母集団推定

心理データ解析演習 2010/06/16

M1 後藤 崇志

はじめに

- 心理学の研究では、人の一般的な心の働きに対する知見を求める
  - しかし、人全体を対象とすることはできず、一部の標本からのデータしか扱うことができない
  - 確率論的な分析を行い、一般化された知見を述べることになる
- 「標本データから母集団パラメータを推論する」ときに何を行っているのかを、理論的な分布を用いた計算から理解する

発表の流れ

- 基礎事項
- 母集団推定（点推定）
- 母数と標本統計量の検定
  - χ^2 分布
  - t 分布
- 母集団推定（区間推定）
- 推定の注意点

基礎事項

推定と検定

- 統計的推定
  - 母集団からとりだした標本の観測値から、未知である母集団の特徴を推し量ること
- 統計的検定
  - 母集団に関する何らかの仮説の真偽を実際の観測値に基づいて判断すること
  - 立証したい仮説と反する帰無仮説(H_0)を確率的基準に従って棄却することで、立証したい仮説(＝対立仮説; H_1)を採択する

確率変数

- 確率変数
  - 数学的な変数であるとともに、とりうる値に確率が付随したもの
  - $\Pr(X = x)$ は、確率変数 X が実測値 x をとるときの確率を表す
  - $\Pr(a \leq X \leq b)$ は、確率変数 X が a 以上 b 以下の値をとる確率を表す

- 離散型確率変数
  - 実測値が整数のように飛び値のもの（ex. サイコロの目）
- 連続型確率変数
  - 実測値が実数のもの（ex.円の中心に立てたペンを無作為に倒した時に、基準線とペンが作る角度の大きさ）
  - 以下は連続型確率変数を用いて説明する

確率分布

- 連続型確率変数の確率の求め方
  - 連続型では1点 xをとる確率は0
  - 確率変数 X が区間 (a, b) に入る確率が以下のように関数 $f(x)$ の定積分で与えられる

$$\Pr(a < X < b) = \int_a^b f(x)dx$$

  - このとき、関数 $f(x)$をXの確率密度関数という

- 確率変数 X が x 以下の値をとる確率は…

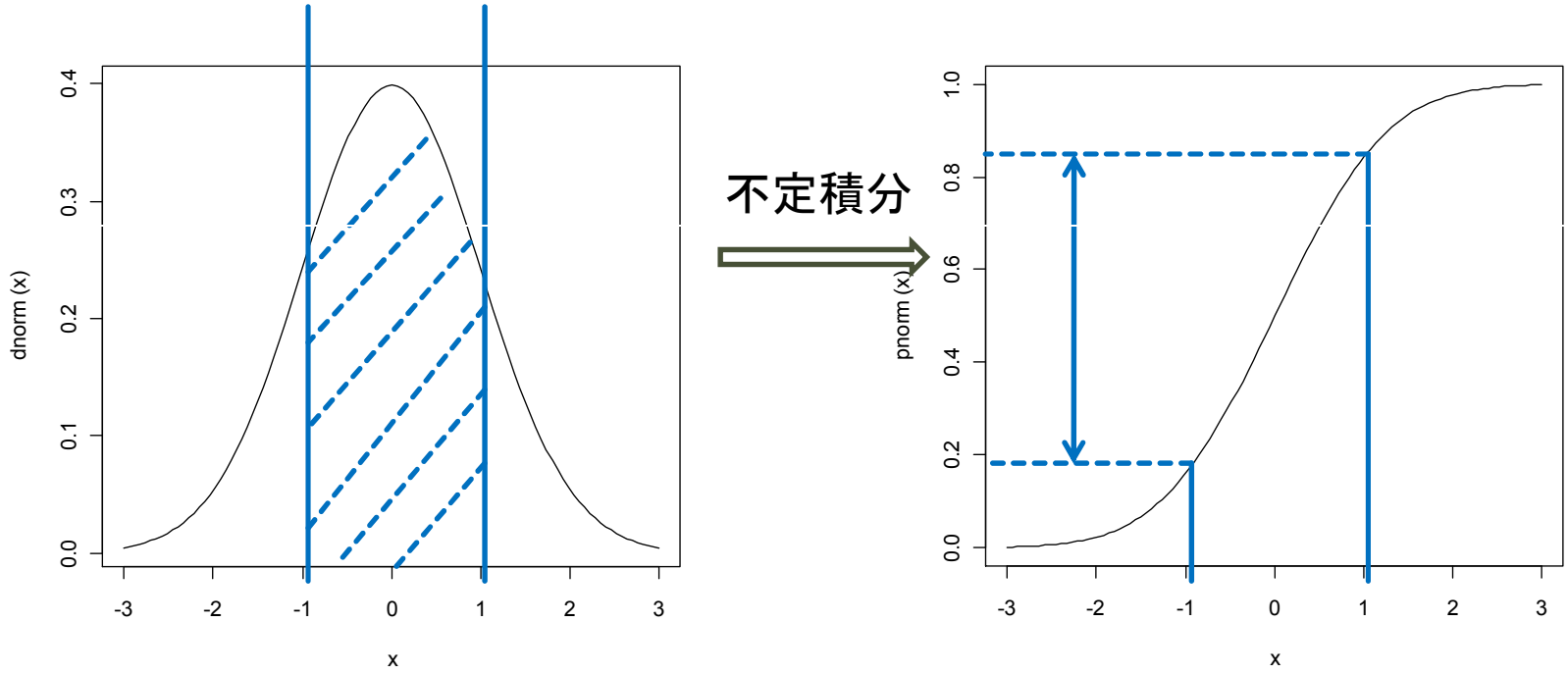
$$F(x) = \Pr(X \le x) = \int_{-\infty}^{x} f(x)dx$$

  - このとき、関数 $F(x)$を X の累積分布関数という
  - 確率変数 X が区間(a, b) に入る確率は、以下のように求められる

$$\Pr(a < X < b) = \int_a^b f(x)dx = F(b) - F(a)$$

確率分布

- 確率密度関数
- 累積分布関数



$\Pr(-1 < x < 1)$ をグラフ上に示すと・・・

#01

正規分布

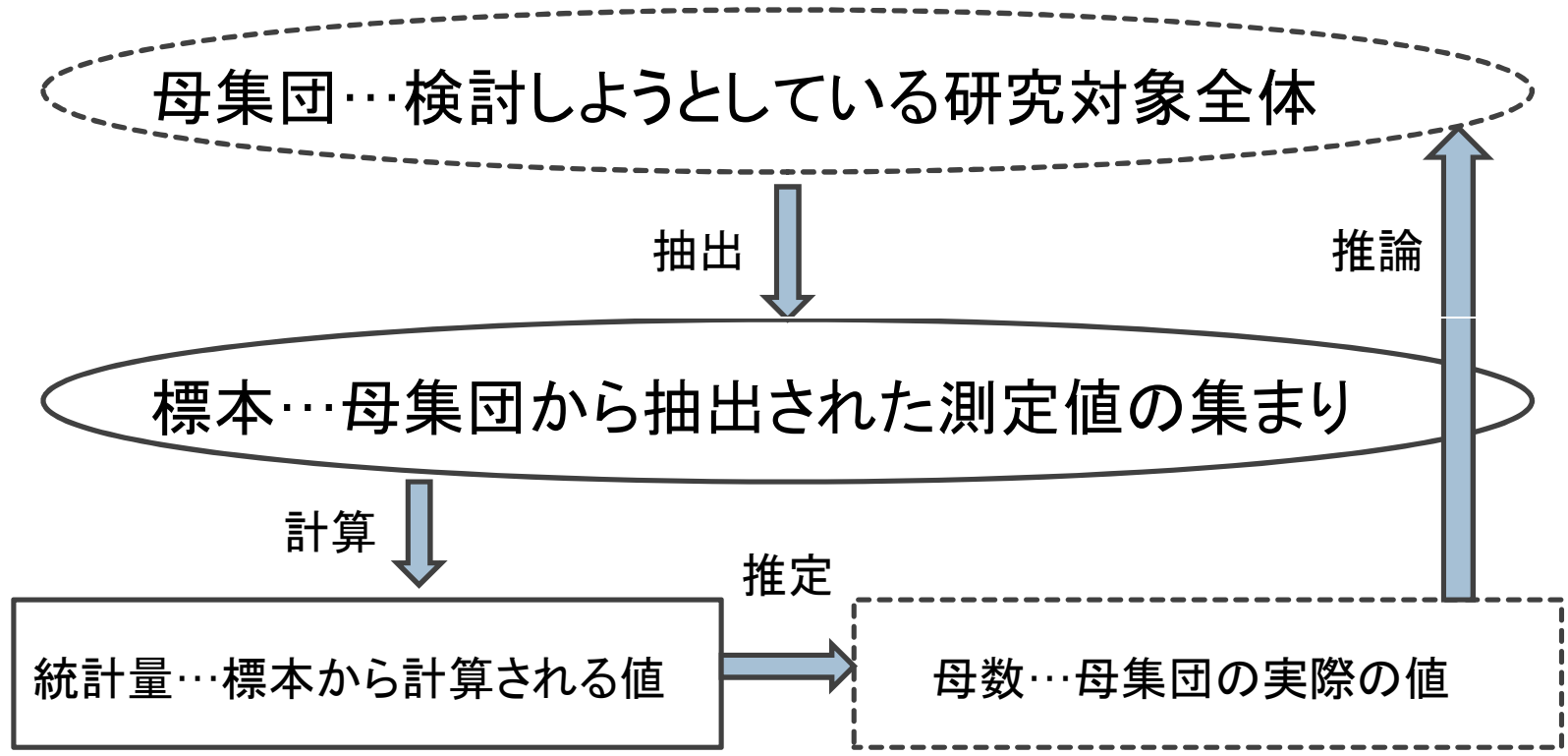
- 左右対称の、釣鐘状、単峰状の分布
- 確率密度関数による定義式

$$f(X)=\frac{1}{\sqrt{2\pi}s}e^{-\frac{1}{2}(\frac{X-\overline{X}}{s})^2} \qquad (-\infty<X<\infty)$$

- 上式の積分で変数 X が任意の a から b までの値をとる確率を求められる
- 正規分布の加法性
  - ともに正規分布に従う2群の集合 (X, Y) のそれぞれから無作為に1つずつ数値を取り出した時、その和および差は正規分布に従う

母集団の推定（点推定）

母集団と標本


母集団…検討しようとしている研究対象全体
抽出
推論
標本…母集団から抽出された測定値の集まり
計算
推定
統計量…標本から計算される値
母数…母集団の実際の値

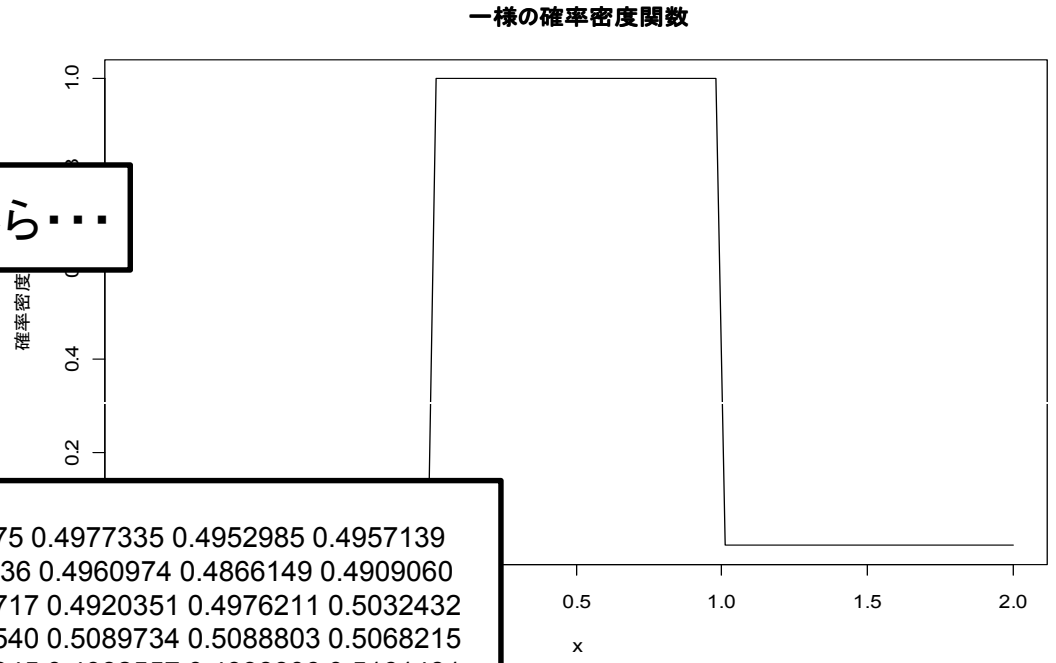
標本平均の分布

- 標本平均 $(\overline{X})$
  - 母集団から取り出したある標本の測定値の平均
  - 標本平均は母平均と常に完全に一致するか？

- **標本抽出を多数回それぞれ独立に繰り返すと…**
  - $(\overline{X})$は、$E(\overline{X})=\mu$を中心にある程度の分散で分布(大数の法則)
  - 標本の大きさがかなり大きければ母集団の分布に関係なく、標本平均の分布は正規分布に近似する(中心極限定理)
  - このとき、分散は、$\sigma_{\overline{X}}^2=\frac{\sigma^2}{n}$で表される

$$\lim_{n\to\infty}\frac{\sigma^2}{n}=?$$

一様分布から標本平均を測定

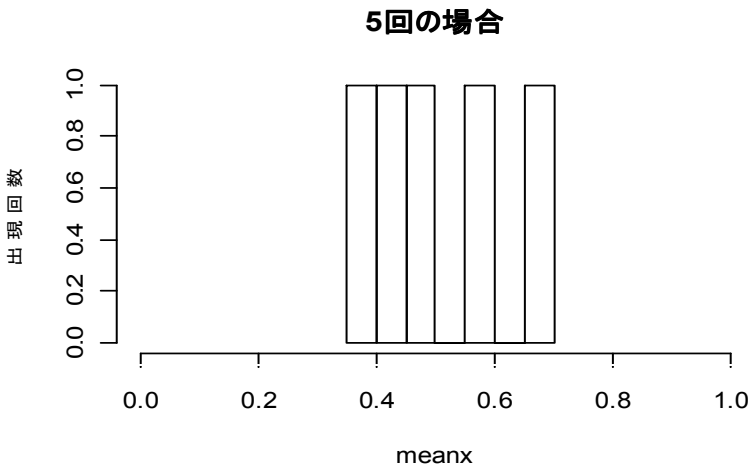


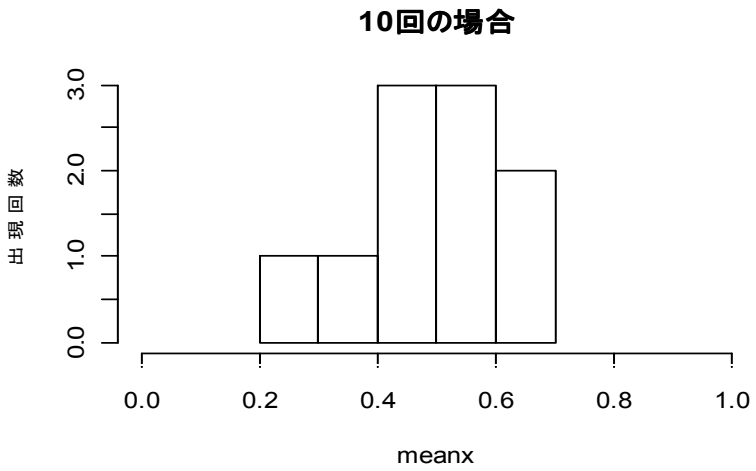
こんな分布をする母集団から・・・

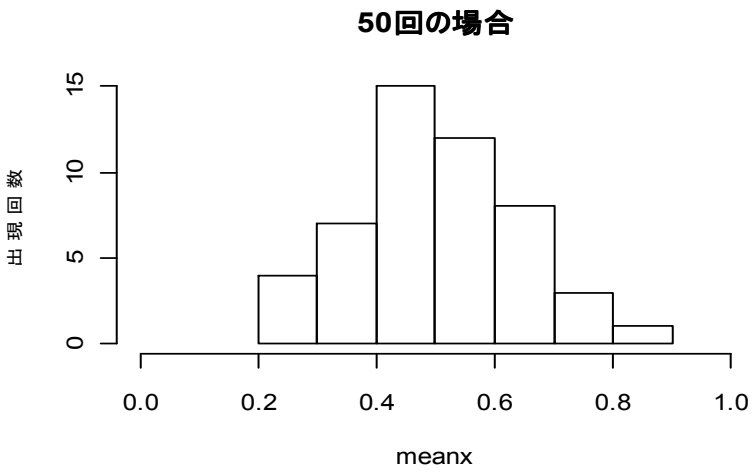
[1] 0.5080536 0.4889859 0.4979629 0.4866375 0.4977335 0.4952985 0.4957139
[8] 0.4768661 0.4938351 0.5120405 0.4822236 0.4960974 0.4866149 0.4909060
[15] 0.5072441 0.5153962 0.4964186 0.4996717 0.4920351 0.4976211 0.5032432
[22] 0.4963519 0.4897058 0.4932324 0.5091540 0.5089734 0.5088803 0.5068215
[29] 0.4938040 0.4987322 0.5004815 0.4899945 0.4982557 0.4938806 0.5161481
[36] 0.5017116 0.5000565 0.5051392 0.5168350 0.4970024 0.4901384 0.4863864
[43] 0.4947760 0.5091333 0.5013613 0.5054796 0.4934738 0.49
[50] 0.4924709 0.5082621 0.5044378 0.5082521 0.4919408 0.49
[57] 0.4866730 0.5100162 0.5090684 0.5017759 0.5038736 0.50
[64] 0.4934492 0.4967303 0.4867970 0.4869280 0.5037817 0.50
[71] 0.4950797 0.5038966 0.5091013 0.5088135 0.4987431 0.48
[78] 0.4834891 0.4822196 0.4936611 0.5137563 0.5025963 0.50
[85] 0.4985442 0.5078990 0.4803685 0.5153157 0.4870976 0.51
[92] 0.5035083 0.5131294 0.5018635 0.5059251 0.4971204 0.4999740 0.5105453
[99] 0.4913706 0.4731109

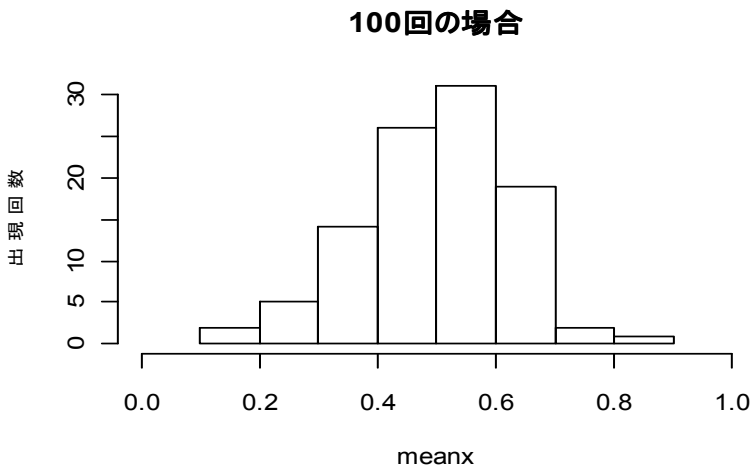
n=100の標本を100回独立に抽出して
それぞれの標本平均を求めてみた

標本平均のヒストグラム
(n =5)

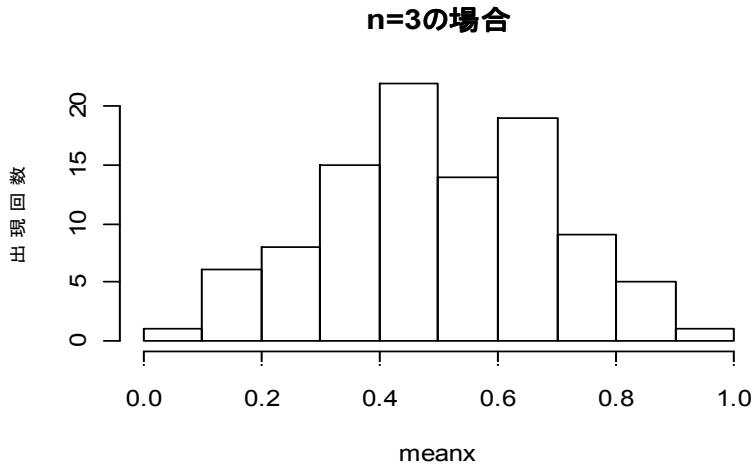


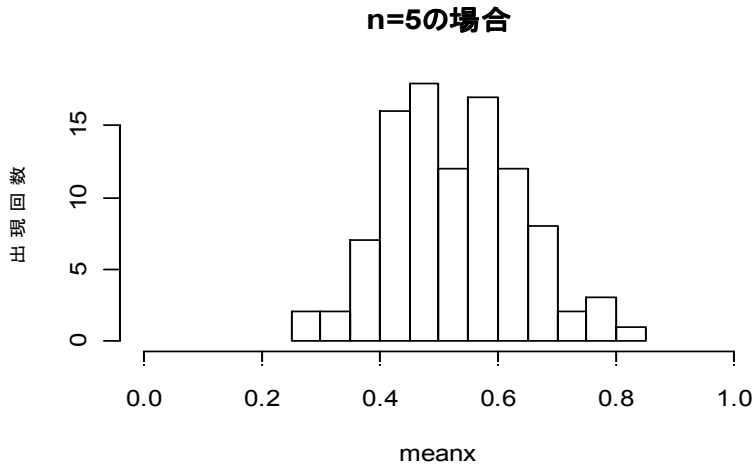


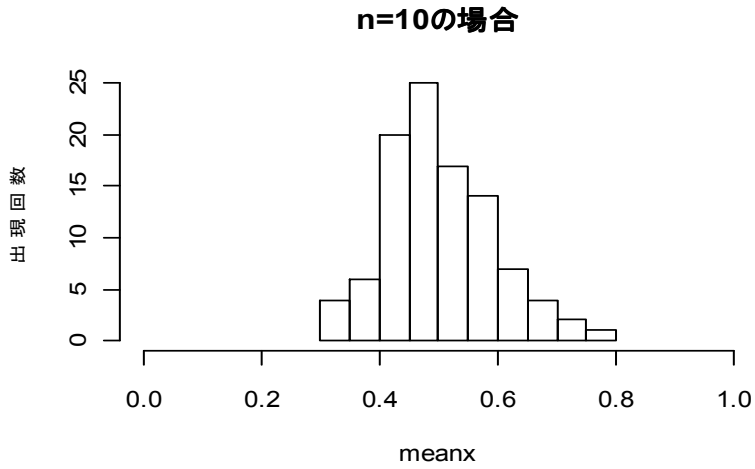


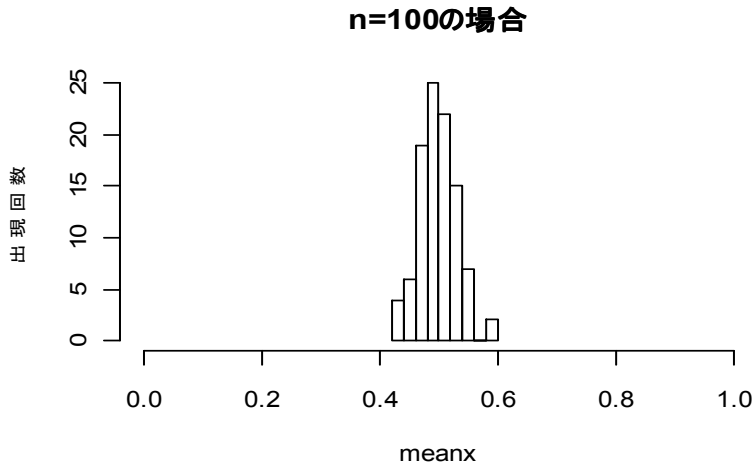


標本平均のヒストグラム (抽出回数=100)









母平均の推定(点推定)

- 標本平均の期待値は母平均に一致する($E(\overline{X}) = \mu$)
  - 統計量の期待値が対応する母数に一致するので、標本平均は母平均の不偏推定値である

- 標本平均の推定値は標本分布の分散の大きさに応じて誤差が伴う
  - 標本分布の標準偏差を標準誤差として指標にする
  - 先のスライドより

$$SE = \frac{\sigma}{\sqrt{n}}$$

母分散の推定（点推定）

- 母分散の不偏推定値

$$\sigma^2 = \frac{\sum_i^n (X_i - \overline{X})^2}{n} + \frac{\sigma^2}{n}$$

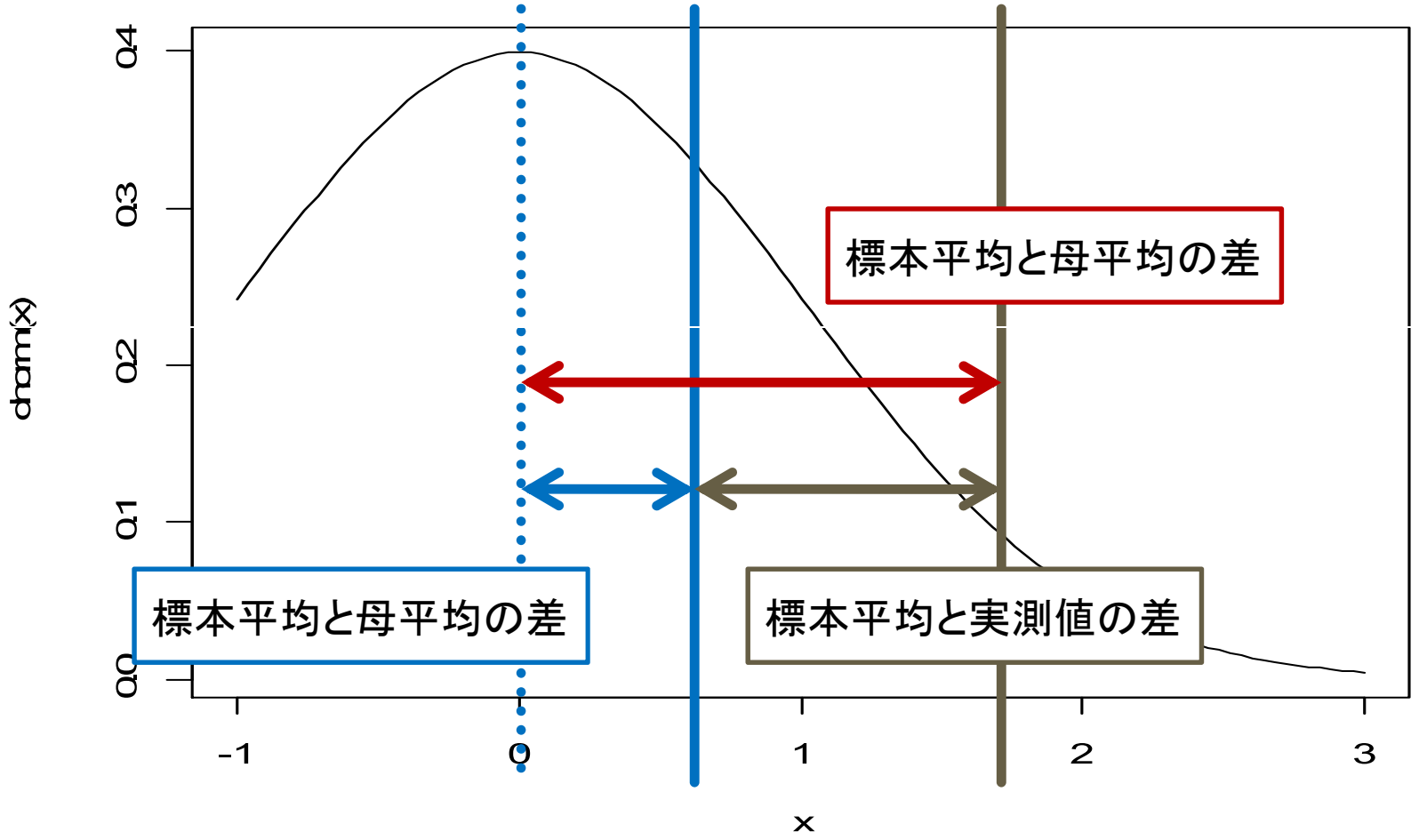
  - 母分散は個々の測定値の母平均(μ)からの偏差の2乗値の平均
  - 母平均の代わりに標本平均$(\overline{X})$を用いるが、標本平均は母平均を中心にある程度の分散をもって分布
  - よって、母分散＝標本の分散+標本平均の分散
  - この式を変形すると…

$$\hat{\sigma}^2 = \frac{\sum_i^n (X_i - \overline{X})^2}{n-1}$$

  - これを用いると、標準誤差の推定値は…

$$SE = \frac{\hat{\sigma}}{\sqrt{n}}$$

母分散の推定（点推定）


dnorm(x)
0.0
0.1
0.2
0.3
0.4
-1
0
1
2
3
x
標本平均と母平均の差
標本平均と母平均の差
標本平均と実測値の差

母数と標本統計量の検定

標準得点と標準正規分布

- 標準得点
  - 変数の尺度を変換して平均値や特定の値になるようにした（標準化）新しい尺度上での各測定値の得点
- z 得点

$$z = \frac{X_i - \overline{X}}{s}$$

  - 平均値が0、標準偏差が1になるように標準化された得点（以下、標準得点は z 得点を指すものとする）

- 標準正規分布
  - 平均値が0、標準偏差が1であるような正規分布
  - 標準化を行っても、確率変数の分布のかたちは変化しないため、もとの確率変数が正規分布に従うならば、標準得点（ z ）は標準正規分布に従う

標準得点による相対位置の推定

- 測定値の相対位置の測定
  - 測定値の標準得点から、標準正規分布を利用して、測定値の相対位置を知ることができる
- 標準得点が z_0をとる場合…
  - $P_{(z \geq z_0)} = p_{z_0}$ は、z_0 に対応する測定値が上位 $p_{z_0} \times 100$%の値であることを示す
  - $P_{(|z| \geq z_0)} = p_{z_0}$ は、値が平均値から標準偏差の z_0倍以上離れている値は全体の $p_{z_0} \times 100$%あることを示す
- 標準得点は、分布全体における測定値の相対的位置を表す測度

ある条件の平均値と母平均との差の検定（母分散が既知）

#07

- ある条件の平均値が母平均(μ)と異なるか
  - 標本平均($\overline{X}$)は、不偏推定値なので、ある条件の母平均は$\overline{X}$であると推定される
  - 帰無仮説は$H_0 : \hat{\mu} = \mu$、対立仮説は$H_1 : \hat{\mu} \neq \mu$
  - 標本平均の分布は、$\overline{X} \sim N(\mu, \quad \sigma^2 / n)$
  - よって標本平均を標準化すると・・・

$$z = \frac{\overline{X} - \mu}{\sqrt{\sigma^2 / n}}$$

  - ここから、$\overline{X}$の、標本平均の分布における相対位置を推定し、標本平均が$\overline{X}$となる確率を求める

＝標準得点と、設定した有意水準に対応する臨界値と比較する

χ^2 分布

- 自由度nの χ^2分布の定義式
  - 正規分布より大きさ1の標本を独立に n回取り出して標準化した値の総和

  ＝互いに独立な、標準正規分布に従う n個の確率変数の総和

$$\chi^2_{(n)} = \frac{\sum_i^n (X_i - \mu)^2}{\sigma^2}$$

  - 期待値はν（自由度）、分散は 2ν
  - 母平均を標本平均で代用すると、自由度は $n-1$となり…

$$\chi^2_{(n-1)} = \frac{\sum_i^n (X_i - \overline{X})^2}{\sigma^2}$$

t 分布

- 自由度 $n-1$ の t 分布の定義式
  - 標準正規分布に従う確率変数 z を分子に、自由度 $n-1$ の χ^2 分布に従う確率変数を自由度で割ったものを分母にしたもの

$$t_{(n-1)} = \frac{z}{\sqrt{\chi^2_{(n-1)} / n-1}}$$

  - この式を変形すると、分散の不偏推定値を用いて標本平均の母平均からの偏差を標準化したものになる

$$t_{(n-1)} = \frac{\overline{X} - \mu}{\sqrt{\hat{\sigma}^2 / n}}$$

  - 期待値は0、分散は $\nu > 2$ のとき $\frac{\nu}{\nu - 2}$

$$\lim_{\nu \to \infty} \frac{\nu}{\nu - 2} = ?$$

ある条件の平均値と母平均との差の検定（母分散が未知）

- ある条件の平均値が母平均 (μ)と異なるか
  - 帰無仮説、対立仮説ともに未知の場合と同じ
  - 母分散が未知なので、不偏推定量 (σ^2) を用いる
  - 標準化した得点は t 分布に従うので、これを用いて相対位置を求める

- t値を、(μ)を左辺にとって変形 $\mu = \overline{X} \pm t_{(n-1)} \frac{\hat{\sigma}}{\sqrt{n}}$

母平均の推定（区間推定）

点推定と区間推定

- 点推定
  - 母数 θ を、ひとつの値 $\hat{\theta}$ によって推定すること
  - 点推定値 $\hat{\theta}$ はn個の確率変数 $X_1, X_2, ,, X_n$ の具体的な実測値 $x_1, x_2, \cdots, x_n$ から求められたものであるため、異なる観測ごとに異なる値をとる
  - 点推定値は期待値を母数とし、ある分散をもった標本分布に従う
  - このときの分散の正の平方根が標準誤差である

- 区間推定
  - 確率値 α について、確率変数 $X_1, X_2, ,, X_n$ から $\Pr(\theta_L < \theta < \theta_H) = 1 - \alpha$ となる区間を求めたとき、この区間を信頼係数 $100 * (1 - \alpha)\%$ の信頼区間という

信頼区間について

- 通常の確率計算では…
  - $\Pr(a < X < b)$ というときに両端の aと bは定数であり、X が aからb の間の値をとることを表す
- 信頼区間においては…
  - $\Pr(\theta_L,\ \theta_U)$ の両端は確率変数であり、Xが θ_Lから θ_Uの間の値をとる確率を表すわけではない
  - $\Pr(\theta_L < \theta < \theta_U)$は「$n$個の観測値から $(\theta_L,\ \theta_U)$ という信頼区間を求める試行を多数回繰り返した場合にその区間が Xを含む確率」を表す
  - 標本数 n が大きいほど信頼区間は狭くなり、より精密な区間推定ができるようになる

母平均の推定（区間推定）

- 母平均(μ)が未知である場合
  - $t_{(n-1)}(\alpha/2)$ を自由度$n-1$のt 分布の$100*\alpha/2$%点の確率点とすると

$$\Pr\left[-\left|t_{(n-1)}(\alpha/2)\right| < \frac{\overline{X}-\mu}{\sqrt{\hat{\sigma}^2/n}} < \left|t_{(n-1)}(\alpha/2)\right|\right] = 1-\alpha$$

  - よって

$$\Pr\left[\overline{X} - \left|t_{(n-1)}(\alpha/2)\right|\frac{\hat{\sigma}}{\sqrt{n}} < \mu < \overline{X} + \left|t_{(n-1)}(\alpha/2)\right|\frac{\hat{\sigma}}{\sqrt{n}}\right] = 1-\alpha$$

  - このとき区間$\left(\overline{X} - \left|t_{(n-1)}(\alpha/2)\right|\frac{\hat{\sigma}}{\sqrt{n}},\ \overline{X} + \left|t_{(n-1)}(\alpha/2)\right|\frac{\hat{\sigma}}{\sqrt{n}}\right)$が($\mu$) の区間推定における$100*(1-\alpha)$%信頼区間となる

サンプルサイズの決定（母集団推定の場合）

- 先のスライドの区間推定の式より、$100*(1-\alpha)$％の信頼区間の幅をciとしたい場合には・・・

$$\left(\overline{X}+\left|t_{(n-1)}(\alpha/2)\right|\frac{\hat{\sigma}}{\sqrt{n}}\right)-\left(\overline{X}-\left|t_{(n-1)}(\alpha/2)\right|\frac{\hat{\sigma}}{\sqrt{n}}\right)=ci$$

  - より、

$$n=\frac{\left(2*t_{(n-1)}(\alpha/2)\right)^2}{ci^2}*\hat{\sigma}^2$$

  - 上式より、信頼区間を狭く（ciを小さく）したい場合にはnは大きくなる

推定の注意点

推定の注意点

- 確率モデルの適用には母集団からのランダムサンプリングが前提となる
  - 確率統計は、それぞれのサンプルが独立に確率的ふるまいをしていることを前提としているため
  - 扱った標本がどのような母集団から抽出されたものか
  - 得られたデータからどの程度まで一般化することができるのか
- 完全なランダムサンプリングは不可能
  - 母集団を限定し、そこから無作為に抽出されたとみなす
  - 参加者の条件間への配置や刺激の提示順序をランダム化することで代替

推定の注意点

- 母集団分布が正規分布に従うことを仮定
  - 正確な分布のかたちを求めることは不可能
  - 実際に正規分布に従うものも多いが、他の分布では確率計算が困難なために正規分布を仮定することも
- 推定・検定によって得られるのは数学的な結果
  - 研究者は結果を解釈し、それに対して心理学的意味を付与していく
  - 自分の用いた分析が、何を前提に、何を比較・分析しているのかを把握することが重要

参考文献（参考URL）

- 森敏昭、吉田寿夫（1990）心理学のためのデータ解析テクニカルブック
- 南風原朝和(2002)心理統計学の基礎
- 岩崎学（2007）確率・統計の基礎
- 青木敏伸(2009)Rによる統計解析
- 竹澤邦夫 R-tips: http://cse.naro.affrc.go.jp/takezawa/r-tips/r.html