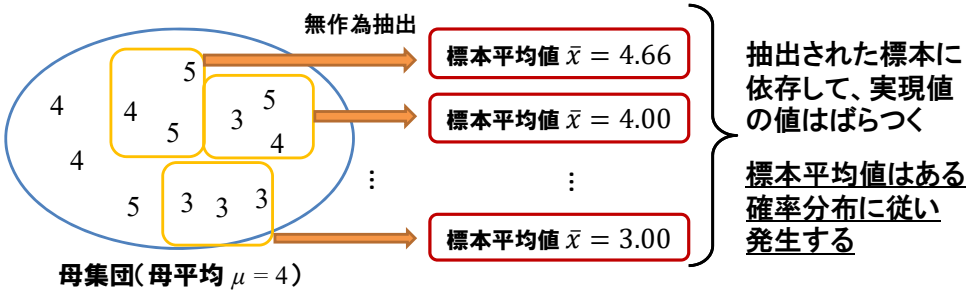
# 統計学入門 補助資料
# ～標本平均～

2022年度1学期:　月曜2限
担当教員:　石垣　司



## 標本分布

- 標本から計算される代表値(など)が従う確率分布
  - 例　$\mu = 4$ の母集団から $n = 3$ の標本抽出
    - 抽出される標本は確率的に決まる




【これ以降の授業での設定】
- 標本 $\{X_1, \dots, X_n\}$ は確率変数
- 標本は独立同一分布から無作為抽出
  - $\{x_1, \dots, x_n\}$ は確率変数の実現値(観測値・データ)



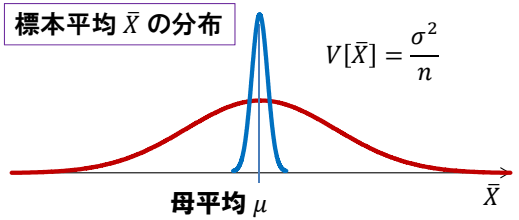
## 用語の整理

- 独立同一分布(i.i.d.)
  - 標本 $\{X_1, \dots, X_n\}$ の各 $X_i$ は独立である
  - 標本 $\{X_1, \dots, X_n\}$ は同一の母集団から抽出
- 母平均, 母分散
  - 母集団の平均 $\mu$ と分散 $\sigma^2$
- 統計量(statistic)
  - 標本から目的の値を計算する関数 $T(X_1, \dots, X_n)$
    - 統計量自体も確率変数
  - 例　標本平均の統計量　$T(X_1, \dots, X_n) = \frac{1}{n}\sum_{i=1}^{n} X_i$



## 標本平均の分布

- 標本平均　$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$
  - $\bar{X}$ は確率変数（サンプルサイズ $n$ の標本抽出を1回の試行とみなすと, 試行ごとに標本平均の実現値は確率的に決まる）
- 標本平均の期待値　$E[\bar{X}] = \mu$　check!
- 標本平均の分散　$V[\bar{X}] = \frac{\sigma^2}{n}$　check!




$n$ が大きいと分散が小さい
$|\mu - \bar{x}|$ が小さくなる確率が高い

$n$ が小さいと分散が大きい
$|\mu - \bar{x}|$ が大きくなる確率が高い

# 標本の誤差の見積もり～信頼区間 #1

- **母集団が正規分布 $N(\mu, \sigma^2)$ のときの標本平均**

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

- **例　正規母集団の母平均 $\mu$ の95%信頼区間**

$$\left[\bar{X} - 1.96\frac{\sigma}{\sqrt{n}}, \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right]$$ check!

  - **標準誤差(SE)　標本平均の分布の標準偏差 $\sqrt{V[\bar{X}]} = \frac{\sigma}{\sqrt{n}}$**
  - “1.96×標準誤差”で誤差の大きさを見積もり
    - $P(-1.96 \le Z \le 1.96) = 0.95$
    - 数値1.96は標準正規分布表より求まる

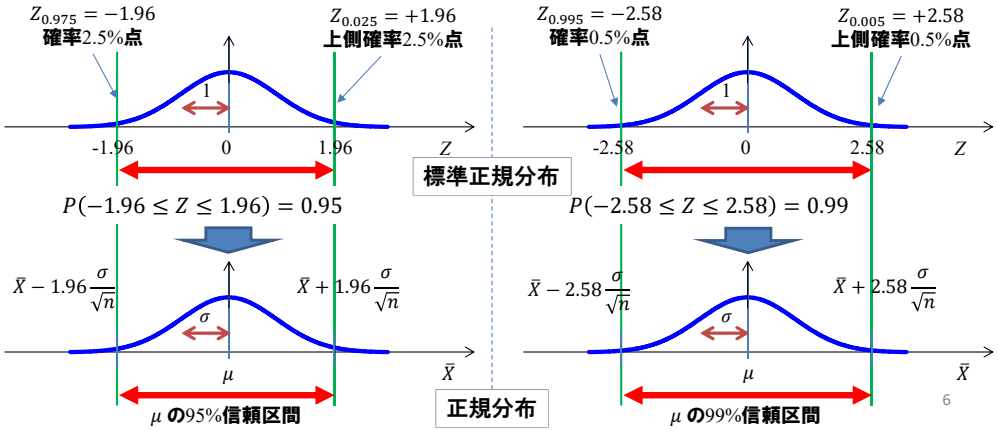


# 標本の誤差の見積もり～信頼区間 #2

- 95%信頼区間

$Z_\alpha$: 標準正規分布の上側確率 $100 \times \alpha$%点

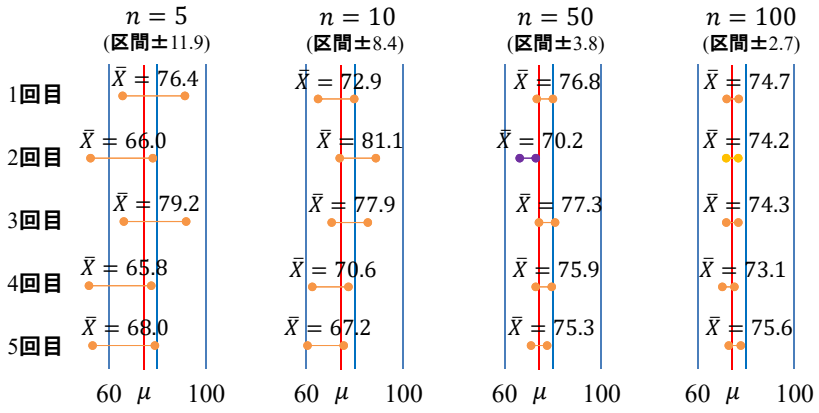
  - **サンプルサイズ $n$ の標本抽出を100回繰り返したら場合, 約95回はその区間$\left[\bar{X} - 1.96\frac{\sigma}{\sqrt{n}}, \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right]$の中に $\mu$ が含まれる**






# 標本の誤差の見積もり～信頼区間 #3

- **例　統計学の授業の成績(2015年度)の平均点の信頼区間**
  - **試験受験者数 $N = 277$, 母標準偏差 $\sigma = 13.5$, 正規母集団を仮定**
  - **復元抽出したサイズ $n$ の標本から各5回, 95%信頼区間を算出**






# 標本の誤差の見積もり～信頼区間 #3

- 95%信頼区間($\mu$ に関する標本平均)の意味・解釈
  - 【正】 サイズ $n$ の標本抽出を100回繰り返した場合, 約95回はその区間の中に $\mu$ が含まれる
  - 1つのサイズ $n$ の標本から計算された95%信頼区間を考える
  - 【誤】 その区間の中に $\mu$ が95%の確率で出現する
    - 母平均 $\mu$ は定数である(確率変数ではない)
  - 【正】 確率95%で, その区間は $\mu$ を含む
    - より正確には「その区間は $\mu$ を含む or 含まないのどちらかであり, “$\mu$ を含む”という事象が生じる確率は95%である」
  - 【誤】 その区間の中心ほど, $\mu$ に近い可能性が高い

# 有限母集団における比率調査

- **例　TV視聴率は“見ている” or “見ていない”の比率**
  - **地域世帯数は既知。どのくらいの家庭での視聴状況を調べれば, 信頼できる数値が分かるか？**
- **ベルヌーイ分布 $\mathrm{Bel}(p)$ でモデル化**
  - **サイズ $N$ の有限母集団からサイズ $n$ の標本を非復元抽出**
    - **$p$ を母比率とよぶ**
  - **有限母集団の標本平均の期待値**　$E[\bar{X}] = \mu$
  - **有限母集団の標本平均の分散**　$V[\bar{X}] = \left(\frac{N-n}{N-1}\right)\frac{\sigma^2}{n}$　check!
  - **有限母集団修正**　$c_N = \frac{N-n}{N-1}$　$(0 \le c_N \le 1)$
    - **有限母集団の $V[\bar{X}]$ ＜ 無限母集団の $V[\bar{X}]$**
    - **非復元抽出の方が $\mu$ を精度よく推定**



# 例　TV視聴率調査の誤差 #1

- **TV視聴率調査の信頼区間**
  - **関東地区の総世帯数**　$N = 20{,}125{,}319$ **世帯**
  - **関東地区の調査対象数**　$n = 900$ **世帯**
  - **有限母集団修正**　$c_N = 0.9999553 \fallingdotseq 1$
  - **標準誤差**　$SE = \sqrt{c_N \frac{\sigma^2}{n}} = \sqrt{\frac{p(1-p)}{900}}$
  - 95%**信頼区間**　$\left[\bar{X} - 1.96\sqrt{\frac{\bar{X}(1-\bar{X})}{900}}, \bar{X} + 1.96\sqrt{\frac{\bar{X}(1-\bar{X})}{900}}\right]$
    - **ここで $\bar{X}$ は標本(データ)から計算された視聴率を意味する**

**世帯数**　総務省住民基本台帳(2017年1月1日)より
**調査対象数**　ビデオリサーチ社HPより　ttps://www.videor.co.jp/tvrating/attention/index.html



# 例　TV視聴率調査の誤差 #2

- **TV視聴率調査の信頼区間**
  - **サンプルサイズ $n = 900$ の標本から計算された比率 $\bar{X}$ ごとの95%信頼区間**

| 標本平均 $\bar{X}$ データから計算した結果 | 5% | 10% | 20% | 30% | 40% | 50% |
|---|---|---|---|---|---|---|
| 95%信頼区間 | 3.6% ～ 6.4% (±1.4%) | 8.0% ～ 12.0% (±2.0%) | 17.4% ～ 22.6% (±2.6%) | 27.0% ～ 33.0% (±3.0%) | 36.8% ～ 43.2% (±3.2%) | 46.7% ～ 53.3% (±3.3%) |

  - **信頼区間の幅は標本平均の値により異なる**
  - **$p = 0.5$ ($\bar{X} = 50\%$)に近いほど信頼区間の幅(許容誤差)が大きい**



# 比率調査のサンプルサイズ $n$ の決め方

1. **信頼区間の水準**(90%, 95%, 99% **など**)**を決め, 係数 $\alpha$ を標準正規分布表から求める**
   - **二項分布の正規近似を用いることで, 比率調査の $\bar{X}$ の分布を正規分布で近似する**
     - 95%**信頼区間の場合,係数** $\alpha = 1.96$
2. **許容できる信頼区間の幅 $\pm r$ を決める**
   - **例 知りたい**95%**信頼区間が**±3%**のとき**, $r = 0.03$
3. **サンプルサイズ $n$ を決める**

$$n = \frac{1}{r^2}\alpha^2 c_N p(1-p)$$ check!

   - **母比率 $p$ に関する情報がない場合, 信頼区間の幅が最大となる $p = 0.5$ を代入する**

# 演習問題

1. **仙台市のTV視聴率調査のサンプルサイズは200世帯とする。あるTV番組の視聴率が20%と算出されたとき, その視聴率の95%信頼区間を求めよ**
   - **仙台市の世帯数は** $N = 532{,}654$ **(2022年4月1日)**
   - $n = 200$ **のときの有限母集団修正** $c_N = 0.99960 \cong 1$
   - **ここで、** $1.96 \cong 2,\ c_N \cong 1, \sqrt{0.02} \cong 0.14$**として計算**

2. **仙台市の各世帯でのペット飼育率を調べたい。実際の飼育率に関わらず95%信頼区間の幅を** $\pm 2\%$ **以下とするためのサンプルサイズ** $n$ **を答えよ**
   - **「実際の飼育率に関わらず」であるので, 信頼区間の幅が最大となる** $p = 0.5$ **を代入する**

**世帯数　仙台市HP(2022年4月1日)より**
**調査対象数　ビデオリサーチ社HPより(2017年時に記載あり。現在では削除)**