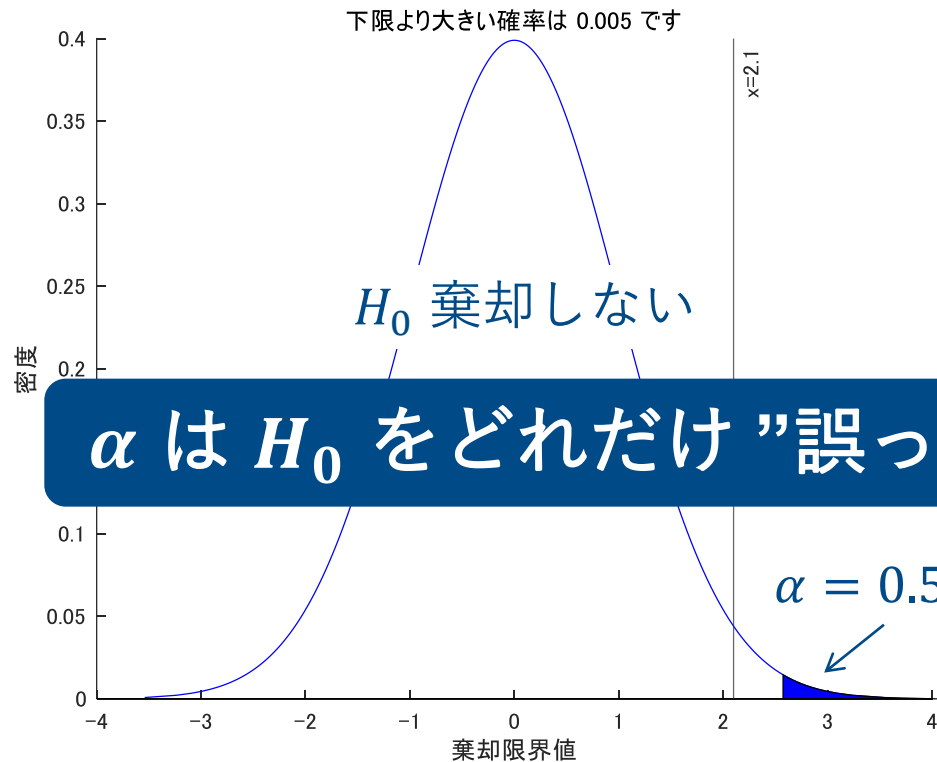
$$H_0 : X \sim N(0, 1^2)$$

平均 分散

α の解釈

- “珍しいライン”の定量化
- 正しい H_0 を棄却してしまう確率

統計値 (サンプル) $x = 2.1$ の場合



α は H_0 をどれだけ “誤って棄却したくないか” で決める

仮説検定

何がオイシイの？

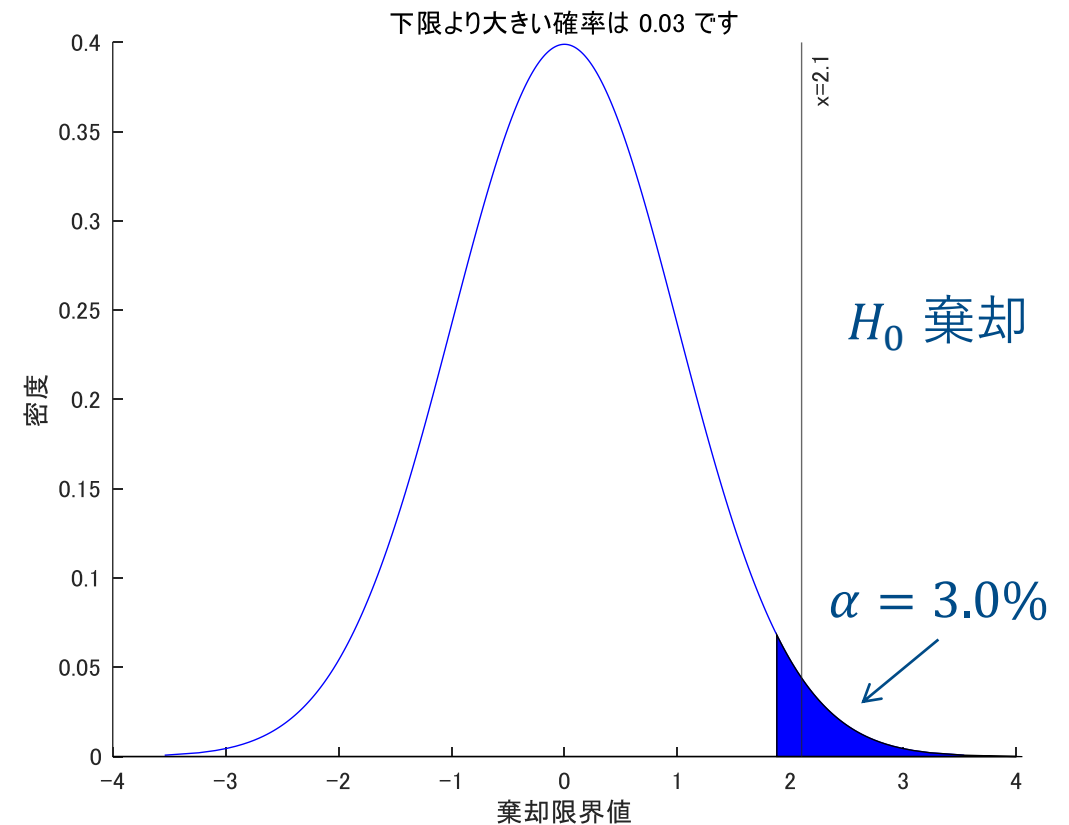
統計量 (確率変数で作られる)

$$H_0 : X \sim N(0, 1^2)$$

平均 分散

- H_0 を曖昧さを持って棄却
- “同じことを何度も繰り返した場合、
3% は間違った判断しますが、今回は棄却”

➡ 下した判断の自信を定量化



仮説検定

母平均仮説検定- 異常検知

直径の平均が 3.70 cm, 標準偏差 0.025 cm の部品を製造している。この部品 10 個を無作為抽出すると平均が 3.72 cmであった。機械は正常に動作しているか? (有意水準 5%)

- 帰無仮説 (仮定する仮説) $H_0: \mu = 3.70$



- 対立仮説 (主張した仮説) $H_1: \mu \neq 3.70$

➡ H_0 を真としてデータを見た際の矛盾を導く
サンプル平均から検定統計量を作成!

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

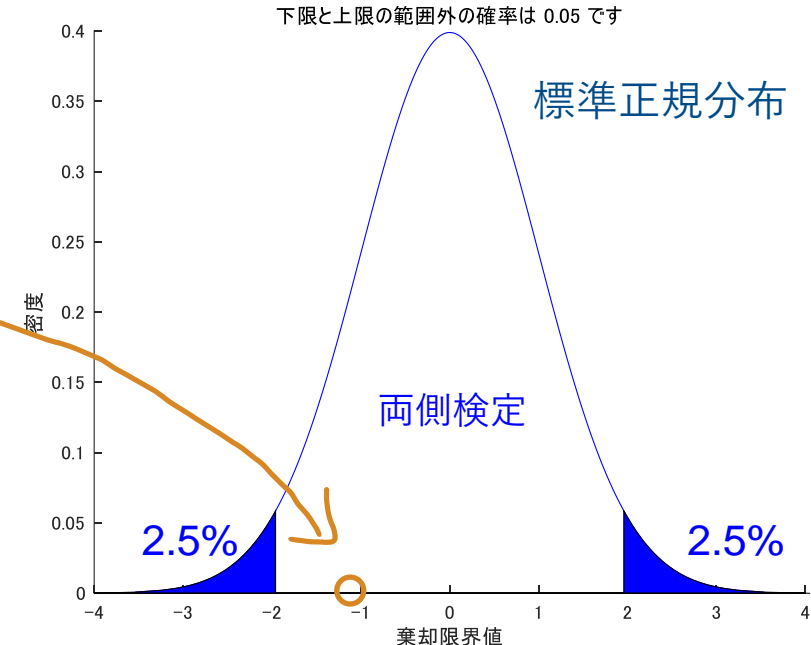
※ 元の分布は正規分布を仮定



$$\frac{\bar{X} - \mu}{\sqrt{\sigma^2/n}} \sim N(0, 1)$$

標準正規分布

今回の検定統計量



仮説検定 (Cont.)

母平均仮説検定- 異常検知

直径の平均が 3.70 cm, 標準偏差 0.025 cm の部品を製造している。この部品 10 個を無作為抽出すると平均が 3.72 cmであった。機械は正常に動作しているか? (有意水準 5%)

```
alpha = 0.05;
[h,p,ci,zval] = ztest(repmat(3.72, 3, 1), ...
    3.7, 0.025, "Tail", "both", "Alpha", alpha)
```

h = 1

p = 0.0114

ci = 2x1

3.7045

3.7355

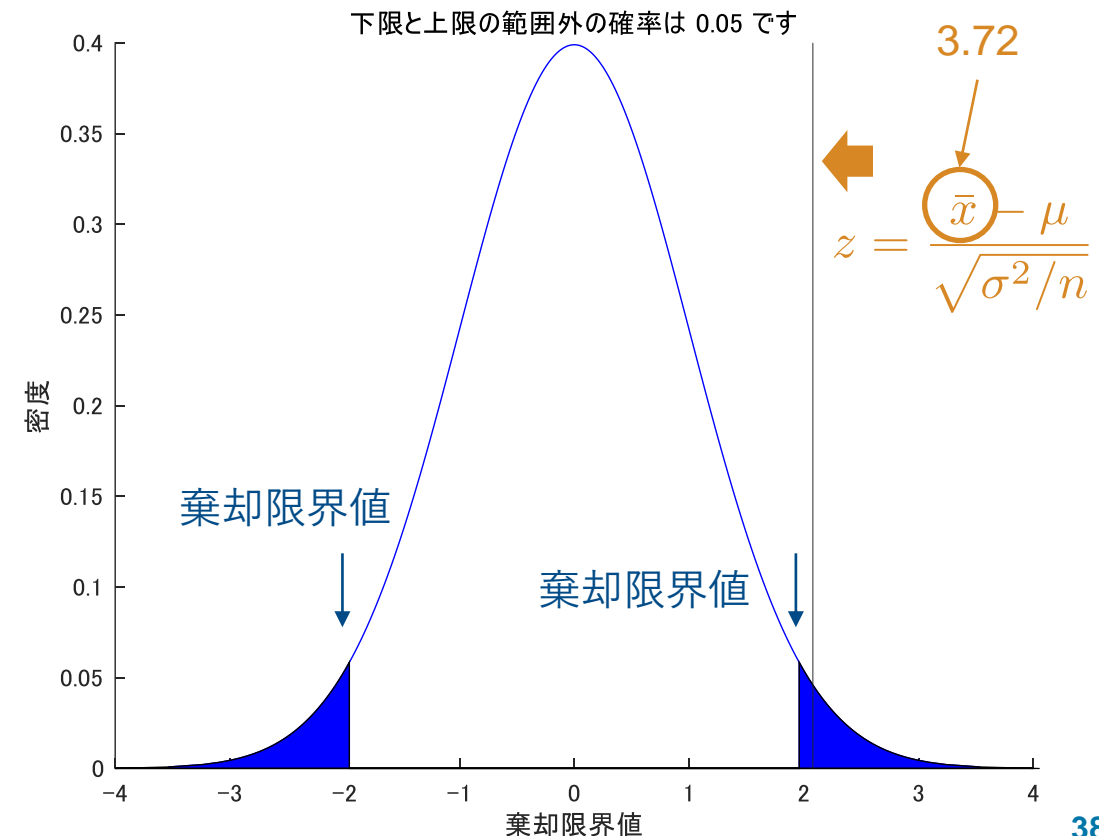
zval = 2.5298

h: 棄却 (1), 棄却しない (0)

p: p値

ci: 平均値の信頼区間

zval: z 値



仮説検定

母集団の性質について標本を用いて判断・検証をする 3ステップ

1. H_0 (帰無仮説) と H_1 (対立仮説) を母集団に対して設定

1. H_0, H_1

2. 帰無仮説を真とした検定統計量を用意,
有意水準 α (棄却域)を決める

$$2. \frac{\bar{X} - \mu}{\sqrt{\sigma^2/n}} \sim N(0, 1)$$

3. サンプルから検定統計値を計算

1. 棄却域に検定統計値が入れば、帰無仮説を棄却
 (“ H_1 は統計的に有意” という)

$$3. z = \frac{\bar{x} - \mu}{\sqrt{\sigma^2/n}}$$

2. 帰無仮説が棄却されない→

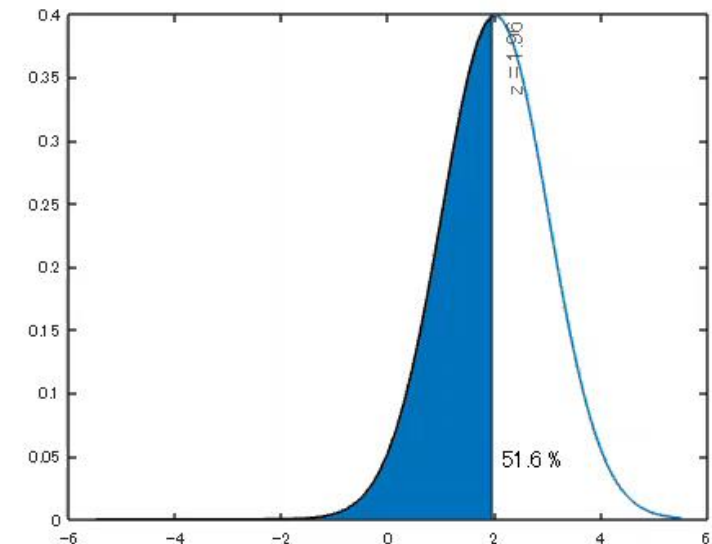
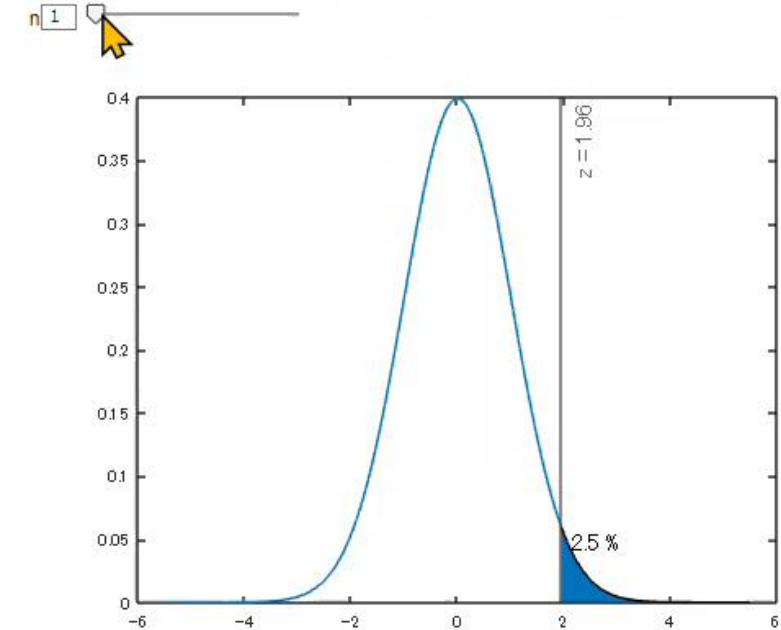
“ H_1 対立仮説が正しいとは言えない” 程度

仮説検定

補足: n (サンプル数) を増やすとどうなるか?

- $H_0: \mu = \mu_0 \quad \frac{\bar{X} - \mu_0}{\sqrt{\sigma^2/n}} \sim N(0, 1)$
- $H_1: \mu = \mu_1 \quad \frac{\bar{X} - \mu_0}{\sqrt{\sigma^2/n}} \sim N\left(\frac{\mu_1 - \mu_0}{\sqrt{\sigma^2/n}}, 1\right)$

正規分布: n が大きくなっていくとどうなるか?



判断 \ 真実	真実	
	H_0 が正しい	H_1 が正しい
H_0 を棄却	Type I error	
H_0 を棄却しない		Type II error

検出力の UP

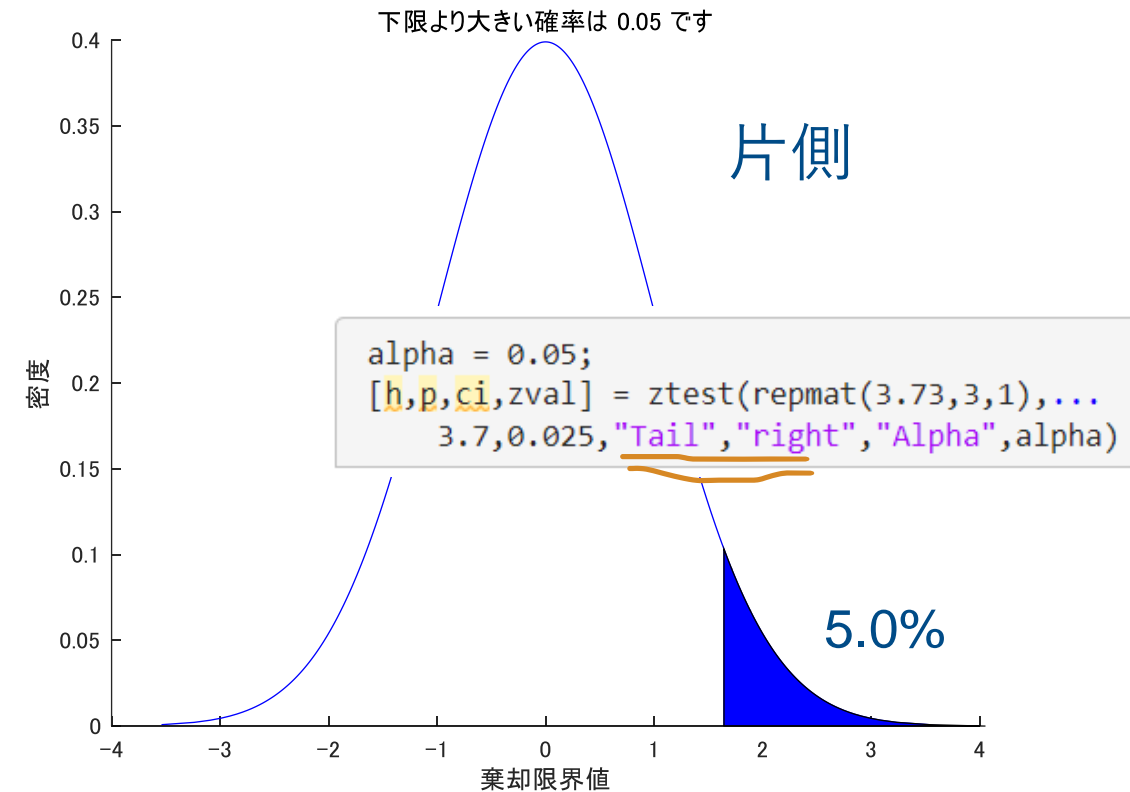
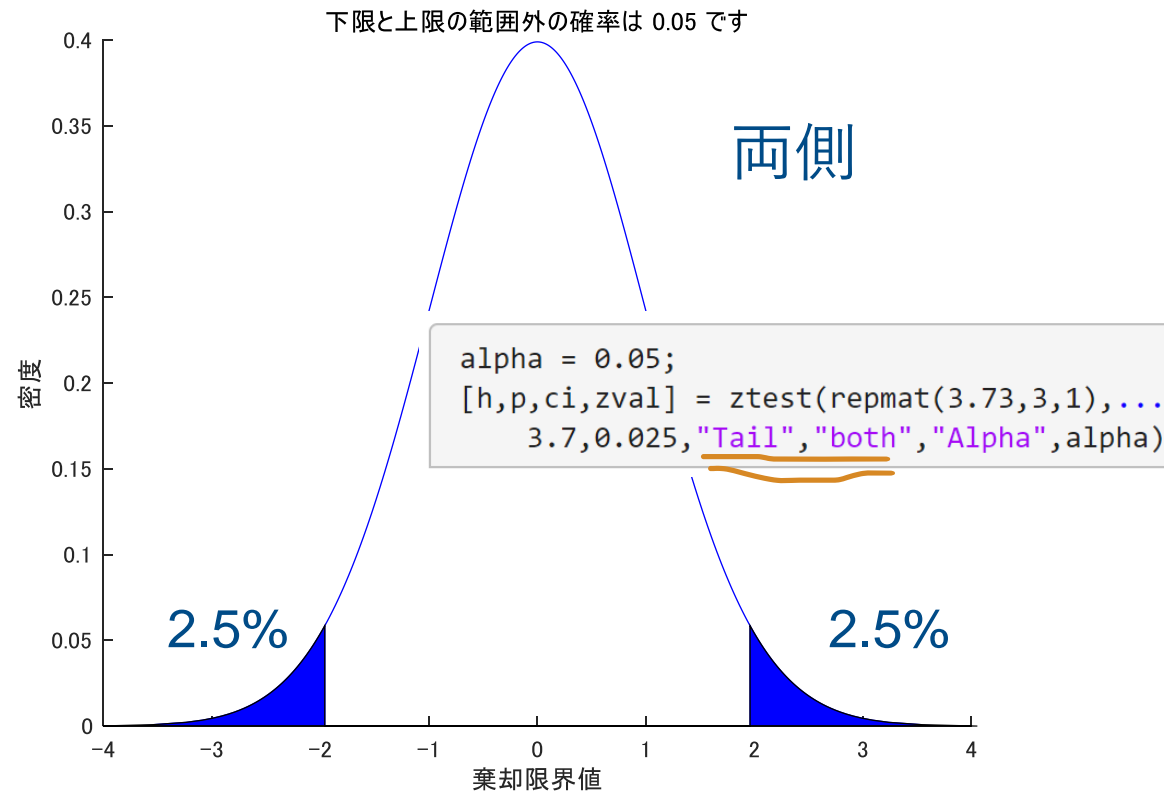
仮説検定

補足: 両側検定・片側検定

“ H_1 をどう考えているか”で決める

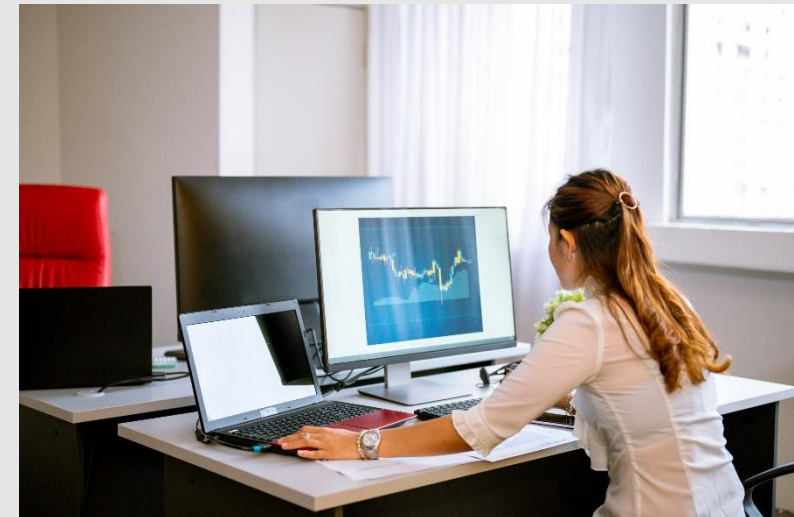
- $H_1: \mu \neq \mu_0$

- $H_1: \mu > \mu_0$



Agenda

- 統計学の違い
- 確率変数とは
- 確率分布
- 区間推定
 - 母平均区間推定
 - 母分散区間推定
- 仮説検定
 - 母平均仮説検定
 - T検定 & ウェルチの検定
 - 一元配置分散分析 (One-way ANOVA)
- 周囲に展開する方法
 - Live Editor
 - アプリ化



仮説検定

等平均仮説 (分散未知) 検定 – いわゆる T 検定



耐熱温度 A 工法 (9 サンプル) = [241 250 230 198 222 245 206 231 249]
 B 工法 (7 サンプル) = [252 233 259 257 224 220 241]

$$X_A \sim N(\mu_A, \sigma_A^2)$$

$$X_B \sim N(\mu_B, \sigma_B^2)$$

A, B の分散は不明だが、等しいとする。A 工法、B 工法に差はないか？

STEP.1 仮説を設定

- 帰無仮説 (仮定する仮説) $H_0: \mu_A = \mu_B$
- 対立仮説 (主張した仮説) $H_1: \mu_A \neq \mu_B$

STEP.2 統計検定量を用意、有意水準を決める

$$T = \frac{\bar{X}_A - \bar{X}_B - (\mu_A - \mu_B)}{\sqrt{\left(\frac{1}{n_A} + \frac{1}{n_B}\right) \cdot \frac{(n_A-1)s_A^2 + (n_B-1)s_B^2}{(n_A-1) + (n_B-1)}}} \sim t(n_A + n_B - 2)$$

平均値の差の標準誤差

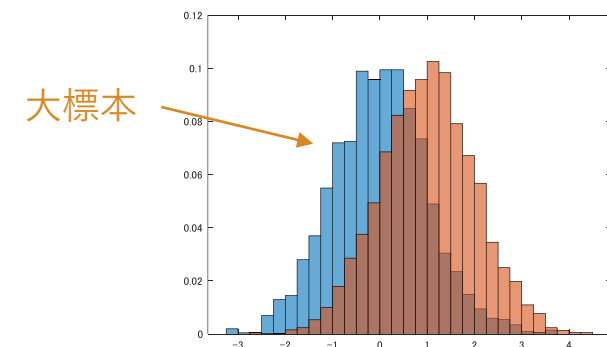
t 分布に従う

STEP.3 統計検定値を計算

$$\bar{x}_A, \bar{x}_B, s_A^2, s_B^2 \rightarrow T$$

母分散未知・小標本

t 分布の活用



仮説検定 (Cont.)

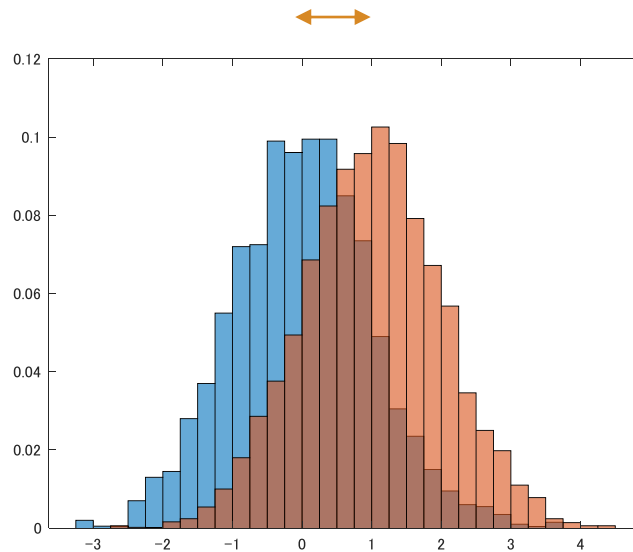
等平均仮説 (分散未知) 検定 - T 値と特徴選択

$$\bar{x}_A, \bar{x}_B, s_A^2, s_B^2$$

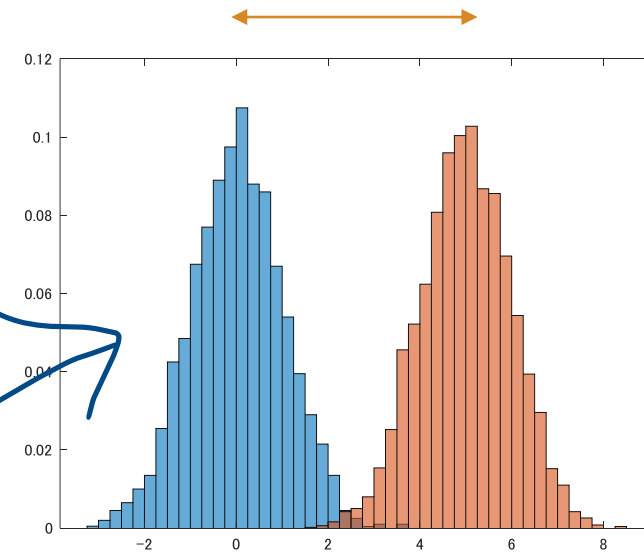


$$T = \frac{\bar{X}_A - \bar{X}_B - (\mu_A - \mu_B)}{\sqrt{\left(\frac{1}{n_A} + \frac{1}{n_B}\right) \cdot \frac{(n_A-1)s_A^2 + (n_B-1)s_B^2}{(n_A-1) + (n_B-1)}}} \sim t(n_A + n_B - 2)$$

t 分布に従う



- $|T|$ 小さめ
- H_0 棄却されず



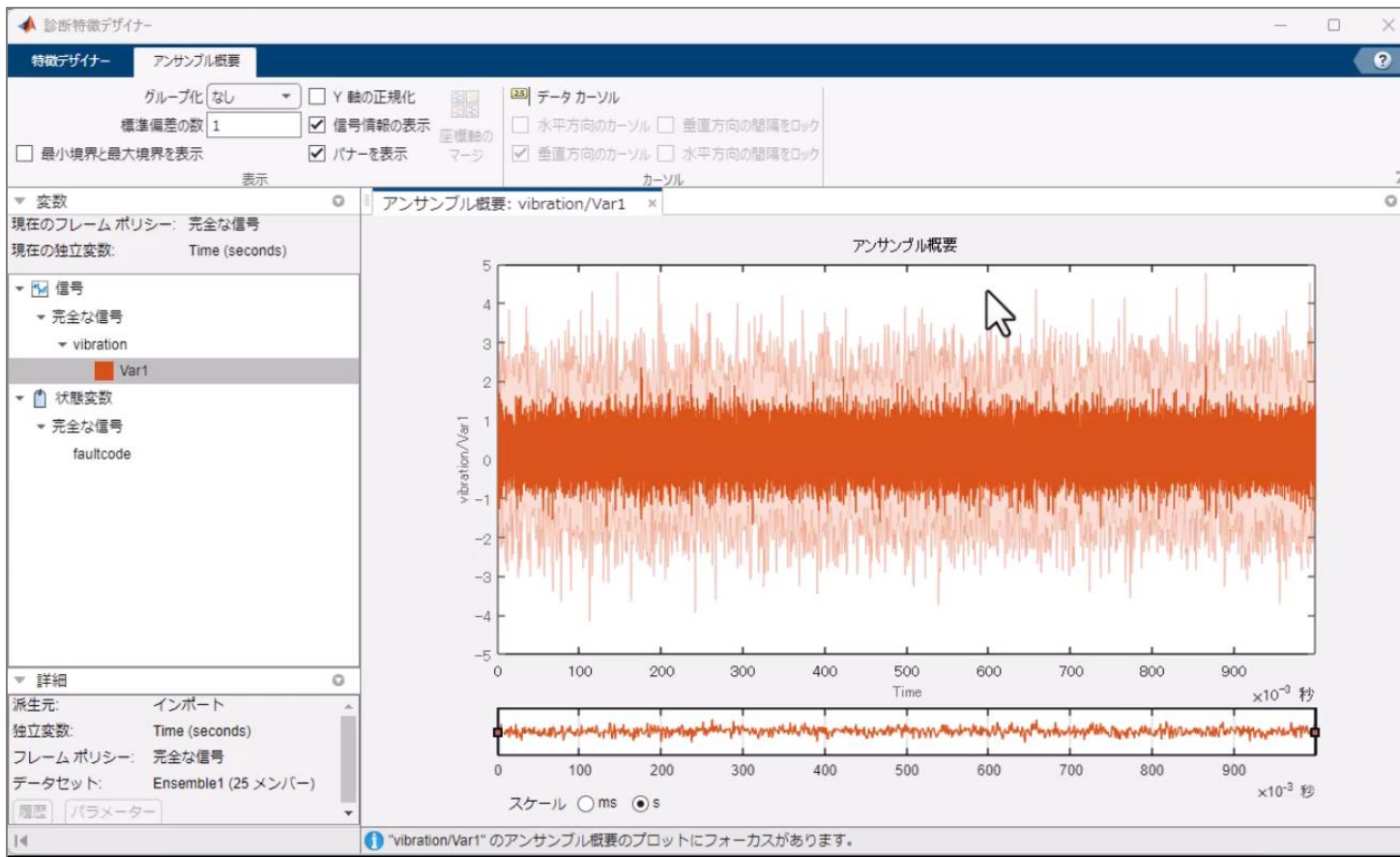
- $|T|$ 大きめ
- H_0 棄却

仮に正常・異常や良品・不良品を見分けるとしたらどちらの特徴量が好ましい？

仮説検定 (Cont.)

等平均仮説 (分散未知) 検定 – 診断特徴デザイナーアプリ

特徴量のランキング付けに t 値を利用 (13 個の特徴量を作成して評価)



MathWorks®

ライブイベント

ラベリング工数大幅削減！アプリで楽するデータラベリングと前処理 - AI モデル検討の前に

開始時間	終了時間
2023年3月23日, 14:00	2023年3月23日, 14:45

概要

機械学習のプロジェクトにおいて前処理が重要なステップであることには異論の余地はなく、学習データの質が悪い場合は、どれだけ有能なアルゴリズムを使用したとしても実用的な精度は期待できません。プロジェクトの早い段階で正しく準備された

- 診断特徴デザイナーアプリの使い方
- ノーコード異常検知
- 特徴量を90個一気に作成!

[詳細はこちらから](#)

仮説検定

等平均仮説 (分散未知) 検定 – 参考



耐熱温度 A 工法 (9 サンプル) = [241 250 230 198 222 245 206 231 249]
 B 工法 (7 サンプル) = [252 233 259 257 224 220 241]

$$X_A \sim N(\mu_A, \sigma_A^2)$$

$$X_B \sim N(\mu_B, \sigma_B^2)$$

A, B の分散は不明かつ、等しくないとする。A 工法、B 工法に差はないか?

ウェルチの検定

$$T = \frac{\bar{X}_A - \bar{X}_B}{\sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}} \sim t(m) \quad \text{where, } m = \frac{\left(\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}\right)^2}{\frac{s_A^4}{n_A^2(n_A-1)} + \frac{s_B^4}{n_B^2(n_B-1)}}$$

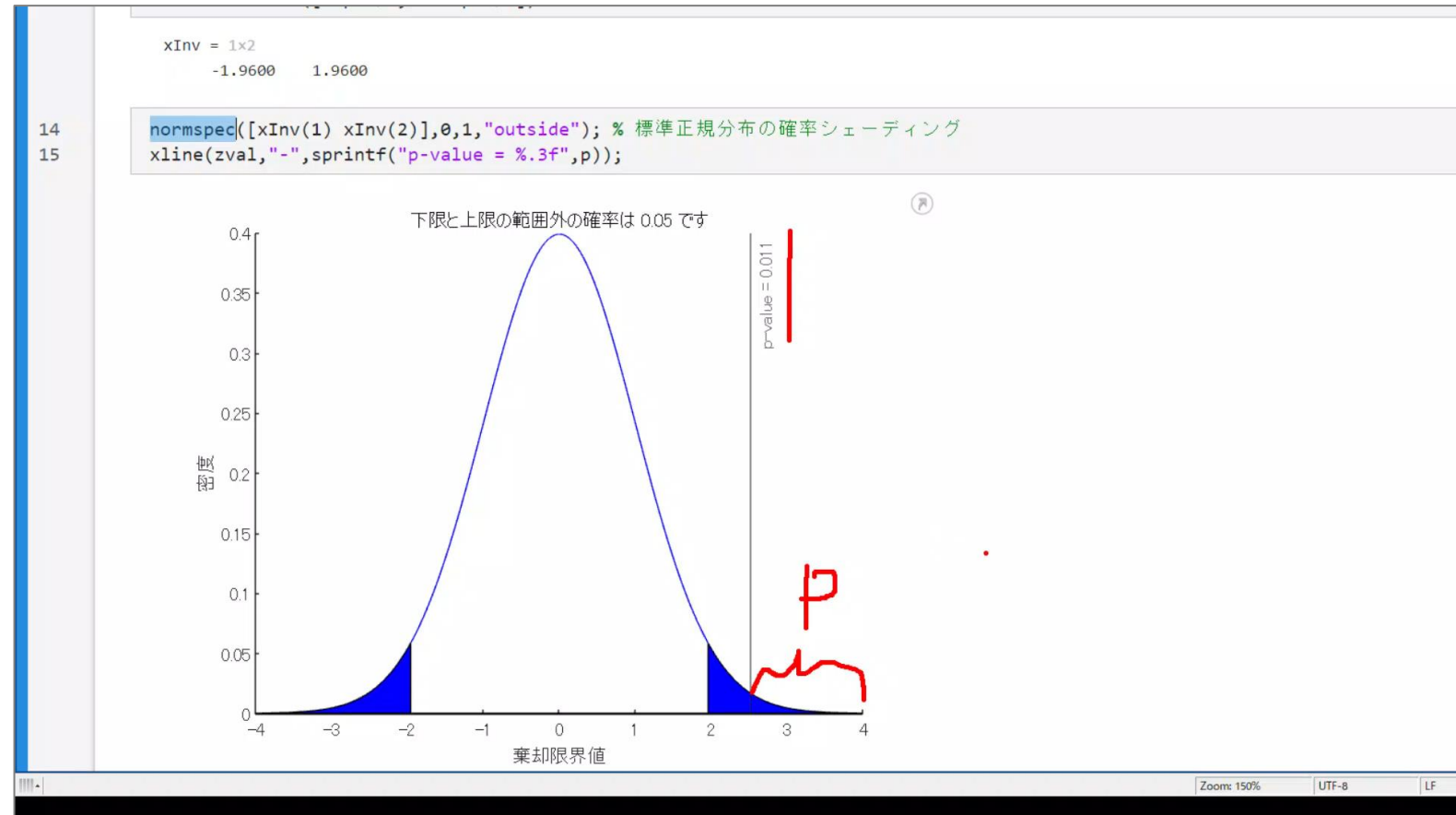
検定統計量 自由度

`[h,p] = ttest2(x,y,'Vartype','unequal')`

MATLAB syntax はほぼ一緒

仮説検定

- 母平均仮説検定
 - 異常検知
- T 検定 & ウェルチの検定

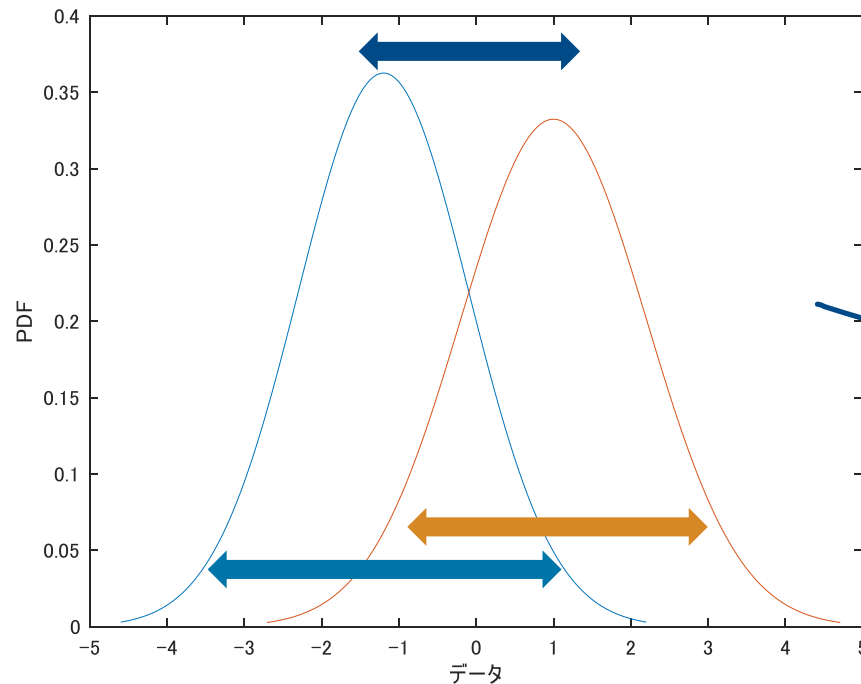


仮説検定

等平均仮説検定 – まとめ (厳密性無視)

$$T = \frac{\bar{X}_A - \bar{X}_B}{\sqrt{\left(\frac{1}{n_A} + \frac{1}{n_B}\right) \cdot \frac{(n_A-1)s_A^2 + (n_B-1)s_B^2}{(n_A-1) + (n_B-1)}}} \sim t(n_A + n_B - 2)$$

$$T = \frac{\text{平均値の差}}{\text{平均値の差の標準誤差}}$$



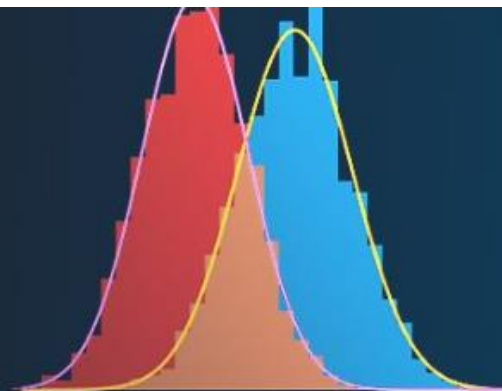
$$T = \frac{\text{平均値の差}}{\text{データのバラつき} + \text{データのバラつき}}$$

平均値の差がデータのバラつきに対して
*** 倍を超えていたら、帰無仮説を棄却する

MATLAB による統計解析

次回開催 7/20-21

※アカデミック割引対象



コースの詳細

この2日間コースの受講により、MATLAB と Statistics and Machine Learning Toolbox を使用して統計解析を行うために必要な関数の使い方を体系的に幅広く習得できます。基本的かつ重要度が高いと考えられる一連の統計的手法 (分布近似、仮説検定、分散分析、回帰、次元削減など) を題材としています。例題と演習ではデータのインポートと整理を行った後、各手法を実行して MATLAB と Statistics and Machine Learning Toolbox で提供される機能の使い方を学べます。

- データのインポートと整理
- データの調査
- 分布
- 仮説検定
- 分散分析
- 回帰
- 多次元データの取り扱い
- 乱数とシミュレーション



レベル: 中級

必要条件:

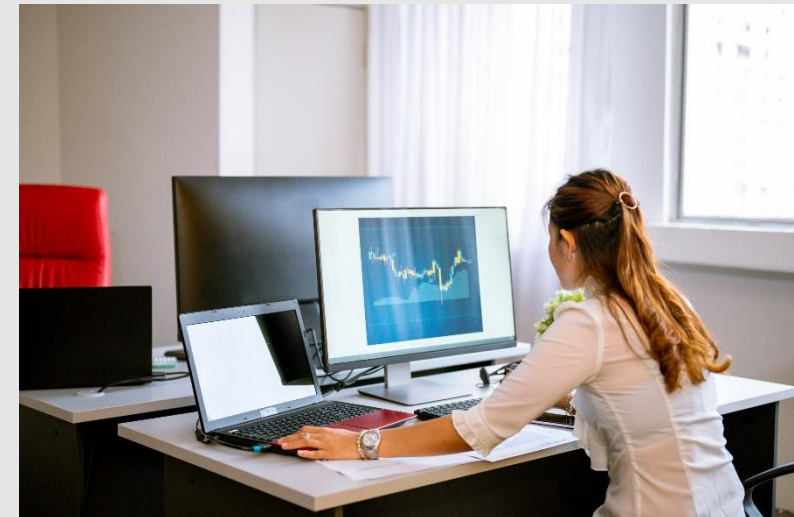
- MATLAB 基礎 コースを受講済み、または同等の MATLAB 使用経験 (特にベクトル、行列、テーブル操作) があり、基本的な統計の知識をお持ちの方

期間: 2 日間

[MATLAB による統計解析](#)

Agenda

- 統計学の違い
- 確率変数とは
- 確率分布
- 区間推定
 - 母平均区間推定
 - 母分散区間推定
- 仮説検定
 - 母平均仮説検定
 - T検定 & ウェルチの検定
 - 一元配置分散分析 (One-way ANOVA)
- 周囲に展開する方法
 - Live Editor
 - アプリ化



仮説検定

一元配置分散分析 (one-way ANOVA)

分散分析は分散を利用して、T 検定の様に母平均を検定する手法

2群の比較 ➡ T 検定

3群以上の比較 ➡ 分散分析 (**AN**alysis **Of** **VA**riance) ✓

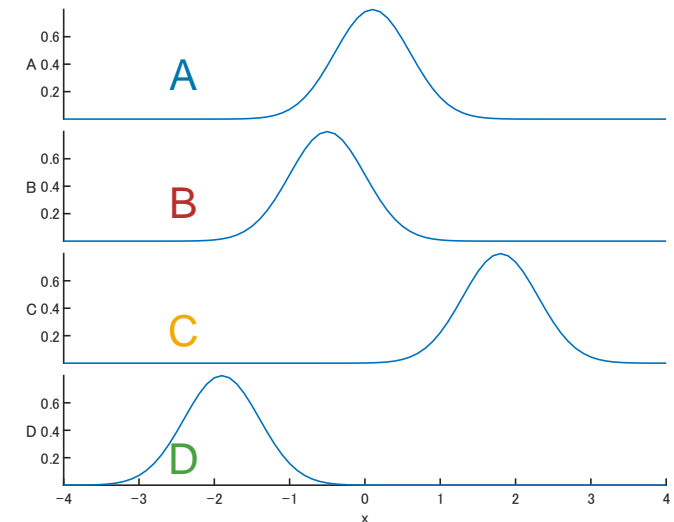
※ T 検定を3群以上くり返し実施するのはダメ!

例えば、正しい確率が 95% (有意水準 5%) の
検定を3回実施し、全部正しい確率は? $\rightarrow 0.95^3 = 85.7\%$

STEP.1 仮説を設定

• 帰無仮説 (仮定する仮説) $H_0: \mu_A = \mu_B = \dots = \mu$

• 対立仮説 (主張した仮説) $H_1: \exists i \mu_i \neq \mu$ 注意



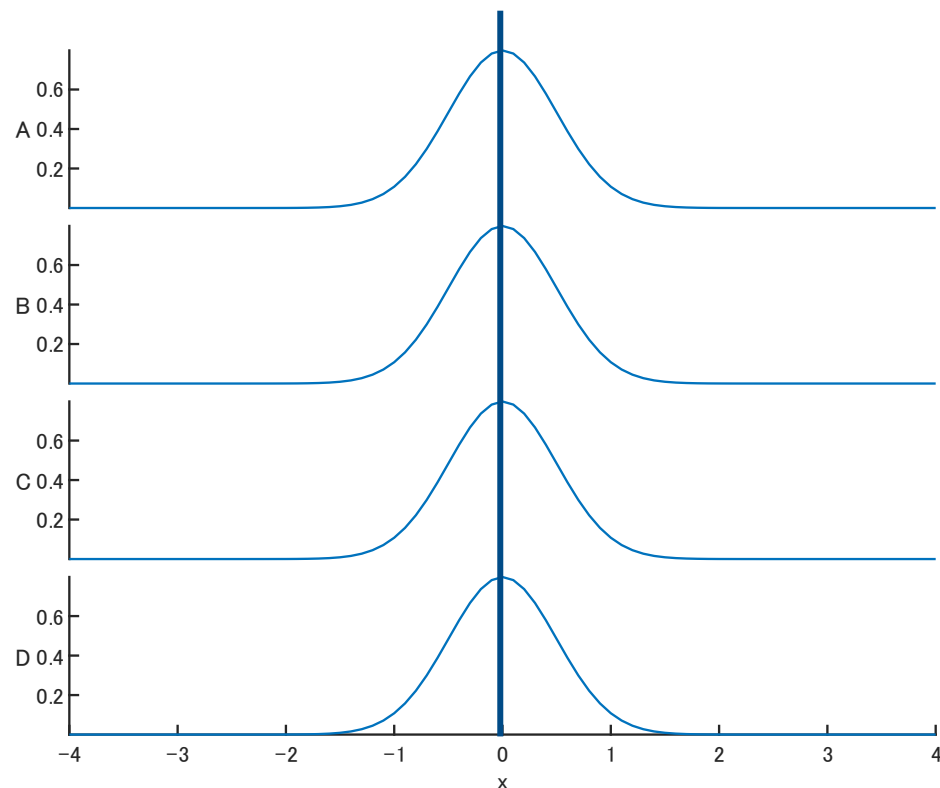
~~“A vs B, B vs C”~~
T 検定を複数やれば良い?

仮説検定 (Cont.)

一元配置分散分析 (one-way ANOVA)

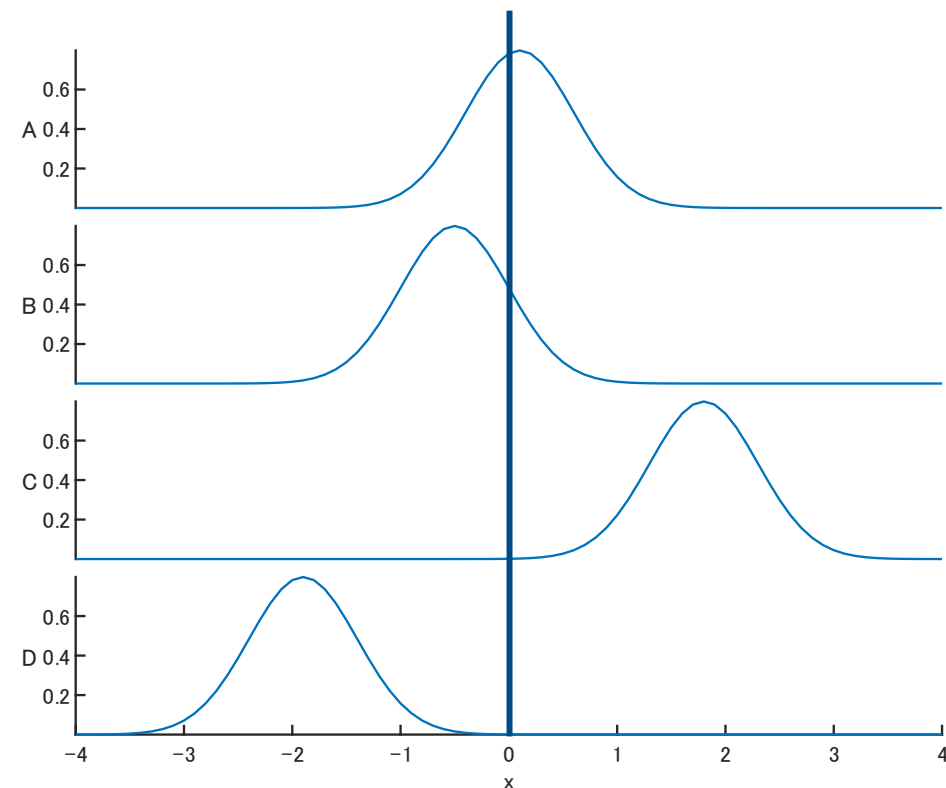
※各グループは等分散を仮定します

- 帰無仮説 (仮定する仮説) $H_0: \mu_A = \mu_B = \dots = \mu$



母平均が全て同じ

- 対立仮説 (主張した仮説) $H_1: \exists i \mu_i \neq \mu$



母平均が異なるものが存在する

仮説検定 (Cont.)

一元配置分散分析 (one-way ANOVA)

STEP.2 統計検定量を用意、有意水準を決める

$$\frac{MSR}{MSE} \sim F_{k-1, N-k}$$

k : グループ数

N : 観測値総数

自由度 ($k-1, N-k$) の F 分布に従う

STEP.3 統計検定値を計算

SSR (グループ間変動和), SSE (グループ内変動和)

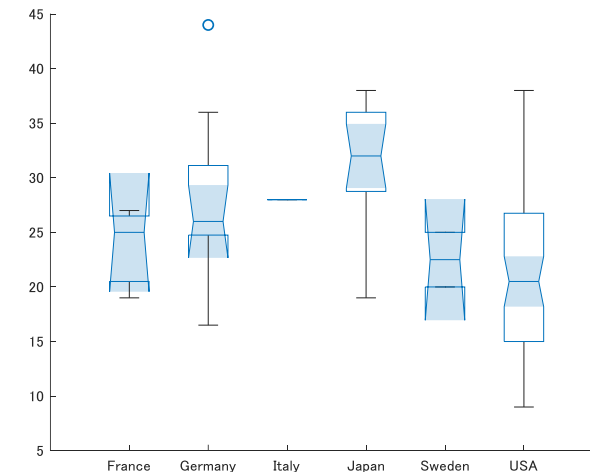
$SST = SSR + SSE$, (総変動和)

	SS	DF	MeanSquares	F	確率 $> F$
Factor1	SSR	$k - 1$	$MSR = SSR / (k - 1)$	MSR / MSE	$P(F_{k-1, N-k}) > F$
Error	SSE	$N - k$	$MSE = SSE / (N - k)$		
Total	SST	$N - 1$			

ANOVA Table

MATLAB で実行

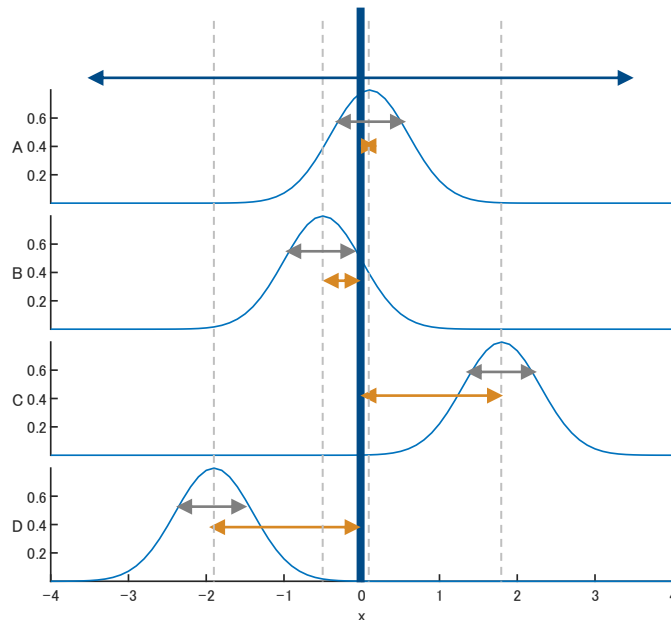
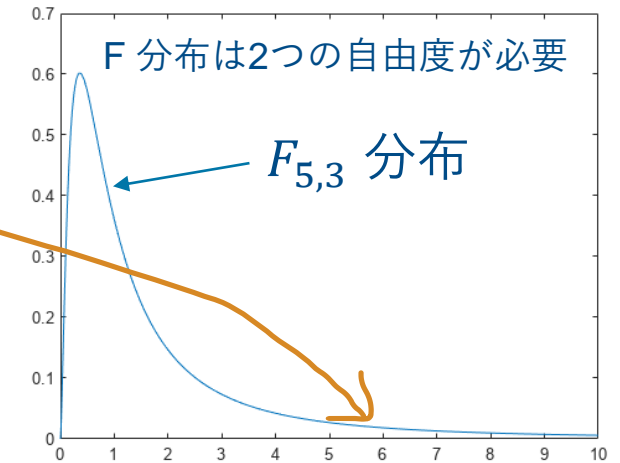
```
[p, anvtbl, stats] = anova(group, y)
```



仮説検定 (Cont.)

一元配置分散分析 (one-way ANOVA)

	Sum of Square	degree of freedom		F 統計値	
	SS	DF	MeansSquares	F	確率 > F
Factor1	SSR	$k - 1$	$MSR = SSR / (k - 1)$	MSR / MSE	$P(F_{k-1, N-k}) > F$
Error	SSE	$N - k$	$MSE = SSE / (N - k)$		p 値
Total	SST	$N - 1$			



$$SST = SSR + SSE$$

総平方和 グループ間変動和 グループ内変動和

$$SSR = \leftarrow \overset{2}{\leftarrow} \overset{2}{\leftarrow} \overset{2}{\leftarrow} \overset{2}{\leftarrow}$$

$$SSE = \leftarrow \overset{2}{\leftarrow} \overset{2}{\leftarrow} \overset{2}{\leftarrow} \overset{2}{\leftarrow}$$

$$SST = \leftarrow \overset{2}{\leftarrow} \overset{2}{\leftarrow} \overset{2}{\leftarrow} \overset{2}{\leftarrow} \overset{2}{\leftarrow}$$

2乗



$$\frac{SSR}{SSE} \propto \frac{MSR}{MSE} \sim F_{k-1, N-k}$$

グループ間のばらつき度
グループ内のデータばらつき度

仮説検定 (Cont.)

一元配置分散分析 (one-way ANOVA)

フランス、ドイツ、イタリア、日本、スウェーデン、アメリカで製造された自動車100台の燃費 (MPG) データがある。製造国による燃費の違いが有ると言えるか? 有るとすれば、どの国が燃費が異なるか?

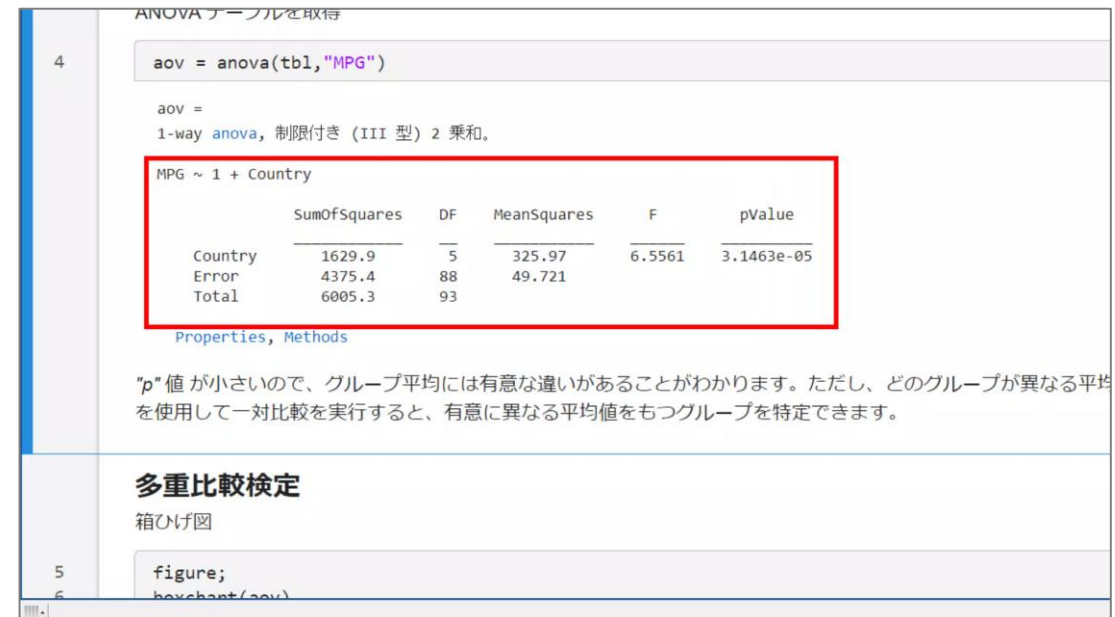
ANOVA

- 帰無仮説 (仮定する仮説) $H_0: \mu_A = \mu_B = \dots = \mu$
- 対立仮説 (主張した仮説) $H_1: \exists i \mu_i \neq \mu$



多重比較検定

米国車と日本車の燃費は異なる



仮説検定 (Cont.)

一元配置分散分析 (one-way ANOVA) – まとめ (厳密性無視)

ANOVA の対立仮説には要注意

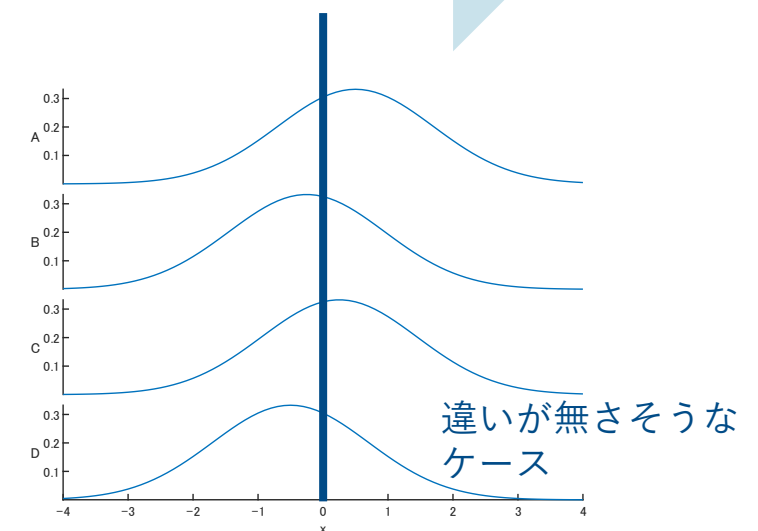
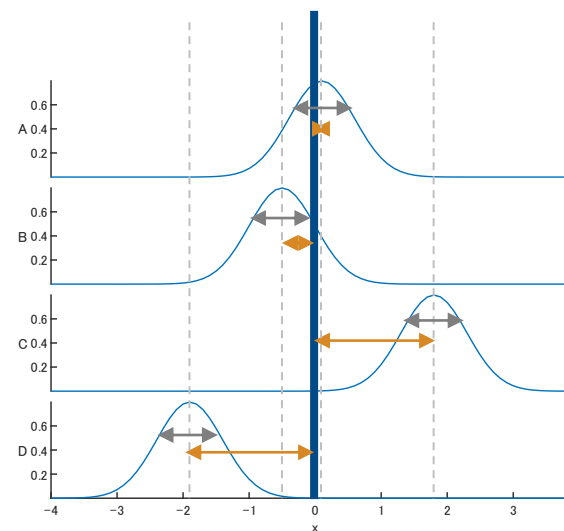
- 帰無仮説 (仮定する仮説) $H_0: \mu_A = \mu_B = \dots = \mu$
- 対立仮説 (主張した仮説) $H_1: \exists i \mu_i \neq \mu$

ANOVA テーブルは
左から右に計算が流れる
検定統計量は F 値

	SS	DF	MeanSquares	F	確率 > F
Factor1	SSR	$k - 1$	$MSR = SSR / (k - 1)$	MSR / MSE	$P(F_{k-1, N-k}) > F$
Error	SSE	$N - k$	$MSE = SSE / (N - k)$		p 値
Total	SST	$N - 1$			

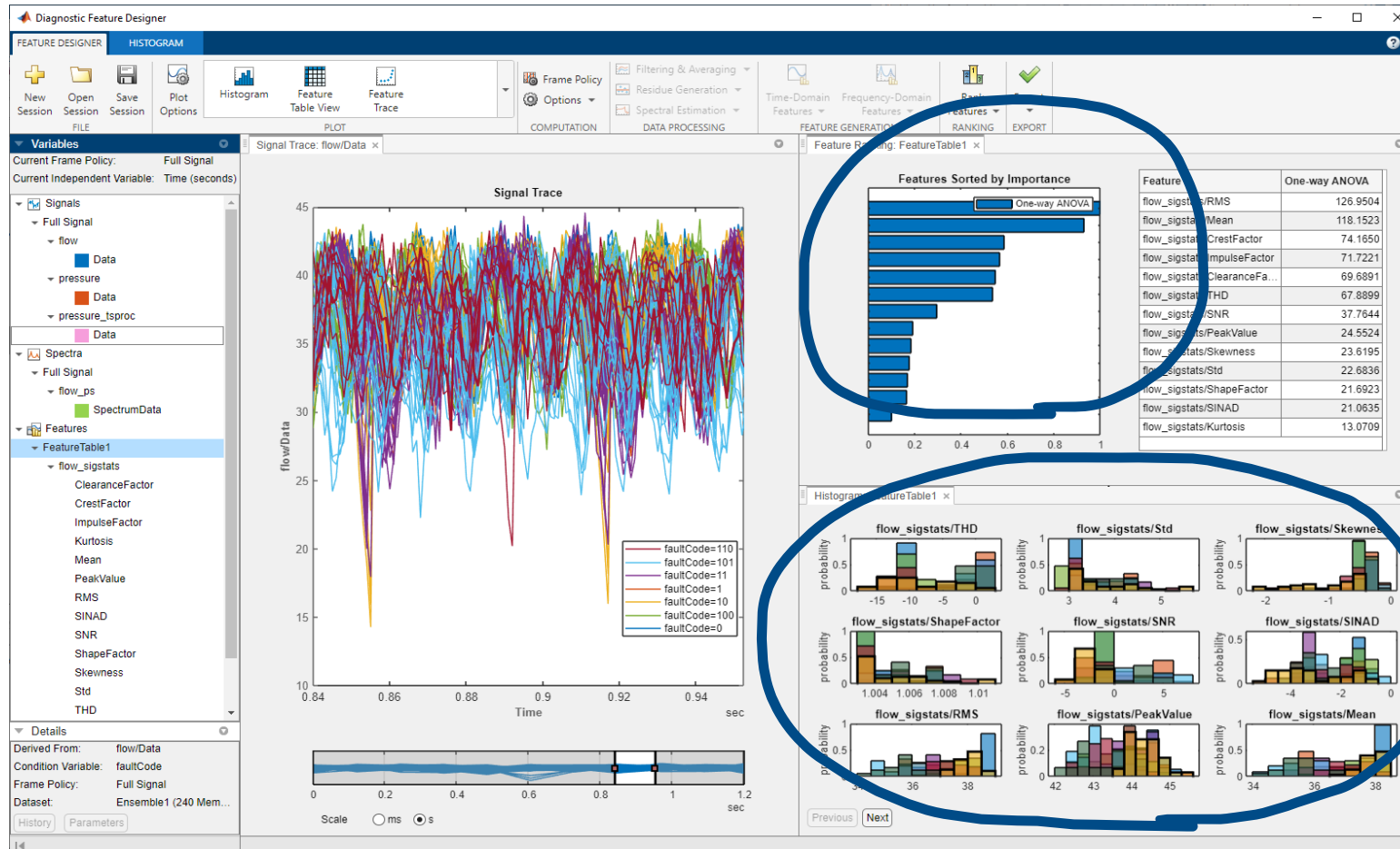
$$F = \frac{\text{グループ間のばらつき度}}{\text{グループ内のデータばらつき度}}$$

グループ内のばらつきに対して
グループ間のばらつきが ** 倍超えたら、
母平均に違いがあるとみなす



仮説検定 (Cont.)

一元配置分散分析 (one-way ANOVA) – 特徴診断デザイナーアプリ



One-way ANOVA を
特徴量選択に利用

特徴量のランキング

MathWorks®

ライブイベント

ラベリング工数大幅削減！アプリで楽するデータラベリングと前処理 - AI モデル検討の前に

開始時間	終了時間
2023年3月23日, 14:00	2023年3月23日, 14:45

概要

機械学習のプロジェクトにおいて前処理が重要なステップであることには異論の余地はなく、学習データの質が悪い場合は、どれだけ高性能なアルゴリズムを使用したとしても実用的な精度は期待できません。プロジェクトの早い段階で正しく準備された

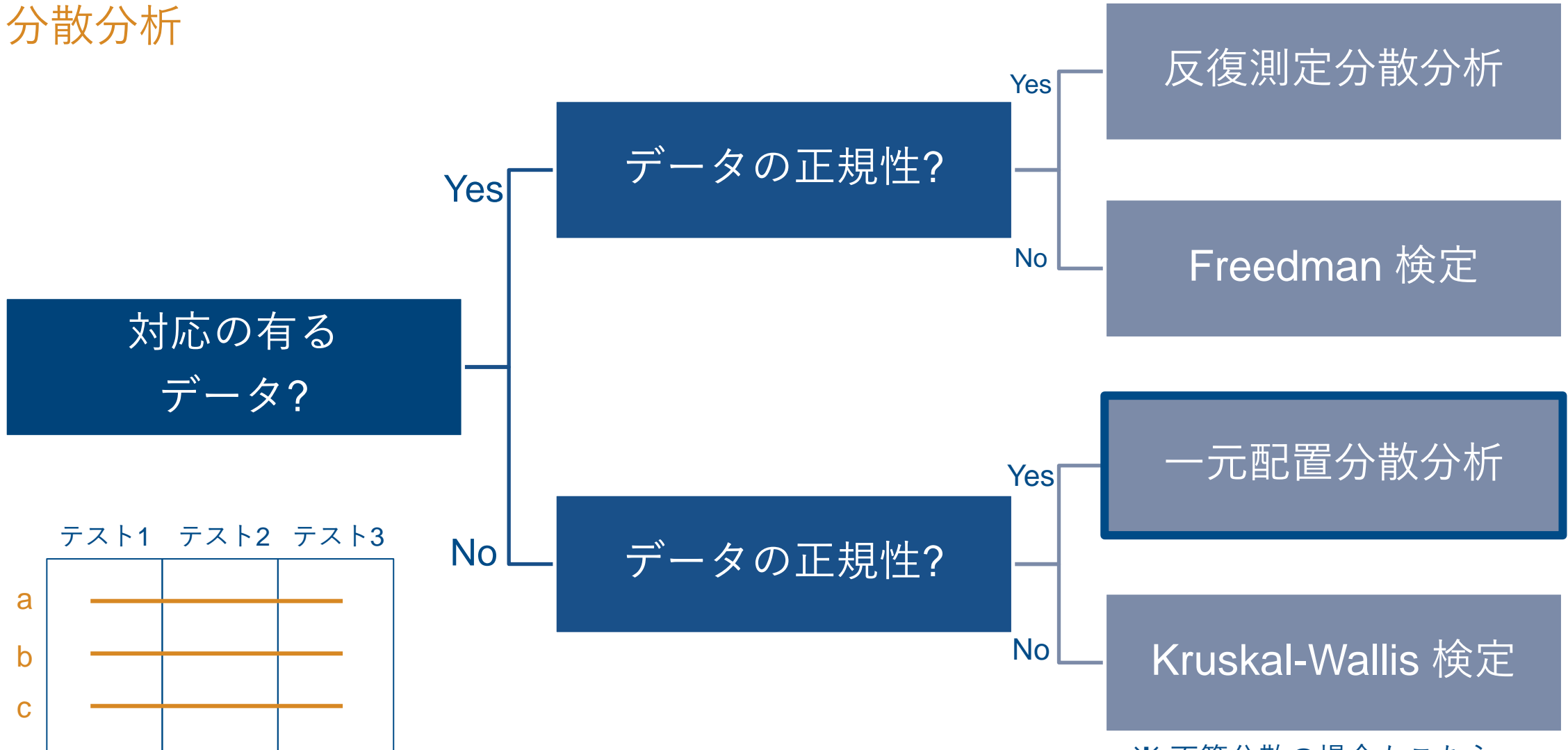
- 診断特徴デザイナーアプリの使い方
- ノーコード異常検知
- 特徴量を90個一気に作成!

[詳細はこちらから](#)

複数クラスのカテゴリ用特徴量

仮説検定 (Cont.)

分散分析



※ 不等分散の場合もこちら
(ノンパラメトリック検定)

仮説検定 (Cont.)

一元配置分散分析後の多重比較検定

```
aov = anova(group,y)
```

ANOVA 実行

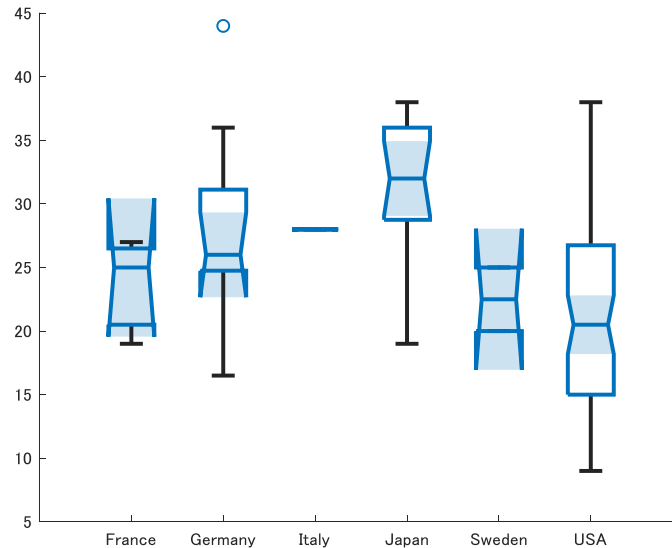


	SumOfSquares	DF	MeanSquares	F	pValue
Factor1	1629.9	5	325.97	6.5561	3.1463e-05
Error	4375.4	88	49.721		
Total	6005.3	93			

どのグループの平均が異なるのか分からない

```
b = boxchart(aov)
```

箱ひげ図



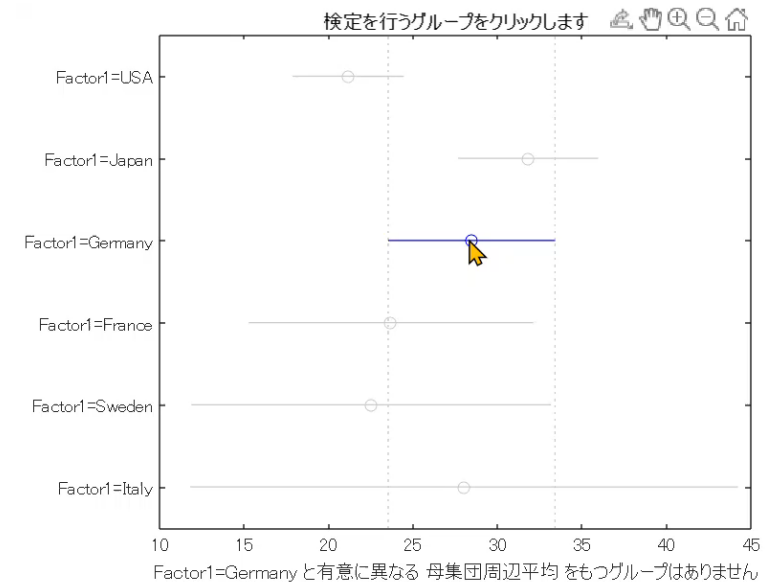
```
C = multcompare(aov)
```

多重比較検定

	Group1	Group2	MeanDifference	MeanDifferenceLower	MeanDifferenceUpper	pValue
1	'USA'	'Japan'	-10.6672	-16.5605	-4.7739	0.0019
2	'USA'	'Germany'	-7.3116	-14.6252	0.0019	0.0501
3	'USA'	'France'	-2.5339	-14.6696	9.6018	0.9902
4	'USA'	'Sweden'	-1.3672	-16.1190	13.3846	0.9998
5	'USA'	'Italy'	-6.8672	-27.5707	13.8364	0.9273
6	'Japan'	'Germany'	3.3556	-5.3064	12.0175	0.8882
7	'Japan'	'France'	8.1333	-4.8596	21.1263	0.4561
8	'Japan'	'Sweden'	9.3000	-6.1647	24.7647	0.5018
9	'Japan'	'Italy'	3.9000	-17.4174	25.0174	0.9952
10	'Germany'	'France'	4.5778	-8.9600	18.4736	0.9113

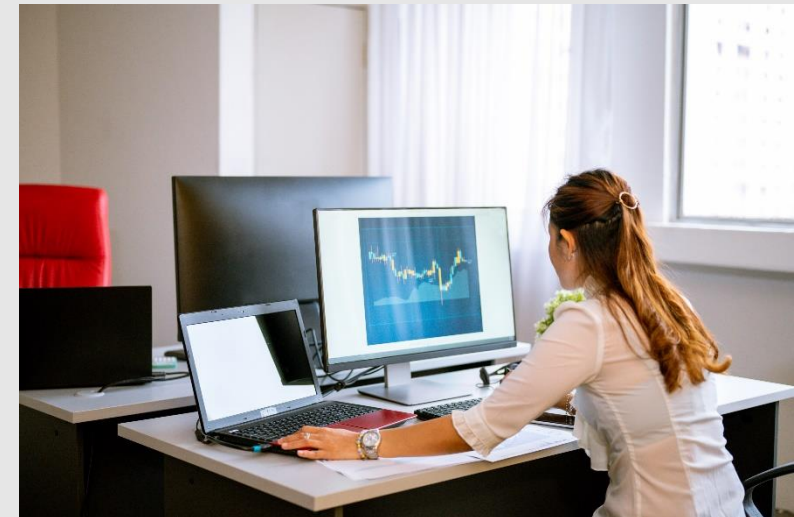
```
f = plotComparisons(aov)
```

多重比較の対話型プロット



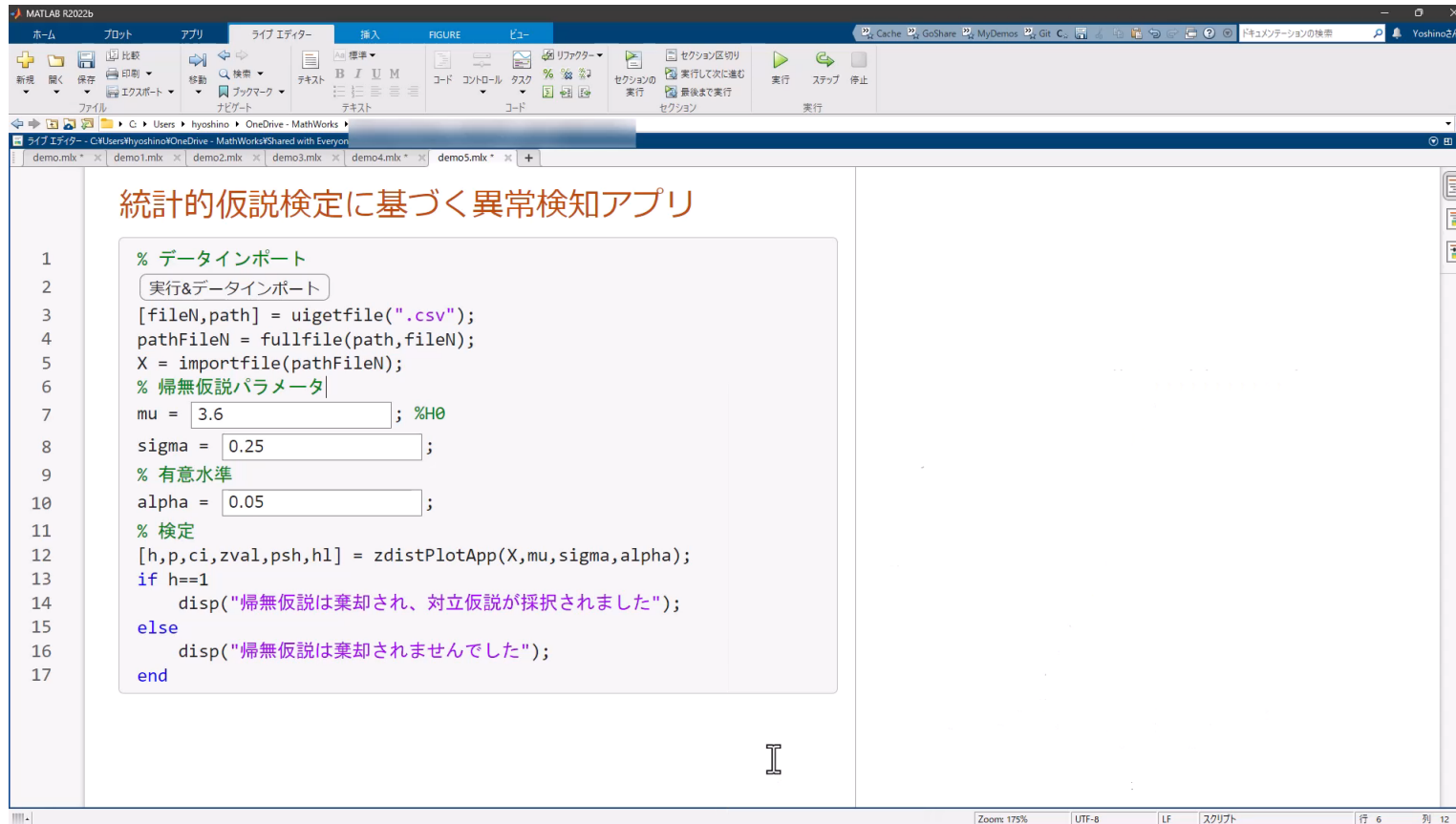
Agenda

- 統計学の違い
- 確率変数とは
- 確率分布
- 区間推定
 - 母平均区間推定
 - 母分散区間推定
- 仮説検定
 - 母平均仮説検定
 - T検定 & ウェルチの検定
 - 一元配置分散分析 (One-way ANOVA)
- 周囲に展開する方法
 - Live Editor
 - アプリ化



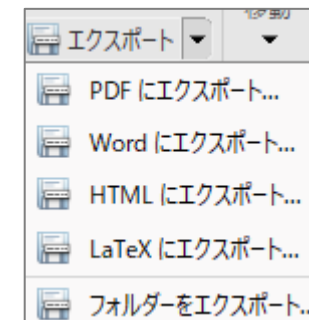
統計的仮説検定アプリ

Live Editor はコードを隠せば、誰でも利用可能なミニアプリに



1	3.733441678
2	3.765847125
3	3.663528828
4	3.741554333
5	3.727969131
6	3.687307793
7	3.709160199
8	3.728565612
9	3.809459923
10	3.789235926
11	

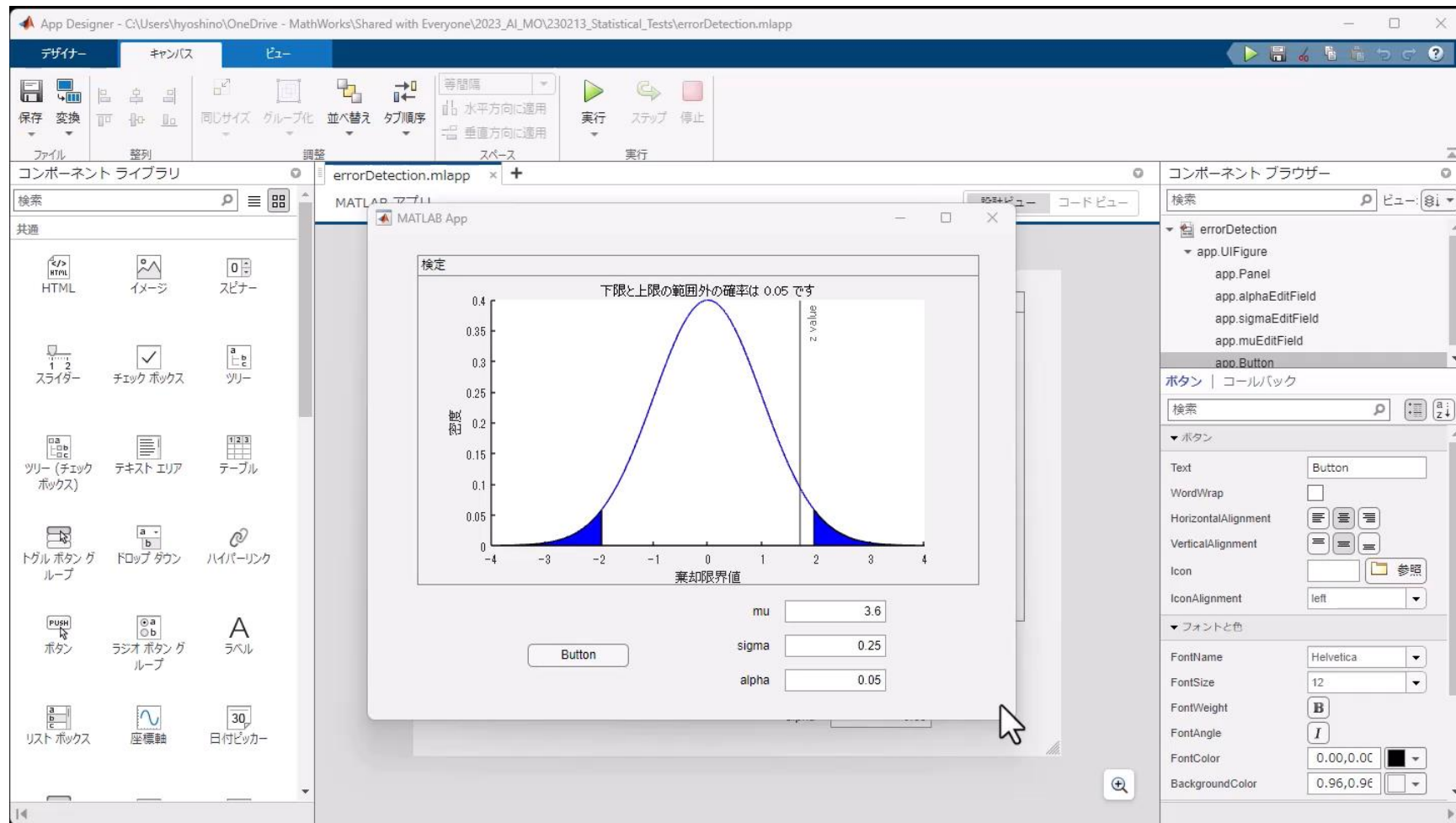
測定結果の csv ファイル



様々な
フォーマットで保存

統計的仮説検定アプリ

App Designer で GUI を加えてスタンドアロンアプリへ (ライセンス不要で動作)



ウェブアプリとしてリモート共有 現場への展開が一瞬で行える

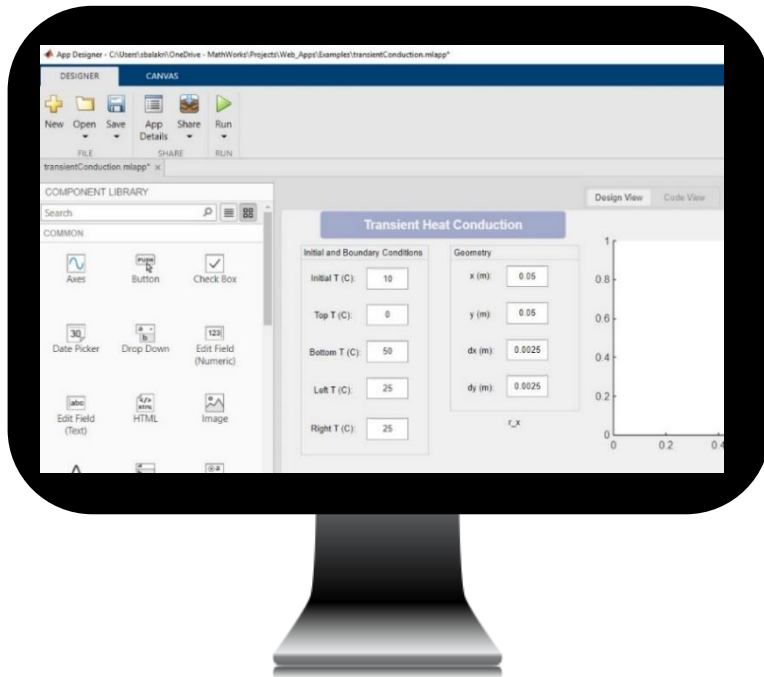
GUI アプリ設計



ホストし、アプリを共有



リモート &
ウェブブラウザで実行



App Designer



MATLAB Web App Server



$$mc^2 \quad \frac{pV}{RT} \quad \epsilon_0 \mathbf{E} + \mathbf{P}$$



確率分布の利用法

確率密度分布の選択方法

1. データが離散か連続かを判断
2. データの分布形がどのような形をしているかを観察
3. データの分布形に近い確率密度関数を候補とする
4. 候補となる確率密度関数のパラメータを推定
5. 推定したパラメータを用いて、データと確率密度関数の適合度を検定する
6. 適合度が高い確率密度関数を選択

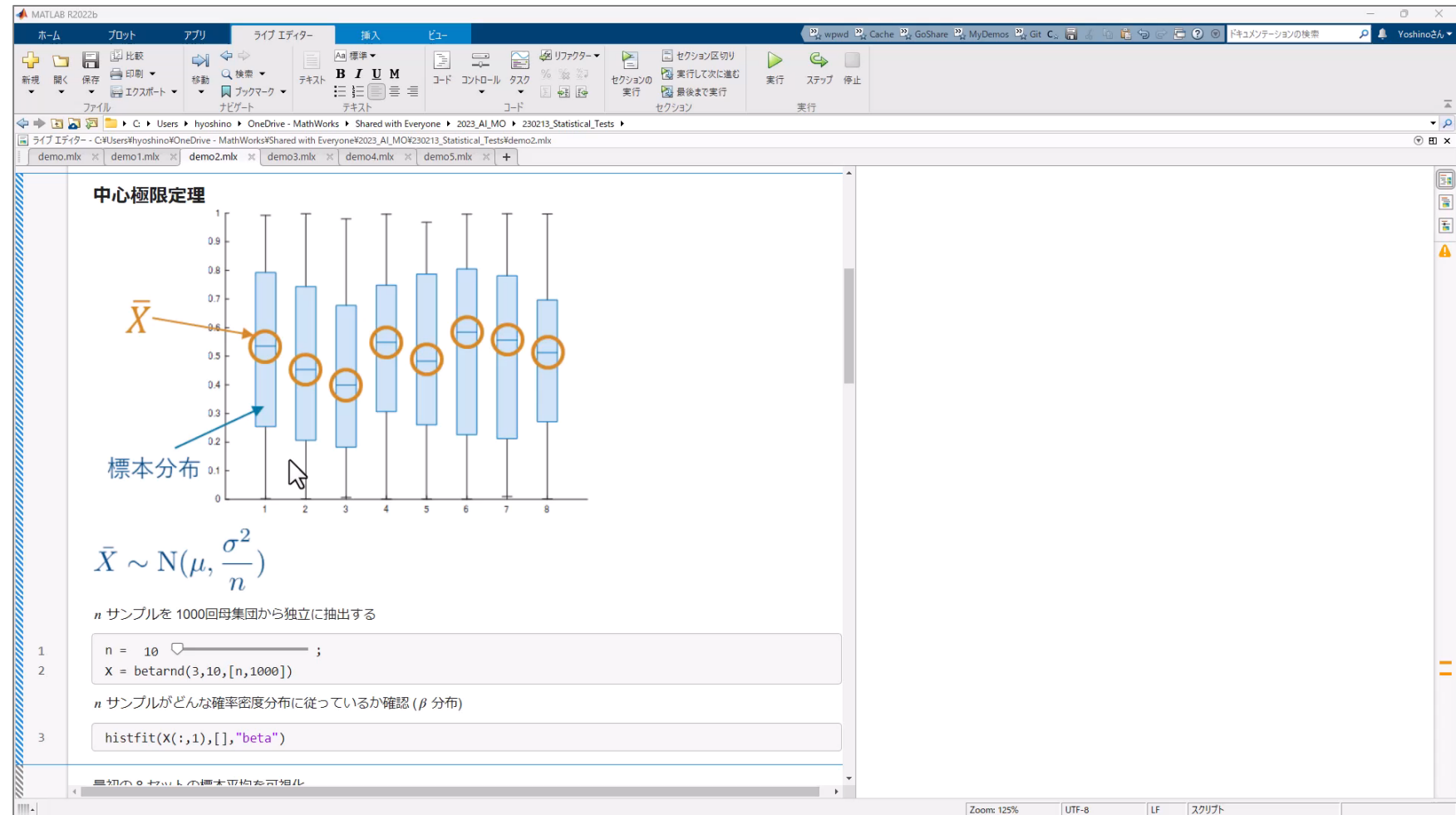
カイ二乗検定適合度検定

E.g.,

- データが連続、分布形が釣り鐘型 >> 正規分布、t 分布 ...
- データが連続、分布形が左右非対称 >> 指数分布、ガンマ分布 ...
- データが離散、成功・失敗のベルヌーイ試行 >> 二項分布、ポアソン分布 ...

区間推定

- 中心極限定理
- 母平均区間推定
- 母分散区間推定

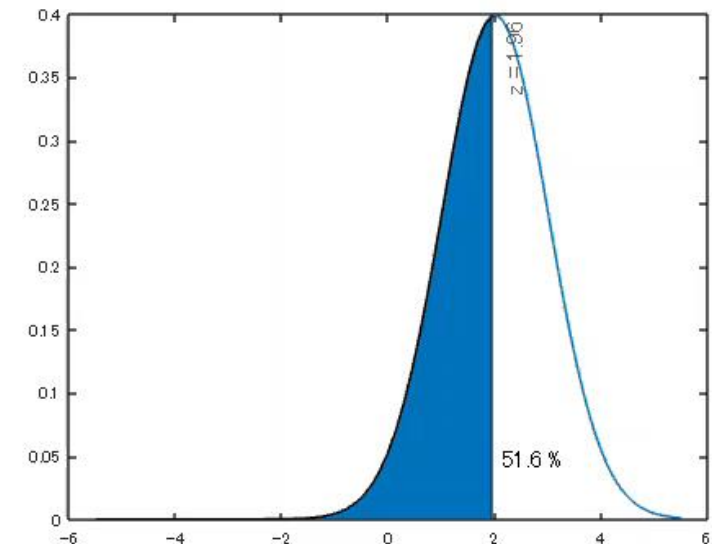
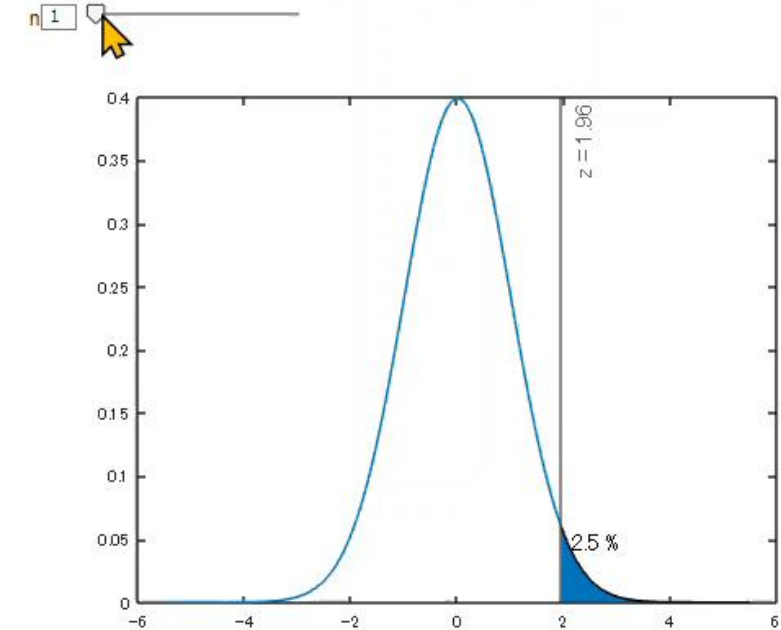


仮説検定

補足: n (サンプル数) を増やすとどうなるか?

- $H_0: \mu = \mu_0 \quad \frac{\bar{X} - \mu_0}{\sqrt{\sigma^2/n}} \sim N(0, 1)$
- $H_1: \mu = \mu_1 \quad \frac{\bar{X} - \mu_0}{\sqrt{\sigma^2/n}} \sim N\left(\frac{\mu_1 - \mu_0}{\sqrt{\sigma^2/n}}, 1\right)$

正規分布: n が大きくなっていくとどうなるか?

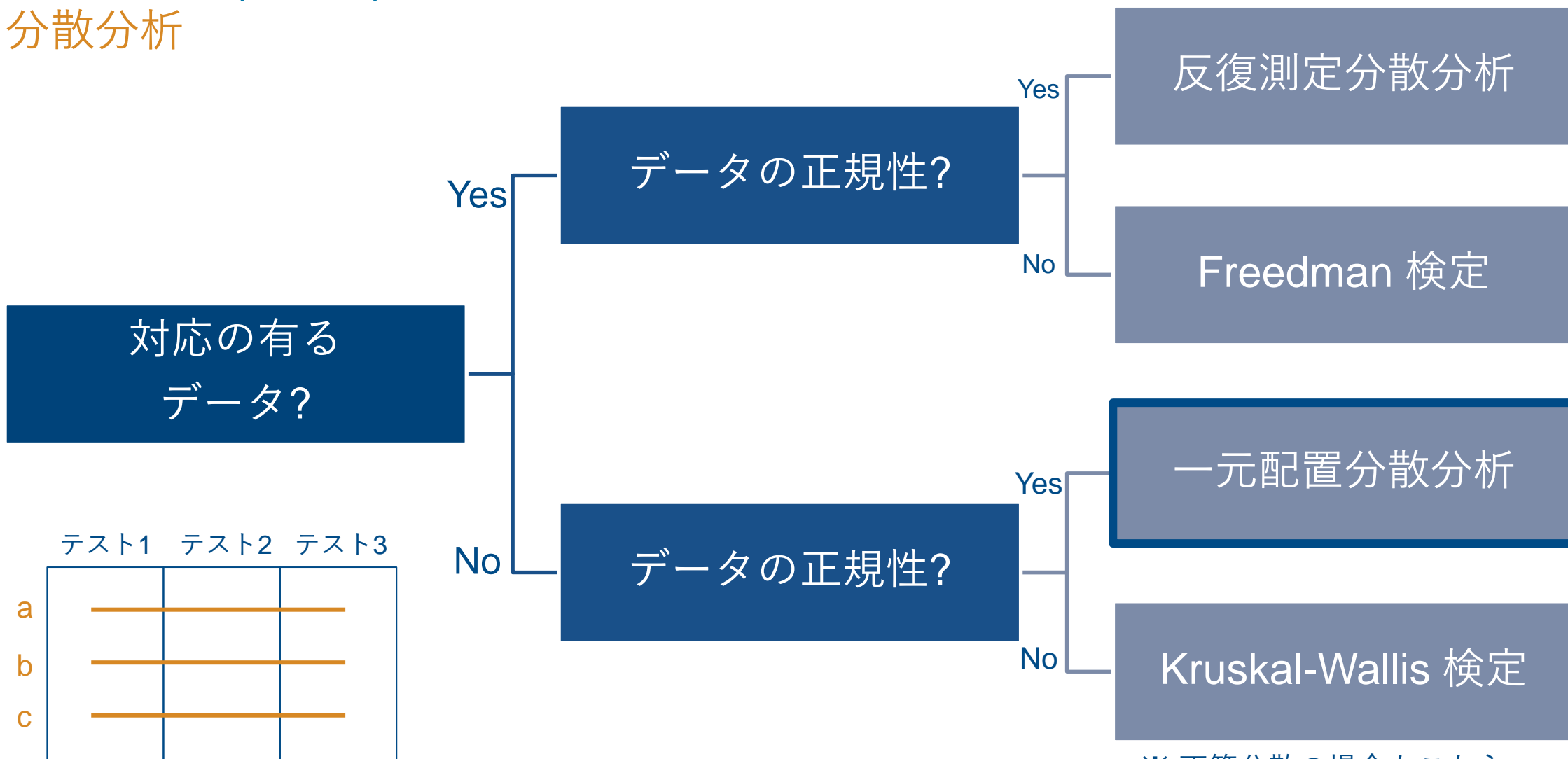


判断 \ 真実	真実	
	H_0 が正しい	H_1 が正しい
H_0 を棄却	Type I error	
H_0 を棄却しない		Type II error

検出力の UP

仮説検定 (Cont.)

分散分析



	テスト1	テスト2	テスト3
a	—	—	—
b	—	—	—
c	—	—	—