Confidential

統計解析 – Day2

確率/統計的推定/仮説検定

AIJobColle

統計解析入門

【Day2】講義内容

- 確率/確率分布
  - 事象、確率、条件付き確率
  - 離散/連続確率分布
  - 期待値と分散

- 統計的推定
  - 母集団、標本
  - パラメータ、点/区間推定
  - 母平均、母比率の推定

- 仮説検定
  - 検定の考え方
  - 帰無/対立仮説、有意水準、2種類の誤り
  - 母平均/比率の差の検定
  - 検定における注意点

事象と確率

標本空間(Sample Space)の部分集合のうち特別に選ばれたものを**事象(Event)**と呼ぶ。
（例）サイコロを1回振る場合、標本空間は{1,2,3,4,5,6}。2の目が出る事象は{2}

和事象：事象A,B,Cのうち、少なくとも1つが起こる事象。$A \cup B \cup C$
（例）サイコロを1回振る場合、2または3の目が出る事象(和事象)は$\{2\} \cup \{3\} = \{2,3\}$

積事象：事象A,B,Cが同時に起こる事象。$A \cap B \cap C$
（例）サイコロを1回振る場合、偶数かつ3の目以下が出る事象(積事象)は$\{2,4,6\} \cap \{1,2,3\} = \{2\}$

事象の起こりやすさを**確率(Probability)**で表現する。

事象Aが起こる確率：$P(A)$
（例）サイコロを1回振る場合、2の目が出る確率は$P(\{2\}) = 1/6$

AとBの和事象が起こる確率：$P(A \cup B)$
（例）サイコロを1回振る場合、2または3の目が出る確率は$P(\{2\} \cup \{3\}) = P(\{2,3\}) = 2/6 = 1/3$

AとBの積事象が起こる確率：$P(A \cap B)$
（例）サイコロを1回振る場合、偶数かつ3の目以下が出る確率は$P(\{2,4,6\} \cap \{1,2,3\}) = P(\{2\}) = 1/6$

(Wiki) https://ja.wikipedia.org/wiki/%E7%A2%BA%E7%8E%87%E8%AB%96#%E5%9F%BA%E7%A4%8E%E6%A6%82%E5%BF%B5

条件付き確率、独立

条件付き確率(Conditional Probability)は、ある事象Bが起こるという条件下での別の事象Aの起こる確率
P(A|B) = P(A∩B) / P(B)

（例）サイコロを2回振り、合計が6以上になる確率。ただし1回目に3の目が出ている。
P(“2回目で合計6以上”|”1回目3が出る”)
= P(“2回目で合計6以上”∩”1回目3が出る”) / P(”1回目3が出る”)
= P(“2回目で3,4,5,6が出る”∩”1回目3が出る”) / P(”1回目3が出る”)
= P(“{3,3},{3,4},{3,5},{3,6}のいずれかが出る”) / P(”{3}が出る”)
= (4/36)/(1/6) = 4/6 = 2/3

事象Bが起こるという条件が、別の事象Aの起こる確率に影響しない場合、事象AとBは**独立(Independent)**
P(A|B) = P(A)
また、P(A∩B) = P(A) P(B)

（例）サイコロを2回振る、2回とも6の目が出る確率は？
P(“1回目6が出る”∩“2回目6が出る”)
= P(”1回目6が出る”) P(”2回目6が出る”)
= 1/6 * 1/6 = 1/36

(Wiki) https://ja.wikipedia.org/wiki/%E6%9D%A1%E4%BB%B6%E4%BB%98%E3%81%8D%E7%A2%BA%E7%8E%87

(応用) バスケット分析

バスケット分析(アソシエーション分析)は、購買トランザクションデータ（PoSデータ）などから、何と何がどれくらい一緒に買われるかを分析する手法です

バスケットデータ（アイテムセットの集合）

バスケット
A, B, C, D
A, E, F, C
G, F, E
H, F, E, I
H, F, J, D

アイテムセット

アイテム
A
B
C
D
E
F
G
H
I
J

「Aを買うとBも買う」の表記：A⇒B

A⇒Bの**支持度(Support)**
- バスケット中、AとBが含まれる確率 P(A∩B)
- バスケット分析では以下の表記が用いられる
  supp(A⇒B) = 事象数(A∩B) / バスケット数
- 支持度が低いと、そもそもどのバスケットにも含まれないアイテムの組み合わせなので、調べる興味度合いは低い

A⇒Bの**確信度(Confidence)**
- Aが含まれるバスケットの内、Bが含まれる確率 P(B|A) = P(A∩B) / P(A)
- バスケット分析では以下の表記が用いられる
  conf(A⇒B) = supp(A⇒B) / 事象数(A)
- Aと同時買いされるアイテムの指標

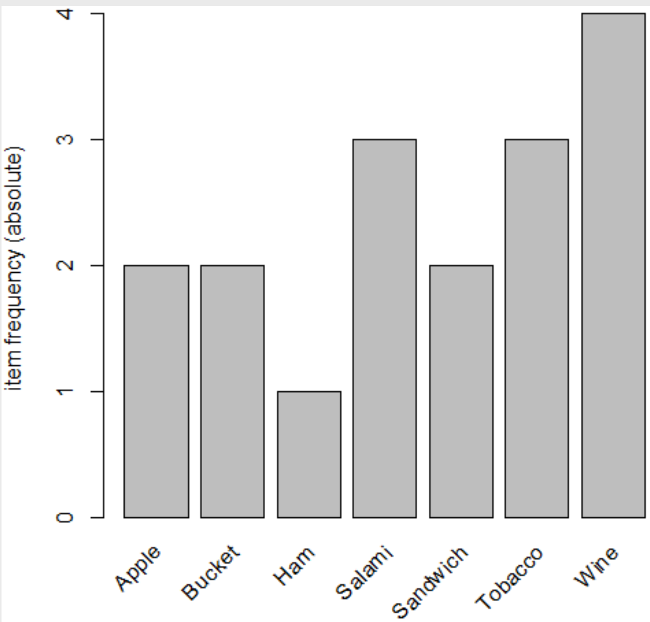
演習【Day2-Exercise1】バスケット分析

- バスケット分析を例に、確率と条件付き確率を確認
- arulesパッケージ（https://cran.r-project.org/web/packages/arules/arules.pdf）

演習の解説

arulesパッケージを利用したバスケット分析
Basket_en2.csv

```
                            items
1          {Apple,Bucket,Salami}
2 {Bucket,Salami,Tobacco,Wine}
3              {Ham,Tobacco,Wine}
4         {Sandwich,Tobacco,Wine}
5 {Apple,Salami,Sandwich,Wine}
```



- BucketとSalamiが同時に買われる確率は？
- Bucketが買われた内、Salamiが買われる確率は？
- Salamiが買われた内、Bucketが買われる確率は？

- Wine ⇒ Tobacco の支持度、確信度は？
- Sandwich,Wine ⇒ Apple の支持度、確信度は？

確率変数と確率関数

変数とその値の出現確率を数式で定義する

事前にとりえる値がわからない変数Xを**確率変数(Random Variable)**と呼び、Xのとりえる各値の出現確率を式で表現したものが**確率関数(Probability Function)**と呼ばれる

例）サイコロを振る

サイコロの目の値が変数X
確率関数はP(X=x)=f(x)=1/6，x=1,2,3,4,5,6

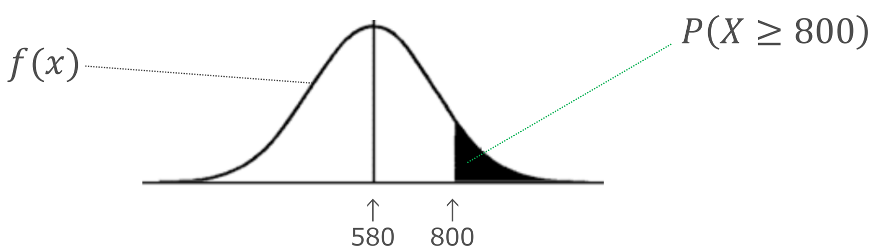
解釈：Xの実現値であるxは、1から6のいずれかの値をとり、各値の出現確率は1/6

例）ランダムに抽出したTOEIC受験者のスコア

TOEICのスコアが、平均580、標準偏差170の正規分布に従っていると仮定する
（正確には最高点、最低点が決まっているので正規分布を仮定できないが説明のため）
ランダムに抽出したあるTOEIC受験者のスコア(X)が800点以上である確率は？

平均580、標準偏差170の正規分布の確率関数： $P(X = x) = f(x) = {}^{1}/_{\sqrt{2\pi}170}\, e^{-(x-580)^2/_{2*170^2}}$

Xが800以上となる確率： $P(X \geq 180) = \int_{x=800}^{\infty} f(x)dx = \int_{x=800}^{\infty} {}^{1}/_{\sqrt{2\pi}170}\, e^{-(x-580)^2/_{2*170^2}}\, dx$



様々な確率分布、関数

確率分布(Probability Distribution)は、確率変数の各々の値に対して、その起こりやすさを記述するもの

変数が離散的な値をとるとき、その変数の確率分布は**離散型**となり、連続値をとるとき、その変数の確率分布は**連続型**となる

確率関数(Probability Function)を用い、確率分布を表現する

代表的な離散型確率分布
- ベルヌーイ分布
- 二項分布
- 多項分布
- ポアソン分布

代表的な連続型確率分布
- 一様分布
- 正規分布
- t分布
- 指数分布

後に学習する統計的仮説検定や、統計モデルで必要な部品となる

(Wiki) https://ja.wikipedia.org/wiki/%E7%A2%BA%E7%8E%87%E5%88%86%E5%B8%83

離散確率分布 – 二項分布

パラメータpならび自然数のパラメータnに対して、自然数を値としてとる確率変数Xが次の確率関数を持つとき、確率変数Xはパラメータ n、p の二項分布B(n, p)に従うという

X~B(n, p) と書く

$$P[X=k] = \binom{n}{k} p^k (1-p)^{n-k} \quad \text{for } k = 0, 1, 2, \ldots, n$$

$$\binom{n}{k} = {}_nC_k = \frac{n!}{k!(n-k)!}$$

解釈：n回の試行のうち、k回成功する確率。一回の試行での成功確率はp

（例）サイコロを3回振る、1回だけ6の目が出る確率は？
（6の目が出る回数k=1、3回の試行n=3、一回の成功確率p=1/6）

P[X=1] = 3 * 1/6 * (5/6)^2 = 75/216 = 25/72

(Wiki) https://ja.wikipedia.org/wiki/%E4%BA%8C%E9%A0%85%E5%88%86%E5%B8%83

連続確率分布 – 正規分布

平均を μ, 分散を $\sigma^2 > 0$ とする正規分布とは、次の形の確率関数を持つ確率分布のことである
この分布を $N(\mu, \sigma^2)$ と表す（μ, σ^2が正規分布におけるパラメータ）

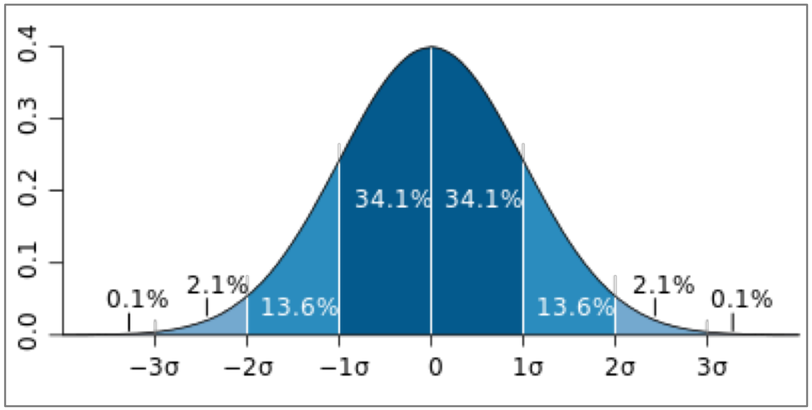
$X \sim N(\mu, \sigma^2)$

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

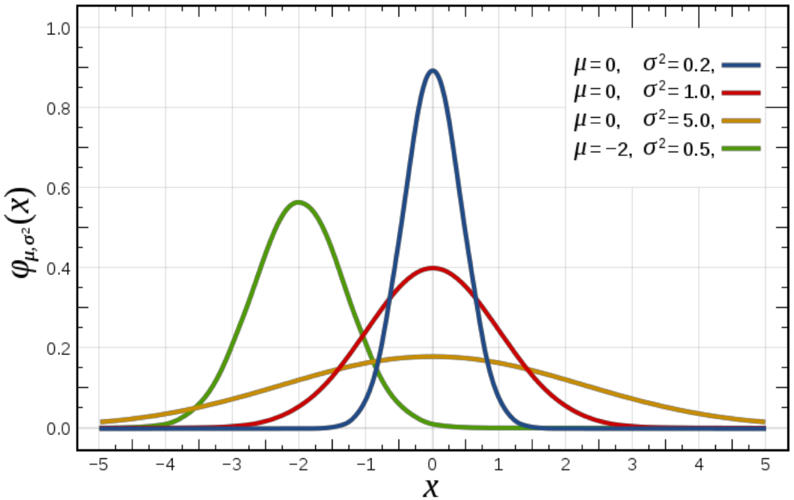
$\mu=0, \sigma^2=1$のとき、この分布は**標準正規分布**と呼ばれる
$Z \sim N(0, 1)$

$$f(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right)$$

標準正規分布における各σ内の確率



μ,σ²を変化させたときの分布
（赤は標準正規分布）

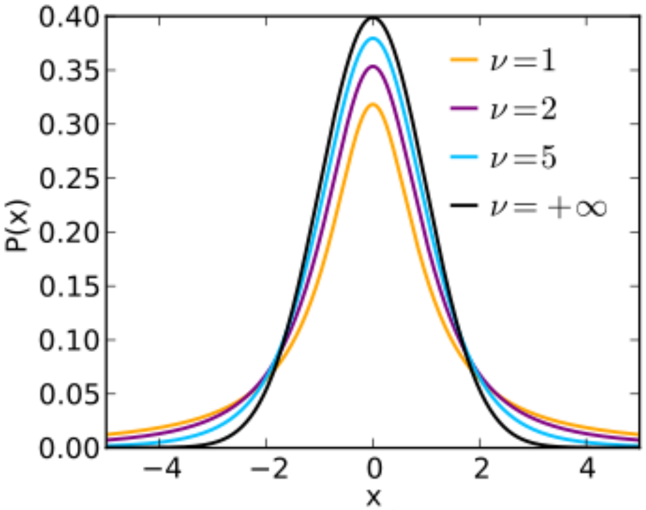


(Wiki) https://ja.wikipedia.org/wiki/%E6%AD%A3%E8%A6%8F%E5%88%86%E5%B8%83

(参考) 連続確率分布 – t分布

t分布(Student's t-distribution)は、連続確率分布の一つであり、正規分布する母集団の平均と分散が未知で標本サイズが小さい場合に平均を推定する問題に利用される。また、2つの平均値の差の統計的有意性を検討するt検定で利用される

$$f(t) = \frac{\Gamma((\nu+1)/2)}{\sqrt{\nu\pi}\,\Gamma(\nu/2)}(1+t^2/\nu)^{-(\nu+1)/2}$$



xもしくはt

vは自由度と呼ばれ、大きくなるに従い分布の裾野が狭まり、正規分布に近づく

”v=データ数-1”なので、データ数が多い時は、正規分布に置き換えて考えてよい

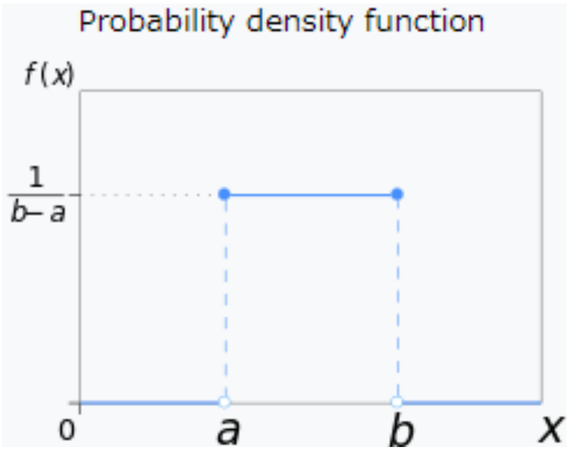
(Wiki) https://ja.wikipedia.org/wiki/T%E5%88%86%E5%B8%83

離散/連続確率分布 – その他の分布

一様分布(Uniform Distribution)

X～N(a, b)

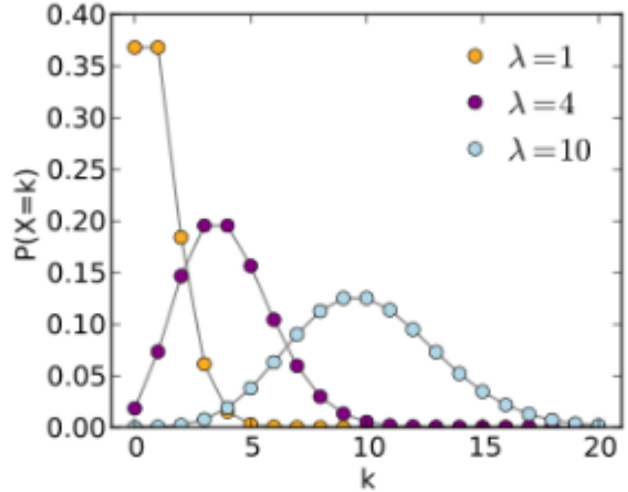
$$f(x) = \begin{cases} \frac{1}{b-a} & \text{for } a \le x \le b, \\ 0 & \text{for } x < a \text{ or } x > b, \end{cases}$$



ポアソン分布(Poisson distribution)

X～Poisson(λ)

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$



Rで利用できる確率分布
http://cse.naro.affrc.go.jp/takezawa/r-tips/r/60.html

期待値と分散

期待値(Expectation)または平均は、確率変数の実現値を確率の重みで平均した値である
多数回の測定を行い測定値の平均を求めると、期待値に近い値になる

確率変数Xの期待される値

離散型確率分布の場合：

$$E(X) = \sum_i x_i P(X = x_i) = \mu$$

連続型確率分布の場合：

$$E(X) = \int_{-\infty}^{\infty} x_i P(X = x_i) = \mu$$

分散(Variance)は、確率変数が期待値からどれだけ散らばっているかを示す指標

確率変数Xの期待されるばらつき

$$V(X) = E((X - E(X))^2) = \sigma^2$$

各確率分布に従う確率変数Xの期待値と分散（確率関数を用い期待値と分散が計算できる）

（例）正規分布

$E(X) = \mu$　　確率変数Xはμ付近の値をとることが期待できる

$V(X) = \sigma^2$

（例）二項分布

$E(X) = np$　　確率変数Xはnp付近の値をとることが期待できる

$V(X) = np(1 - p)$

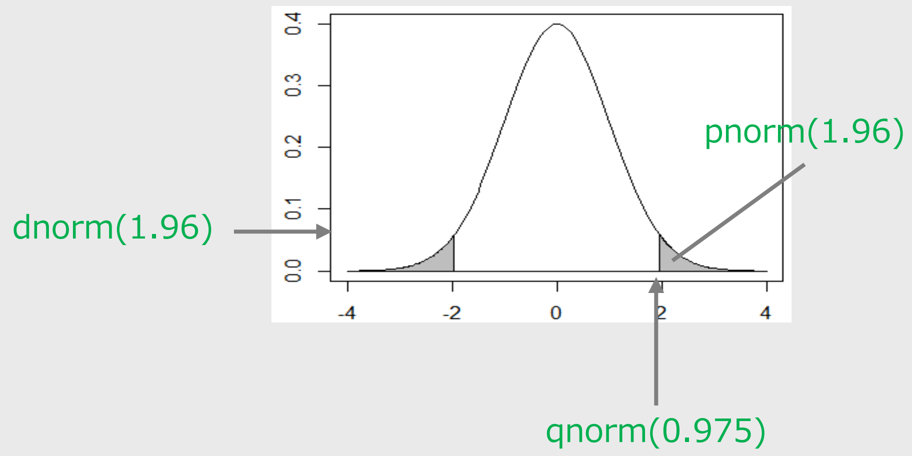
(Wiki) https://ja.wikipedia.org/wiki/%E6%9C%9F%E5%BE%85%E5%80%A4

演習【Day2-Exercise2】
確率分布の生成

- 乱数を確率分布から生成し確認
- 確率分布の生成には各確率分布に対応したパラメータを指定することを理解

演習の解説

- 正規分布の理解
  - 標準正規分布
    - 95%をカバーする範囲、分位点
    - σ×1, σ×2, σ×3がカバーする確率
  - TOEICの例
    - TOEICのスコアが、平均580、標準偏差170の正規分布に従っていると仮定する（正確には最高点、最低点が決まっているので正規分布を仮定できないが説明のため）ランダムに抽出したあるTOEIC受験者のスコアが800点以上である確率は？

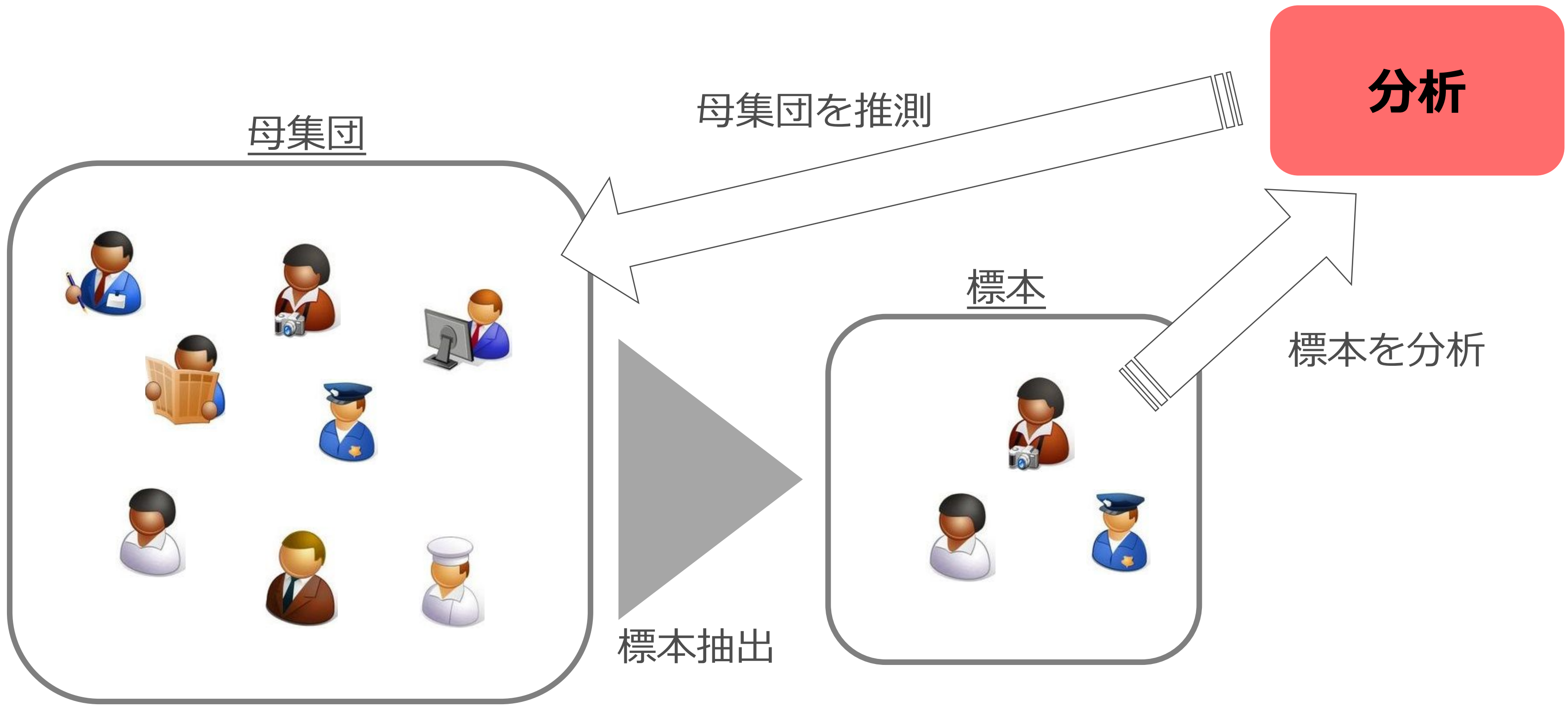


- 二項分布の理解
  - サイコロを3回振る、1回だけ6の目が出る確率は？dbinom(1,3,1/6)
  - 1回のコンタクトで契約が取れる確率が4%、100件の客先に1回ずつコンタクトした場合、5件以上契約が取れる確率は？10件以上契約が取れる確率は？

母集団と標本

母集団(Population)とは分析により知りたい対象全体
標本(Sample)とは母集団からの部分集合

分析で、標本から母集団を推測することが目的
（母集団が大きすぎるなどの何らかの理由で全数調査ができない場合）



パラメータ、点推定、区間推定

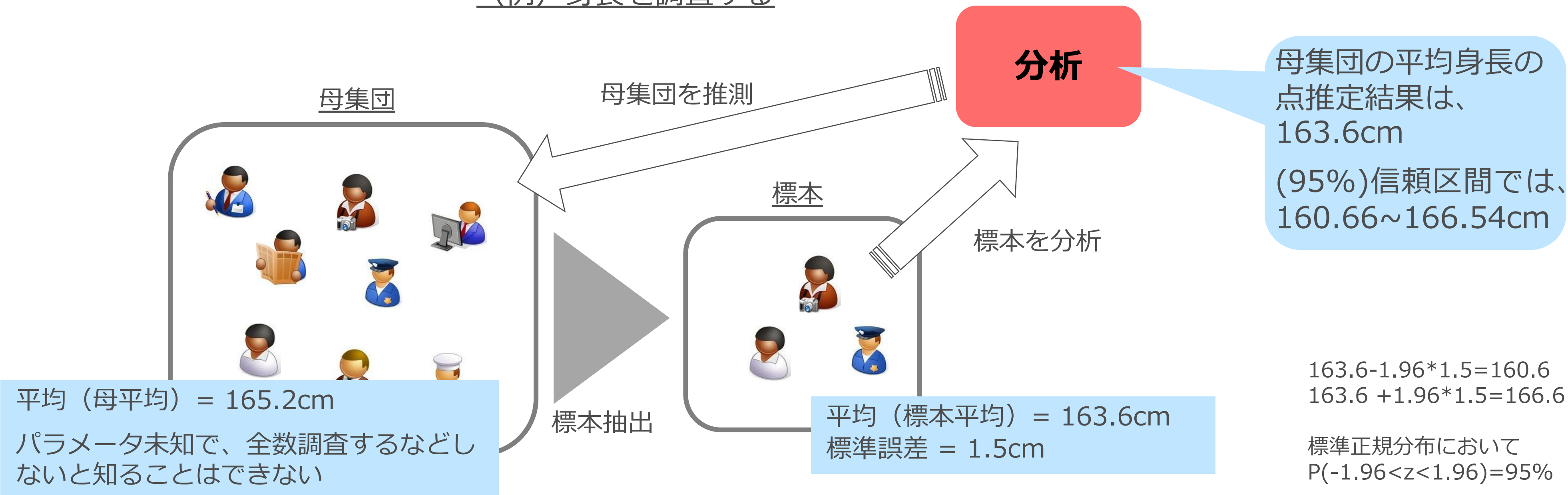
バラメータ(母数)とは母集団の特徴を説明する何らかの値（通常、未知）

標本から1つの値でパラメータを示すことを、**点推定(Point Estimate)**と言う。

また、点推定は標本からの計算なので、母集団の真の値に対し、点推定自体ばらつきを持つ。
標準誤差(Standard Error)を使い、このばらつきの大きさを示す。

点推定と標準誤差を組み合わせ、区間でパラメータを示す場合、**信頼区間(Confidence Interval)**を用いる。信頼区間を用いると、ある程度の信頼性をもって、この範囲にパラメータは収まるだろうということを示せる

（例）身長を調査する



163.6-1.96*1.5=160.6
163.6 +1.96*1.5=166.6

標準正規分布において
P(-1.96<z<1.96)=95%

母平均の推定

標本$x_1, x_2, \cdots, x_n$が、正規分布$N(\mu, \sigma^2)$に従うと仮定する（$x_1, x_2, \cdots x_n \sim N(\mu, \sigma^2)$）

母集団の平均(**母平均**)μの推定が目的

この場合、母平均の点推定は $\bar{x} = \sum_i x_i / n$（**標本平均**）となる。
また、標本平均も正規分布に従う $\bar{x} \sim N(\mu, \sigma^2/n)$

標準誤差は$\sqrt{\sigma^2/n} = \sigma/\sqrt{n}$と定義される。標準誤差とは推定値$\bar{x}$そのものの標準偏差（$\bar{x}$自体のばらつき）

<u>100(1-α)%信頼区間の公式</u>：

$$\bar{x} - z_{\alpha/2}\,\sigma/\sqrt{n} \le \mu \le \bar{x} + z_{\alpha/2}\,\sigma/\sqrt{n}$$

$z_{\alpha/2}$ … 標準正規分布の分位点

※ 95%信頼区間の場合$z_{\alpha=0.05/2} = 1.96$となる。$P(-1.96 \le z_{\alpha=0.05/2} \le 1.96) = 0.95$

ただし、母集団の標準偏差(**母標準偏差**)σは通常未知

σを標本から計算される標準偏差(**標本標準偏差**) $s = \sqrt{v} = \sqrt{\sum_i (x_i - \bar{x})^2} / \sqrt{n-1}$で代用する

<u>100(1-α)%信頼区間の公式（標本標準偏差で代用）</u>：

$$\bar{x} - t_{\alpha/2}(n-1)\, s/\sqrt{n} \le \mu \le \bar{x} + t_{\alpha/2}(n-1)\, s/\sqrt{n}$$

$t_{\alpha/2}(n-1)$ … 自由度n-1のt分布の分位点

点推定と区間推定 – 前ページの詳細

各標本が正規分布に従う $x_1, x_2, \cdots x_n \sim \mathrm{N}(\mu, \sigma^2)$

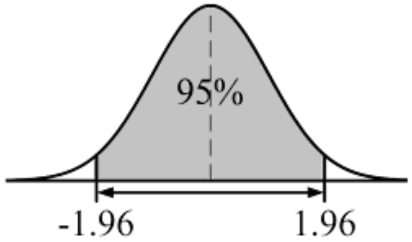
標本平均： $\bar{x} = {}^{\sum_i x_i}/_{n}$

標本平均の期待値と分散：

$$E(\bar{x}) = E\left({}^{\sum_i x_i}/_{n}\right) = {}^{1}/_{n} \sum_i E(x_i) = {}^{1}/_{n} \sum_i \mu = \mu$$

$$V(\bar{x}) = V\left({}^{\sum_i x_i}/_{n}\right) = {}^{1}/_{n^2} \sum_i V(x_i) = {}^{1}/_{n^2} \sum_i \sigma^2 = {}^{\sigma^2}/_{n}$$

よって、$\bar{x} = {}^{\sum_i x_i}/_{n} \sim \mathrm{N}(\mu, {}^{\sigma^2}/_{n})$ **真の平均μの点推定は $\bar{x}$**



平均を引き標準偏差で割り$\bar{x}$を**標準化**：$z = {}^{\bar{x}-\mu}/_{\frac{\sigma}{\sqrt{n}}} \sim \mathrm{N}(0, 1)$

$\mathrm{P}(-1.96 \leq z \leq 1.96) = 0.95$

$\Leftrightarrow \mathrm{P}(\bar{x} - 1.96 \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + 1.96 \frac{\sigma}{\sqrt{n}}) = 0.95$ **真の平均μの95%信頼区間は $[\bar{x} - 1.96 \frac{\sigma}{\sqrt{n}}, \bar{x} + 1.96 \frac{\sigma}{\sqrt{n}}]$**

<u>100(1-α)%信頼区間の公式</u>：

$$\bar{x} - z_{\alpha/2}\, {}^{\sigma}/_{\sqrt{n}} \leq \mu \leq \bar{x} + z_{\alpha/2}\, {}^{\sigma}/_{\sqrt{n}}$$

$z_{\alpha/2}$ … 標準正規分布の分位点

ただし、真の標準偏差σは通常未知
σを標本標準偏差sに置き換える

<u>100(1-α)%信頼区間の公式（標本標準偏差で代用）</u>：

$$\bar{x} - t_{\alpha/2}(n-1)\, {}^{s}/_{\sqrt{n}} \leq \mu \leq \bar{x} + t_{\alpha/2}(n-1)\, {}^{s}/_{\sqrt{n}}$$

$t_{\alpha/2}(n-1)$ … 自由度n-1のt分布の分位点

母比率の推定

標本xが、二項分布$B(n,p)$に従うと仮定する（$x \sim B(n,p)$）

母集団の比率(**母比率**)pの推定が目的

この場合、母比率の点推定は $\hat{p} = {}^{x}/_{n}$（**標比率**）となる

標準誤差は$\sqrt{{}^{\hat{p}(1-\hat{p})}/_{n}}$と定義される。標準誤差とは推定値$\hat{p}$そのものの標準偏差（$\hat{p}$自体のばらつき）

100(1-α)%信頼区間の公式：

$$\hat{p} - z_{\alpha/2}\sqrt{{}^{\hat{p}(1-\hat{p})}/_{n}} \leq p \leq \hat{p} + z_{\alpha/2}\sqrt{{}^{\hat{p}(1-\hat{p})}/_{n}}$$

$z_{\alpha/2}$ … 標準正規分布の分位点

※ 95%信頼区間の場合$z_{\alpha=0.05/2} = 1.96$となる。$P(-1.96 \leq z_{\alpha=0.05/2} \leq 1.96) = 0.95$

演習【Day2-Exercise3】 母平均、母比率の推定

- 標本から、母平均、母比率の推定を行ってみる
- 母集団はシミュレーションにより生成、標本抽出をランダムに行う（母集団と標本の違いを確認）
- 区間推定を理解する

演習の解説

ある地域の20代男性(母集団)の年収を調査する
真の分布は正規分布、母平均は300万、母標準偏差は60万とする ← 本当はわからない
全数調査できないとし、標本抽出して母集団を推測する

母集団

母集団の数 = 20,000
正規分布
- 母平均 = 300万
- 母標準偏差 = 60万

全数調査しないと本当はわからない情報
標本から平均と標準偏差を推測

抽出する標本数を10、
100、1000と変化させて、
区間推定を行う

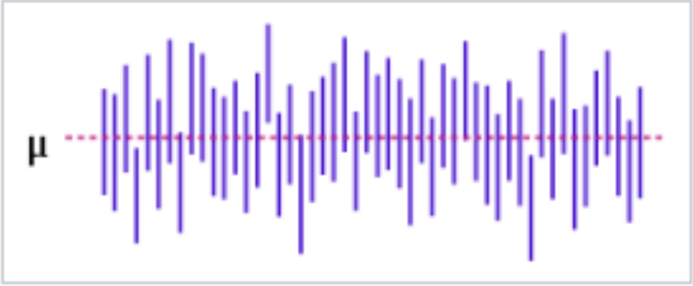
標本

標本平均 = ？
平均の95%信頼区間 = ？
標本標準偏差 = ？

信頼区間に関する注意点

95%信頼区間とは「95%の確率で真のパラメータ(母数)がその区間に入る」と解釈できるわけではない

正しくは、「**同じ状況において、観測を複数回繰り返し95%信頼区間を求めた場合、95%はその区間に真のパラメータを含む**」と、やや不自然な解釈となる



95%信頼区間の例。50の信頼区間のうち3つには母数 μ が含まれていない。

後ほど学習する、**ベイズ統計**による信頼区間(**信用区間**)を用いれば、上記の「**95%の確率で真のパラメータがその区間に入る**」と自然な解釈ができる

(Wiki) https://ja.wikipedia.org/wiki/%E4%BF%A1%E9%A0%BC%E5%8C%BA%E9%96%93

統計的仮説検定

仮説検定(hypothesis testing)とは、母集団分布のパラメータ(母数)に関する仮説を標本から検証する統計学的方法のひとつ。

統計的仮説検定においては、仮説が正しいと仮定した上で、それに従う母集団から、実際に観察された標本が抽出される確率を求め、その値により判断を行う。その確率が十分に（予め決めておいた値より）小さければ、その仮説を棄却する（すなわち仮説は成り立ちそうもないと判断する）。

手順

1. 仮説(帰無仮説)を立てる
   - **帰無仮説(H0)**：棄却できるか調べる仮説
   - **対立仮説(H1)**：帰無仮説が棄却された場合、採択
2. 仮説を正しいとして、データが観測される確率(**p値**)を計算する
3. 計算された確率に基づき、仮説を棄却するか、受容するか判断する（**有意水準**で判断）

例

新しいマーケ施策AとBの効果検証のため、100人の顧客に施策を実施し（A,B各50名ずつ）、売り上げを比較した

1. 仮説を立てる
   - **帰無仮説**：AとBに差はない
   - **対立仮説**：AとBに差はある
2. p値=0.042
3. p値が有意水準より小さいので(有意水準を0.05に設定済み)、仮説は正しいとは言えないと判断。帰無仮説を棄却

(Wiki) https://ja.wikipedia.org/wiki/%E4%BB%AE%E8%AA%AC%E6%A4%9C%E5%AE%9A

2種類の誤り

帰無仮説が正しいときに、これを棄却してしまう誤りを**第1種の誤り(Type I Error)**という。これを犯す確率をαで表す。このαを**危険率**とも呼び有意水準に等しい。

誤った帰無仮説を棄却しない誤りのことを**第2種の誤り(Type II Error)**という。これを犯す確率をβで表す。 **1 - β**は検定力あるいは**検出力(Power)**と呼び、誤った帰無仮説を正しく棄却できる確率となる。

	検定でH0を受容	検定でH1を受容
H0が真実	正しい判断	**第1種の誤り**
H1が真実	**第2種の誤り**	正しい判断

(Wiki) https://ja.wikipedia.org/wiki/%E4%BB%AE%E8%AA%AC%E6%A4%9C%E5%AE%9A

母平均の検定

標本$x_1, x_2, \cdots x_n$が、正規分布$\mathrm{N}(\mu, \sigma^2)$に従うと仮定する（$x_1, x_1, \cdots x_1 \sim \mathrm{N}(\mu, \sigma^2)$）

母平均μがμ_{H0}かどうか検定（**H0: $\boldsymbol{\mu}$ =** μ_{H0} ）

標本平均：$\bar{x} = {\sum_i x_i}/_n \sim \mathrm{N}(\mu, {\sigma^2}/_n)$

標本標準偏差：$s = \sqrt{{\sum_i (x_i - \bar{x})^2}/_{n-1}}$

100(1-α)%信頼区間の公式（標本標準偏差で代用）：

$\bar{x} - t_{\alpha/2}(n-1)\ {s}/_{\sqrt{n}} \leq \mu \leq \bar{x} + t_{\alpha/2}(n-1)\,{s}/_{\sqrt{n}}$　　　$t_{\alpha/2}(n-1)$ ⋯ 自由度n-1のt分布の分位点

確率で表記しなおすと

$$P\left(\bar{x} - t_{\alpha/2}(n-1)\ {s}/_{\sqrt{n}} \leq \mu \leq \bar{x} + t_{\alpha/2}(n-1)\,{s}/_{\sqrt{n}}\right) = 1 - \alpha$$

$$\Leftrightarrow P\left(-t_{\alpha/2}(n-1)\ \leq {\mu - \bar{x}}/_{{s}/_{\sqrt{n}}} \leq t_{\alpha/2}(n-1)\right) = 1 - \alpha$$

$$\Leftrightarrow P\left({\mu - \bar{x}}/_{{s}/_{\sqrt{n}}} \leq -t_{\alpha/2}(n-1),\ t_{\alpha/2}(n-1) \leq {\mu - \bar{x}}/_{{s}/_{\sqrt{n}}}\right) = \alpha$$

μをμ_{H0}に置き換えた値$\left|{\mu_{H0} - \bar{x}}/_{{s}/_{\sqrt{n}}}\right|$が**t-値**と呼ばれ（通常、絶対値をとる）、自由度n-1のt分布に従う

自由度n-1のt分布において、このt-値が観測される確率が**p値**となる

1. 検定の統計量（t-値）を計算
2. p値の計算：対応する確率分布（t分布）における、1で求めた統計量が観測される確率
3. 仮説の棄却か受容か判断

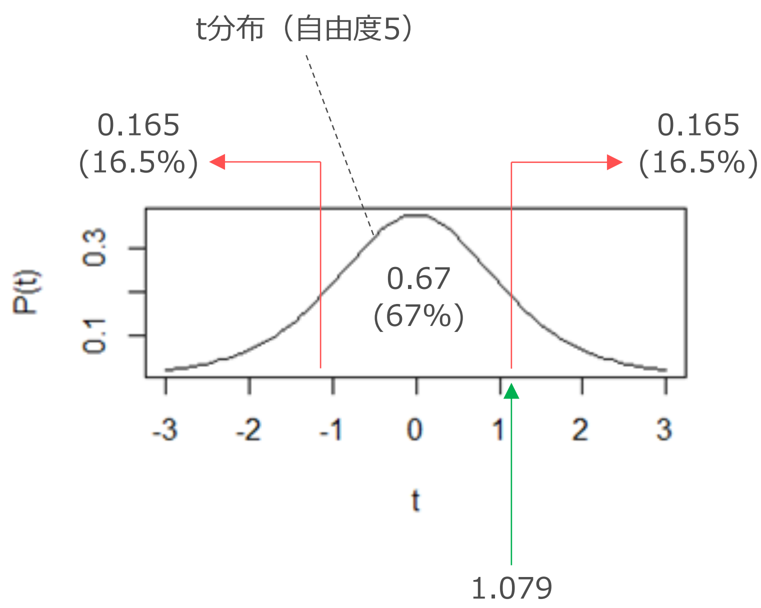
母平均の検定 – 例

あるポテトチップスの内容量が60gとの表記がある。
本当に60gか調べるため、6袋購入し実際に重さを測ってみて、以下の結果を得た。
62.1, 59.7, 60.5, 57.5, 64.6, 62.0

1. 検定の統計量（t-値）を計算
2. p値の計算：対応する確率分布（t分布）における、1で求めた統計量が観測される確率
3. 仮説の棄却か受容か判断

平均（$\bar{x}$） = 61.07
標準偏差（s） = 2.42
仮説（μ_{H0}） = 60

1. t-値（ $\left|\frac{\mu_{H0}-\bar{x}}{s/\sqrt{n}}\right|$ ） = 61.7-60 / 2.42/2.45 = 1.079
2. p値（$P\left(t(n-1) \leq \left|\frac{\mu-\bar{x}}{s/\sqrt{n}}\right|\right)$） = 0.165*2 = 0.33
3. p値が有意水準より大きいので(有意水準を0.05に設定済み)、帰無仮説を受容




```
> x <- c(62.1, 59.7, 60.5, 57.5, 64.6, 62.0)
> t.test(x, mu=60)

        One Sample t-test

data:  x
t = 1.0787, df = 5, p-value = 0.33
alternative hypothesis: true mean is not equal to 60
95 percent confidence interval:
 58.52481 63.60853
sample estimates:
mean of x
 61.06667
```

母平均の差の検定

標本$x_{1,1}, x_{1,2}, \cdots x_{1,n}$が、正規分布$N(\mu_1, \sigma^2)$に従うと仮定する（$x_{1,1}, x_{1,2}, \cdots x_{1,n} \sim N(\mu_1, \sigma^2)$）
標本$x_{2,1}, x_{2,2}, \cdots x_{2,m}$が、正規分布$N(\mu_2, \sigma^2)$に従うと仮定する（$x_{2,1}, x_{2,2}, \cdots x_{2,m} \sim N(\mu_2, \sigma^2)$）
※ 共通の分散を仮定

母平均μ_1とμ_2に差があるかの検定（**H0：$\boldsymbol{\mu_1 - \mu_2 = 0}$**）

t-値：$\left| {}^{(\overline{x_1} - \overline{x_2}) - \mu_{H0}} / {}_{\sqrt{1/n + 1/m}\,\hat{\sigma}} \right|$ （自由度n+m-2のt分布に従う）

ここで$\hat{\sigma}^2 = {}^{\sum(x_{1,i} - \overline{x_1})^2 + \sum(x_{2,i} - \overline{x_2})^2} / {}_{n+m-2}$は標本分散、 $\overline{x_1}, \overline{x_2}$は各標本平均、 μ_{H0}は仮定する両群の平均の差（通常0）

p値：$P\left(t(n + n - 1) \leq \left| {}^{(\overline{x_1} - \overline{x_2}) - \mu_{H0}} / {}_{\sqrt{1/n + 1/m}\,\hat{\sigma}} \right|\right)$

1. 検定の統計量（t-値）を計算
2. p値の計算：対応する確率分布（t分布）における、1で求めた統計量が観測される確率
3. 仮説の棄却か受容か判断

母比率の差の検定

標本x_1が、二項分布$B(n_1, p_1)$に従うと仮定する（ $x_1 \sim B(n_1, p_1)$ ）
標本x_2が、二項分布$B(n_2, p_2)$に従うと仮定する（ $x_2 \sim B(n_2, p_2)$ ）

母比率p_1とp_2に差があるかの検定（**H0：$\boldsymbol{p_1 - p_2 = 0}$**）

z値：$\left| {(\widehat{p_1} - \widehat{p_2}) - \mu_{H0}} \Big/ {\sqrt{\widehat{p_1}(1-\widehat{p_1})/n_1 + \widehat{p_2}(1-\widehat{p_2})/n_2}} \right|$ （標準正規分布に従う）

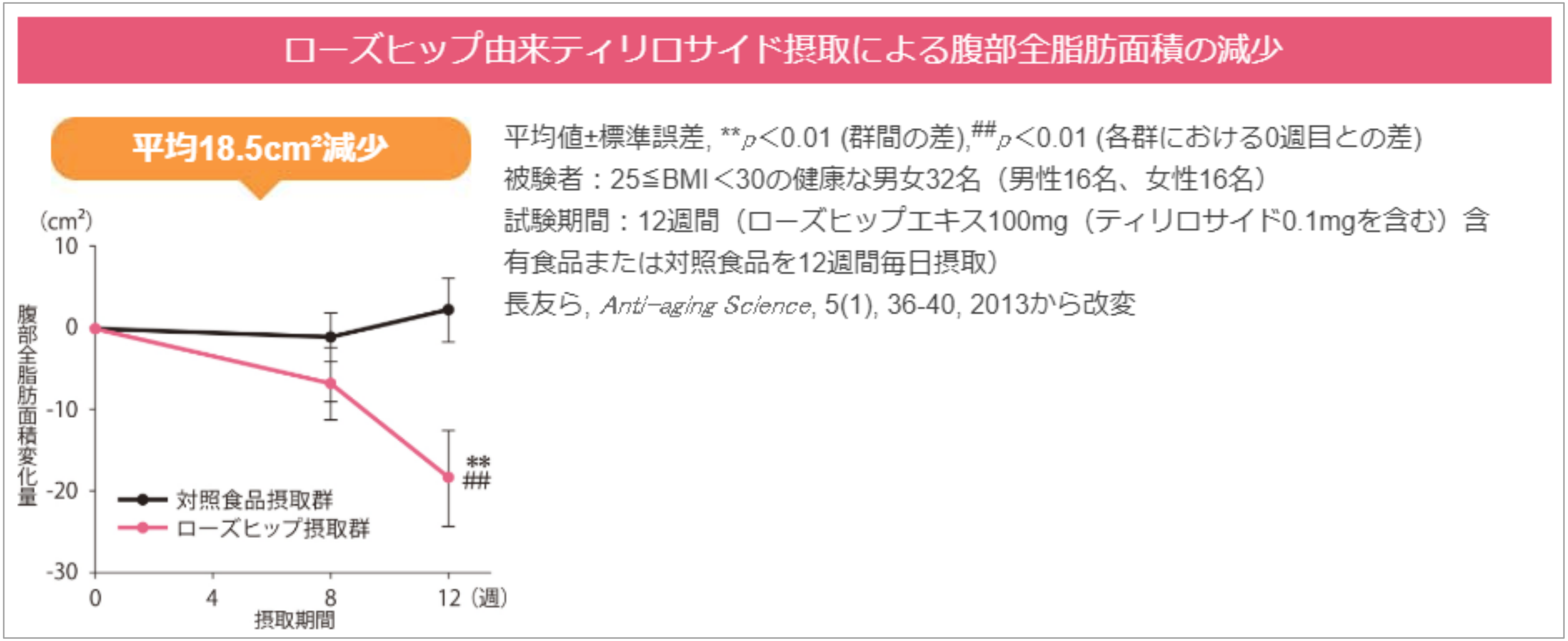
ここで$\widehat{p_1}, \widehat{p_2}$は標本比率（$\widehat{p_1} = x_1/n_1$, $\widehat{p_2} = x_2/n_2$）、 μ_{H0}は仮定する両群の平均の差（通常0）

p値：$P\left(z \leq \left| {(\widehat{p_1} - \widehat{p_2}) - \mu_{H0}} \Big/ {\sqrt{\widehat{p_1}(1-\widehat{p_1})/n_1 + \widehat{p_2}(1-\widehat{p_2})/n_2}} \right| \right)$

1. 検定の統計量（z値）を計算
2. p値の計算：対応する確率分布（標準積分布分布）における、1で求めた統計量が観測される確率
3. 仮説の棄却か受容か判断

標準誤差や検定が使われている例

"森下仁丹: ヘルスエイド®"
http://healthaid.jintan.jp/feature/rosehip.html



t.test()、prop.test()

母平均値の差の検定（2群）

t.test(グループ1, グループ2)

グループ1, グループ2のデータをベクトルで投入

例（Exercise4で実施）

新しいマーケ施策AとBの効果検証のため、20人の顧客に施策を実施し（A,B各10名ずつ）、売り上げ(円)を比較した

テスト結果

Aにおける購入平均金額: 1171.5円
Bにおける購入平均金額: 1567.0円

帰無仮説：AとBに差はない
対立仮説：AとBに差はある

母比率の差の検定（2群）

prop.test(反応数, 標本数)

反応数, 標本数をベクトルで投入

例（Exercise4で実施）

新しいマーケ施策AとBの効果検証のため、100人の顧客に施策を実施し（A,B各50名ずつ）、反応を比較した

テスト結果

Aにおける反応人数: 5
Bにおける反応人数: 13

帰無仮説：AとBに差はない
対立仮説：AとBに差はある

演習【Day2-Exercise4】
検定

- 母平均、母比率の差の検定を行ってみる
- t.test()、prop.test() の使い方を理解する
- いくつかの母集団をシミュレーションで生成、標本サイズを変えながらランダムに標本を抽出（検定により、母集団の違いを判断できるかを確認）
- 検定結果（p値）はサンプルサイズに依存することを理解する

演習の解説

シミュレーションにより、検定の性質を理解する

シミュレーションA

母集団A

正規分布
- 母平均 ＝ 30
- 母標準偏差 ＝ 10

母集団B

正規分布
- 母平均 ＝ 35
- 母標準偏差 ＝ 10

標本A

標本平均 ＝ ?

標本B

標本平均 ＝ ?

検定

母平均、標本数を変えながらシミュレーションを実施

実際に母平均に差がある場合に、検定で棄却できるかを確認する
- 実際に差があっても、サンプル数が少ない場合は棄却できない
- 小さい差を検定により棄却する場合、沢山のサンプル数が必要

シミュレーションB

母集団

正規分布
- 母平均 ＝ 30
- 母標準偏差 ＝ 10

標本A

標本平均 ＝ ?

標本B

標本平均 ＝ ?

検定

母平均、母標準偏差、標本数を変えながらシミュレーションを実施

実際に母平均に差がない場合に、検定で棄却してしまう場合があるかを確認する
- 実際に差がないが、棄却してしまう場合がある（第1種の誤り(Type I Error)）。誤って棄却してしまう割合は有意水準に等しい

p値に関する注意点

p値に関する12の誤解

Goodman, S. (2008), A Dirty Dozen: Twelve P-Value Misconceptions

Table 1 Twelve *P*-Value Misconceptions

1	*If P = .05, the null hypothesis has only a 5% chance of being true.*
2	*A nonsignificant difference (eg, P ≥.05) means there is no difference between groups.*
3	*A statistically significant finding is clinically important.*
4	*Studies with P values on opposite sides of .05 are conflicting.*
5	*Studies with the same P value provide the same evidence against the null hypothesis.*
6	*P = .05 means that we have observed data that would occur only 5% of the time under the null hypothesis.*
7	*P = .05 and P ≤.05 mean the same thing.*
8	*P values are properly written as inequalities (eg, "P ≤.02" when P = .015)*
9	*P = .05 means that if you reject the null hypothesis, the probability of a type I error is only 5%.*
10	*With a P = .05 threshold for significance, the chance of a type I error will be 5%.*
11	*You should use a one-sided P value when you don't care about a result in one direction, or a difference in that direction is impossible.*
12	*A scientific conclusion or treatment policy should be based on whether or not the P value is significant.*

ASA(American Statistical Association)による統計的有意性、p値に関する声明

（原文）
https://www.amstat.org/asa/files/pdfs/P-ValueStatement.pdf
（訳）
http://biometrics.gr.jp/news/all/ASA.pdf