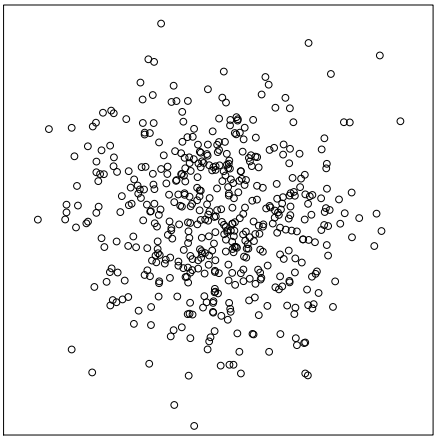
空間モデリング

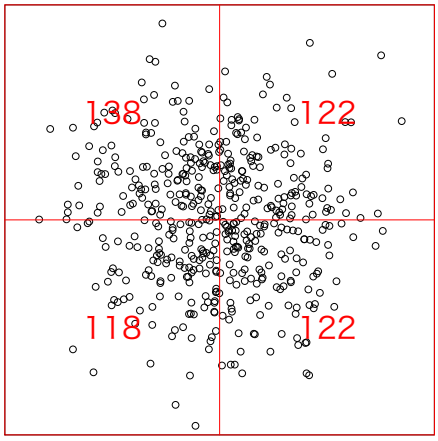
古谷知之

講義内容

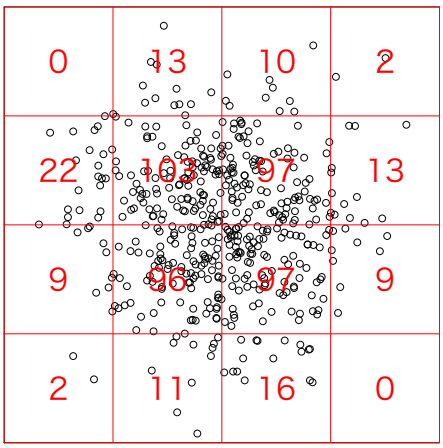
- 空間集計単位問題：MAUP
- 平均・分散・標準偏差、標準化
- コルモゴロフ・スミルノフ検定
- 等分散性の検定
- 平均値の差の検定
- ジニ係数とローレンツ曲線
- 変動係数
- 地域特化係数

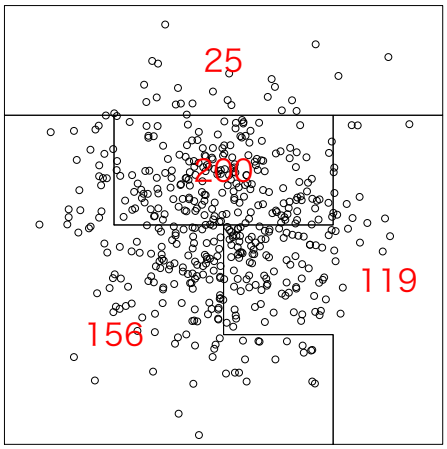
空間集計単位問題



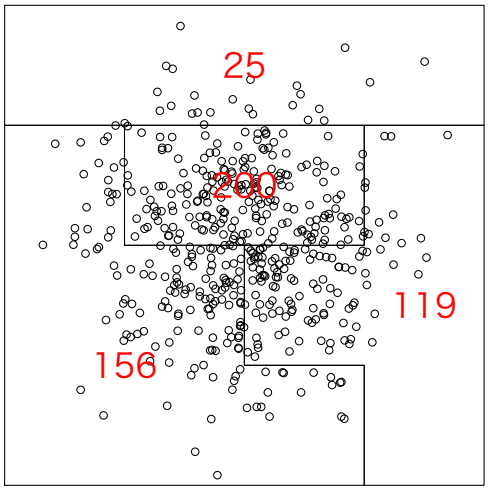


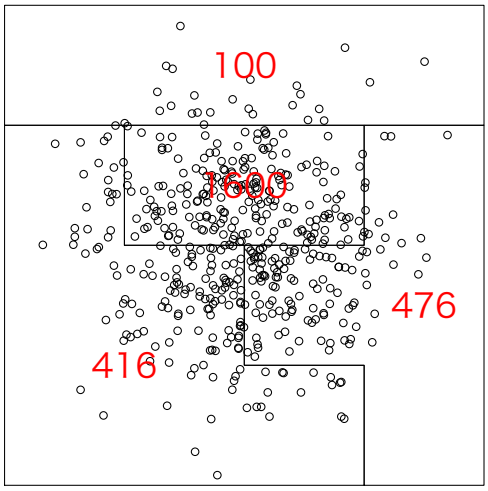
空間集計単位問題



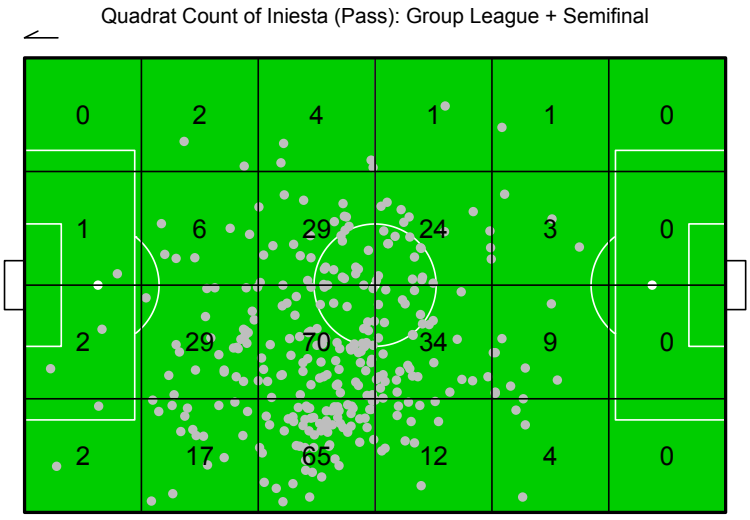


度数（左）と密度（右）

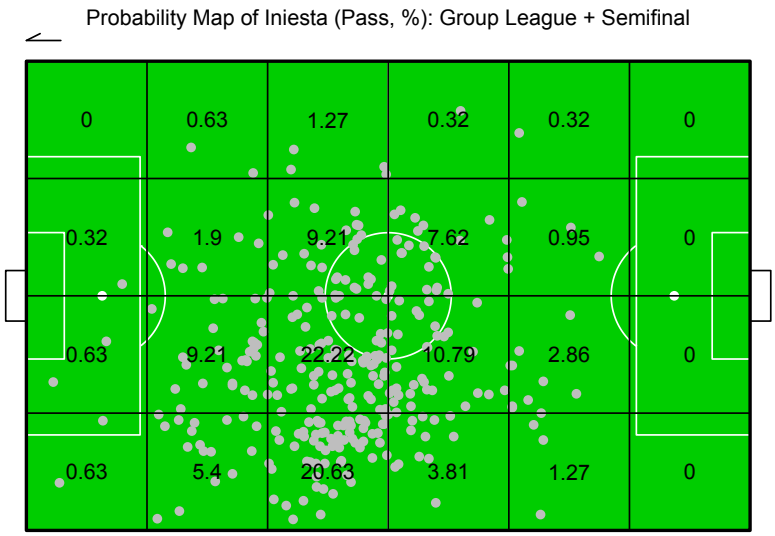




コドラートによる集計（度数）



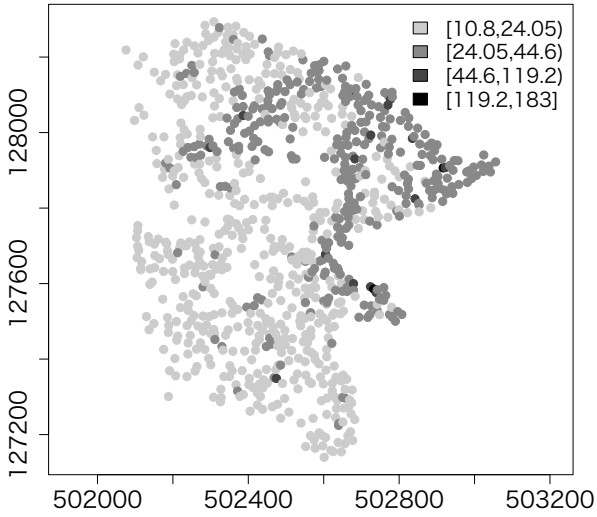
コドラートによる集計（密度）



演習課題 1

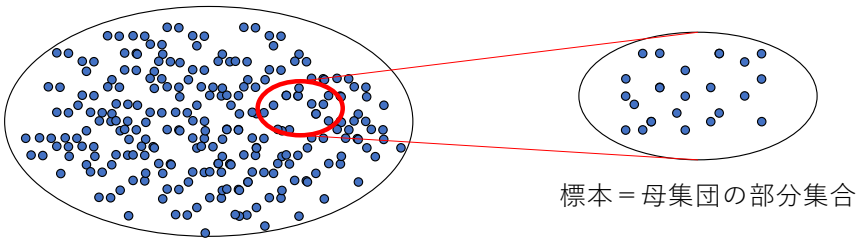
- SFC-SFSから「コドラートカウント用シート」をダウンロードし、サッカースペイン代表イニエスタ選手のパス位置をコドラート法を用いて度数と密度を計算しなさい
- コドラート以外の空間区分を提案し、度数と密度を計算しなさい

住宅地地価の分布

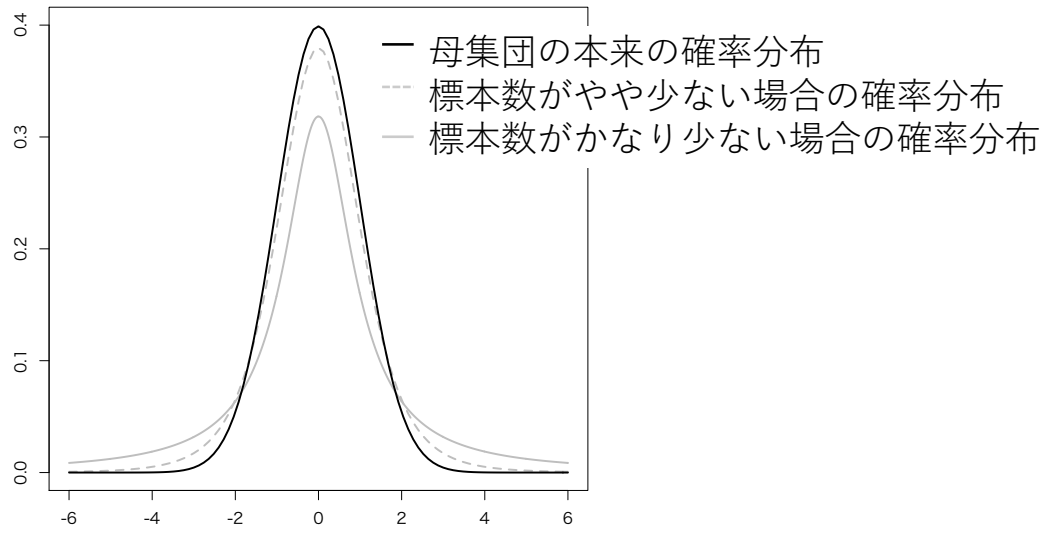


母集団と標本

- 関心のある集合全体のことを母集団という
- 母集団の部分集合を標本（サンプル）という
- 母集団から標本を選ぶことを抽出といい、母集団のどの要素も同様に確からしく抽出されることを無作為抽出という



標本数が少ないと…



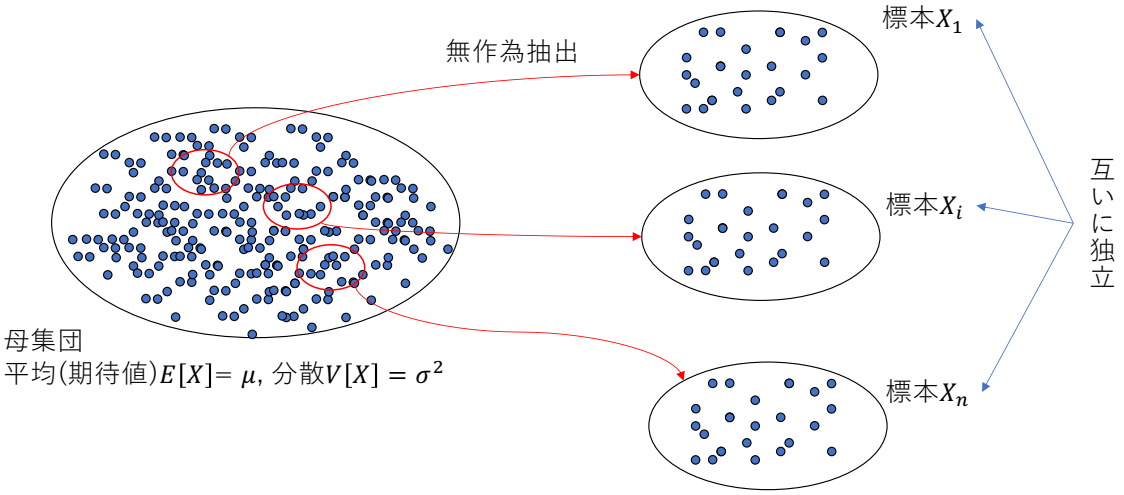
何をやりたいのか？

- 大量のデータが得られない場合、標本データから母集団の統計的性質を把握したい
- 標本データと母集団データとの統計的性質の関係を把握することで、標本データを用いた統計分析を有効なものとみなす

- そこでまず、標本と母集団の平均と分散（標準偏差）との関係について理解する
- どの標本を取り出しても母集団と同じ統計的性質をもつことが大事だが、その確証がないので、標本が母集団とどの程度異なる統計的性質を持つのかを把握する

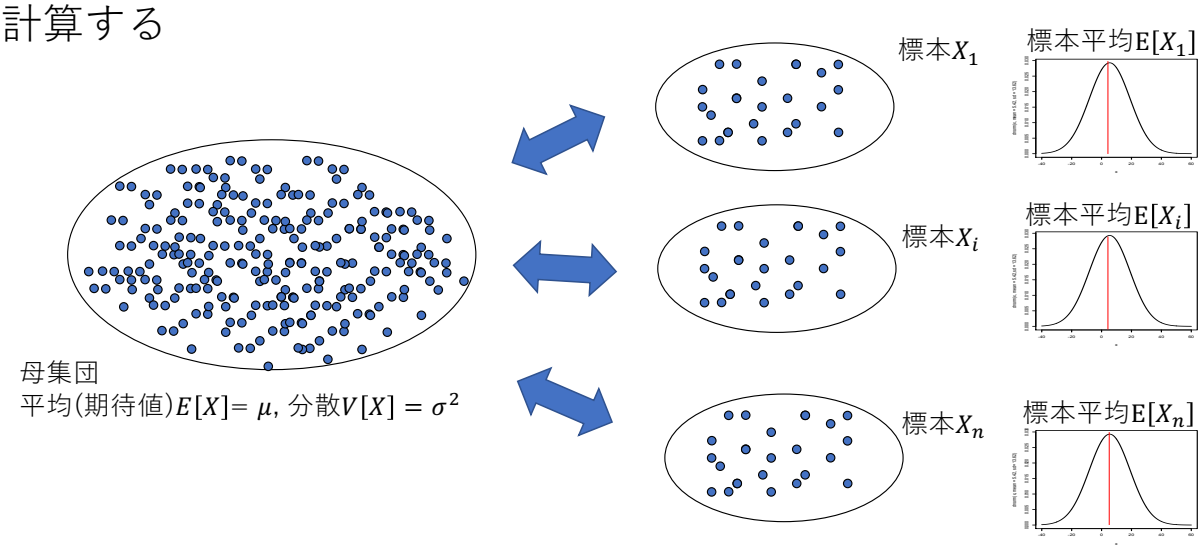
母集団の平均と標本の平均と分散

- 母集団から（互いに独立な）どの標本を（何度も）抽出しても、同じ様に母集団を再現できるのか？



母集団の平均と標本の平均と分散

- 母集団と標本との統計的性質の関係を把握する上で、各標本の平均と分散をそのまま扱うのではなく、標本平均の平均と分散を計算する



標本の平均

- 根元事象ωであるn個の標本$\{X_1, \dots, X_n\}$について、実現値$\{x_1, \dots, x_n\}$が得られたとする
- 標本$\{X_1, \dots, X_n\}$の平均$\overline{\mathrm{X}}$を以下の確率変数で捉えることにする

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{n}(X_1 + \cdots + X_n)$$

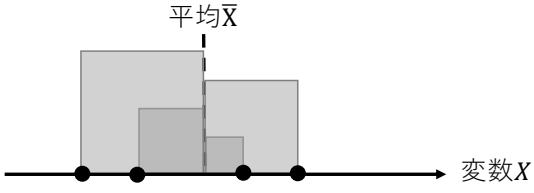
- ここで、確率変数X_iは同じ母集団から無作為抽出されたものとする。さらに、任意の確率変数X_iとX_jは互いに独立とする

標本の分散

- 標本$\{X_1, \dots, X_n\}$の分散s^2は平均$\overline{\mathrm{X}}$と標本X_iとの解離度$(\overline{\mathrm{X}} - X_i)$の平方和を標本数で割った値となる

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}\left((\overline{\mathrm{X}} - X_i)\right)^2$$

- イメージ的には、平均と各標本との差分を一辺とする正方形の面積の総和を標本数で割った値



- 母分散σ^2と標本分散s^2との関係性については後ほど考察する

標本平均の平均

母集団の期待値$E[X]=\mu$、分散$Var[X]=\sigma^2$とすると、
標本X_iの期待値は標本平均$E[X_i]=\mu$となる
標本平均$\bar{X}$の期待値は、

$$E[\bar{X}]=\mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}X_i\right]=\frac{1}{n}E\left[\sum_{i=1}^{n}X_i\right]=\frac{1}{n}\sum_{i=1}^{n}E[X_i]$$

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

となる。標本数nが大きくなるほど、標本平均は期待値に近づく。

標本平均の分散

他方、標本平均$\bar{X}$の分散は、

$$Var[\bar{X}]=\mathrm{Var}\left[\frac{1}{n}\sum_{i=1}^{n}X_i\right]=\frac{1}{n^2}Var\left[\sum_{i=1}^{n}X_i\right]$$

$$=\frac{1}{n^2}\left[\sum_{i=1}^{n}Var[X_i]+2\sum_{i\neq j}Cov[X_i,X_j]\right]$$

$$=\frac{1}{n^2}\sum_{i=1}^{n}s^2=\frac{s^2}{n}$$

となる。
標本数nが大きくなるほど、分散は小さくなる。

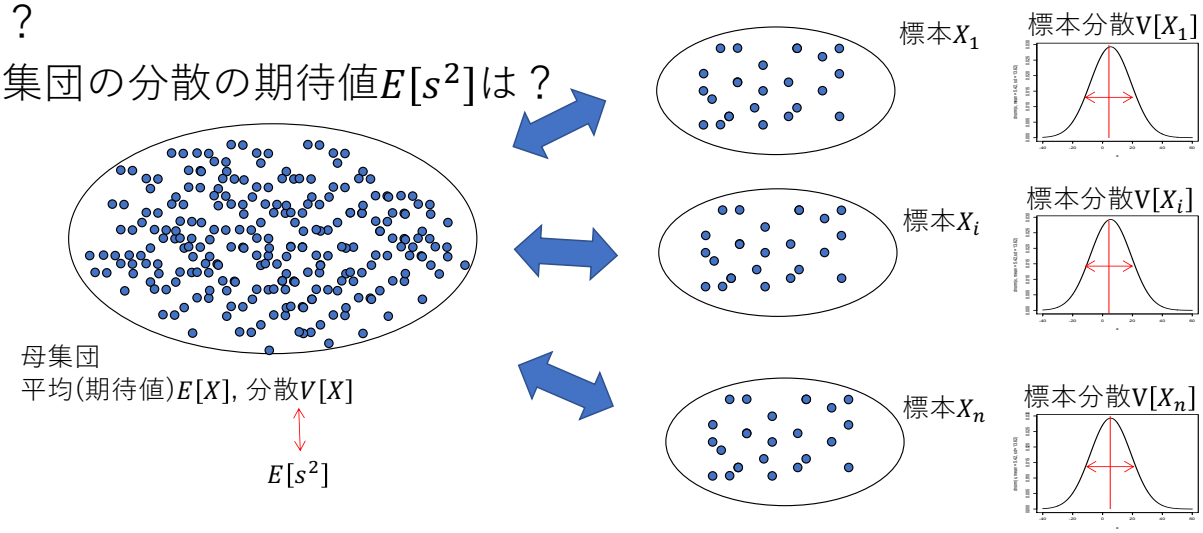
これまでに分かったことを整理すると…

- 母集団の期待値$E[X]=\mu$, 分散$Var[X]=s^2$とすると、標本X_iの期待値$E[X_i]=\mu$, 分散$Var[X_i]=s^2$となる
- 標本平均$\bar{X}$の平均（期待値） $E[\bar{X}]=\mu$である。標本数nが大きくなるほど母集団の平均（母平均に近づく）
- 標本平均$\bar{X}$の分散$Var[\bar{X}]=\frac{s^2}{n}$である。標本数$n$が大きくなるほど小さくなる

- では、母分散σ^2と標本分散s^2との関係性はどうなのか。
- 標本分散はどのような値にあることが期待されるか？＝標本分散の期待値$E[s^2]$はどのような値になるか？

母集団の分散と標本の分散

- 母集団の分散は、標本の分散s^2を用いて再現できるのか？
- 標本分散s^2と母集団の分散σ^2との間にどの程度の誤差があるのか？
- 母集団の分散の期待値$E[s^2]$は？



標本分散の期待値

- 標本分散の定義式に$\overline{\mathrm{X}} - X_i = \overline{\mathrm{X}} - \mu - (X_i - \mu)$を代入する

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}\left((\overline{\mathrm{X}} - X_i)\right)^2 = \frac{1}{n}\sum_{i=1}^{n}\left((\overline{\mathrm{X}} - \mu)^2 - 2(\overline{\mathrm{X}} - \mu)(X_i - \mu) + (X_i - \mu)^2\right)$$

$$= (\overline{\mathrm{X}} - \mu)^2 - 2(\overline{\mathrm{X}} - \mu)\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu) + \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2$$

$$= (\overline{\mathrm{X}} - \mu)^2 - 2(\overline{\mathrm{X}} - \mu)\frac{1}{n}(X_1 - \mu + X_2 - \mu + \cdots X_n - \mu) + \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2$$

$$= (\overline{\mathrm{X}} - \mu)^2 - 2(\overline{\mathrm{X}} - \mu)\frac{1}{n}(n\overline{\mathrm{X}} - n\mu) + \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2$$

$$= (\overline{\mathrm{X}} - \mu)^2 - 2(\overline{\mathrm{X}} - \mu)^2 + \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2$$

標本分散の期待値

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}\left((\overline{\mathrm{X}} - X_i)\right)^2 = \cdots$$

$$= (\overline{\mathrm{X}} - \mu)^2 - 2(\overline{\mathrm{X}} - \mu)^2 + \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2$$

$$= \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 - (\overline{\mathrm{X}} - \mu)^2$$

- 標本分散の期待値$E[s^2]$を求めると、次のようになる

$$E[s^2] = \frac{1}{n}E[(X_i - \mu)^2] - E[(\overline{\mathrm{X}} - \mu)^2]$$

標本分散の期待値

- 母集団の分散$\sigma^2 = \frac{1}{n}E[(X_i - \mu)^2]$であることから、
$$E[s^2] = \sigma^2 - E[(\overline{\mathrm{X}} - \mu)^2]$$
- つまり、標本分散の期待値$E[s^2]$は母分散σ^2より$E[(\overline{\mathrm{X}} - \mu)^2]$だけ小さい
- 標本分散にこの誤差分だけ修正を加えれば、標本分散を利用して母分散を推定できるようになる
- 平均μ、分散σ^2の母集団について、さらに次の関係が成立する
$$E[(\overline{\mathrm{X}} - \mu)^2] = \frac{1}{n}E[(X_i - \mu)^2] = \frac{1}{n}\sigma^2$$
- このことから、
$$E[s^2] = \sigma^2 - \frac{1}{n}\sigma^2 = \frac{n-1}{n}\sigma^2$$

となる

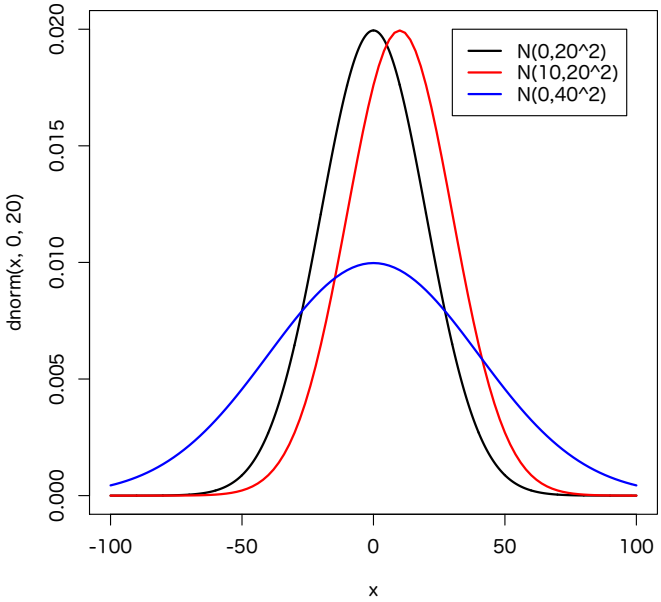
不偏分散

- 従って母分散は以下のようにして推定される
$$\sigma^2 = \frac{n}{n-1}E[s^2] = \frac{n}{n-1}\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{\mathrm{X}})^2\right)$$
$$= \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{\mathrm{X}})^2$$
- つまり、標本から分散を計算するときには、nで割るのではなく$n-1$で割ると母分散と等しくなる
- これを不偏分散$\hat{\sigma}^2$という
- 標本分散s^2は母分散σ^2に対して$E[(\overline{\mathrm{X}} - \mu)^2]$だけ偏りがある（小さい）が、不偏部分散$\hat{\sigma}^2$はそのような偏り（誤差）がない

正規分布

- 平均μ、分散σ^2(標準偏差σ)の正規分布$N(\mu, \sigma^2)$は、以下のように表される

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

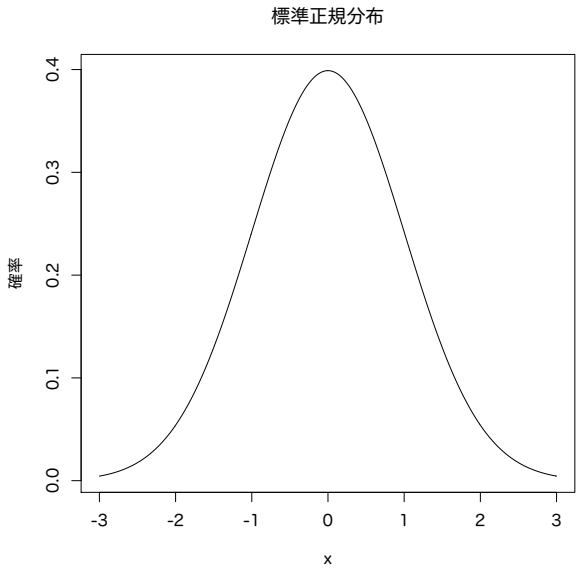


データの標準化

- 単位が異なる複数の変数を用いる場合や、単位に意味がない変数（例：５段階評価等）を用いる場合
- 偏回帰係数を比較するために、独立変数と従属変数を平均0・標準偏差1となるデータに標準化する
- 変数xに対する標準化データz_xは以下のように得られる

$$z_x = \frac{x - \bar{x}}{sd(x)}$$

- 標準化されたデータは、右図のような標準正規分布に従う



空間データの特徴を捉える方法

- 推測統計学では、以下の方法によりデータの特徴を捉えることができる
- コルモゴロフ・スミルノフ検定：分布の正規性を判断
- 等分散性の検定：正規分布を仮定するのが適当と判断されたデータについては属性データの分散が等しいかどうかにを検定
- 平均値の差の検定：地域属性の分布（平均値）の差をパラメトリックに検定
- ウィルコクソンの順位和検定：データが正規分布に従わない場合や標本数が少ない場合に地域属性の平均値の差をノンパラメトリックに検定

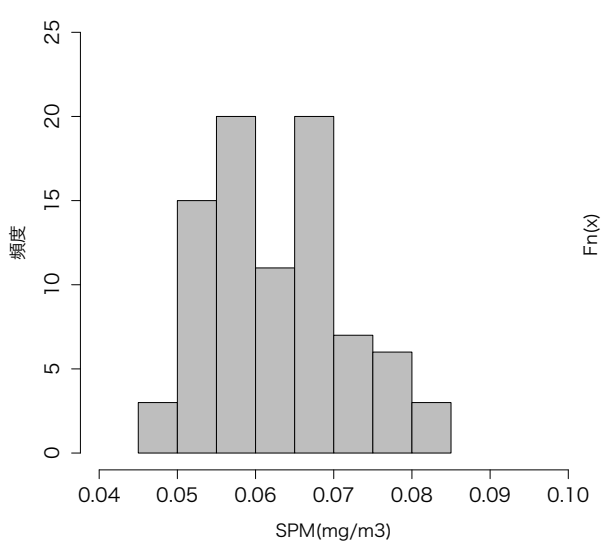
コルモゴロフ・スミルノフ検定

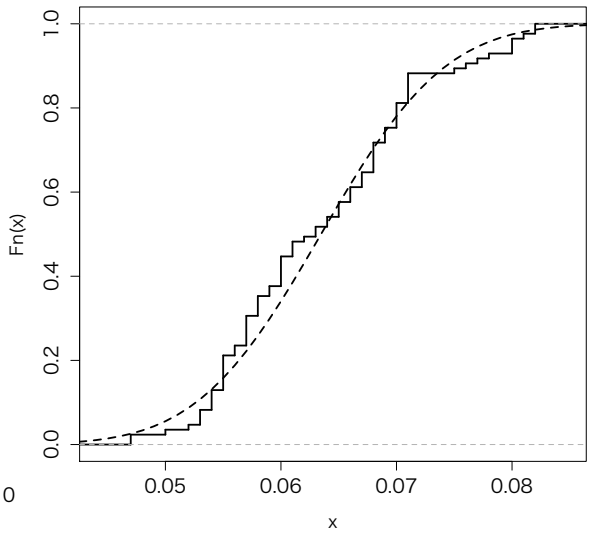
- コルモゴロフ・スミルノフ検定（KS検定）は、異なる二標本のデータの分布が、一致するかどうかを調べるために用いられる。
- 空間データ分析では、与えられた標本が正規分布に従うかどうかを判断する際に、コルモゴロフ・スミルノフ検定が、しばしば適用される。
- コルモゴロフ・スミルノフ検定では、階級区分毎に標本データの累積相対度数と正規分布の累積相対度数の差をとり、累積相対度数の差の絶対値が最も大きい値Dを用いて、以下のχ^2検定統計量を計算する

$$\chi^2 = 4D\frac{n_1 n_2}{n_1 + n_2}$$

- ここで、 n_1は標本データの個数、 n_2は正規分布のデータ数（ここでは標本データの個数と同じ）

コルモゴロフ・スミルノフ検定





SPM観測データの累積度数分布表

階級区分	標本データの経験累積度数	標本データの経験累積相対度数(A)	正規分布の累積度数	正規分布の累積相対度数(B)	(A)-(B)
0-0.050	3	0.035	5	0.055	0.020
0.051-0.055	18	0.212	13	0.158	0.054
0.056-0.06	38	0.447	29	0.341	0.107
0.061-0.065	49	0.576	49	0.572	0.005
0.066-0.07	69	0.812	66	0.780	0.031
0.071-0.075	76	0.894	78	0.914	0.020
0.076-0.08	82	0.965	83	0.975	0.010
0.081-0.085	85	1.000	85	0.995	0.005
0.086-0.09	85	1.000	85	0.999	0.001
0.091-0.095	85	1.000	85	1.000	0.000

等分散性の検定

- 等分散性の検定は、 F検定とも呼ばれる
- 二つの標本1と標本2について、次式であらわされる F値を計算

$$F = s_1^2 / s_2^2$$

ただし、 $s_1^2 > s_2^2$である。

- 二標本の自由度をもとに、 F分布表から得られる有意水準αのときのF_α値を求め、帰無仮説H_0「分散が等しい」を棄却できるかどうかを判定する。

平均値の差の検定

- ２つの標本分布（標本１と標本２）に対して、いずれも分布が正規分布に従い、互いに分散が等しいと考えられる分布については、平均値の差の検定を適用することで、地域属性の分布を比較できる
- 帰無仮説H_0「２つの標本の母平均が等しい」に対して、以下の検定統計量t値を計算し、そのp値が統計的に有意かどうかで帰無仮説を棄却するかどうかを判断する方法

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1 - 1} + \frac{s_2^2}{n_2 - 1}}}$$

ジニ係数

- 地域iの属性をx_i、地域属性の標本平均を$\bar{x}$とすると、ジニ係数Gは次式で表される

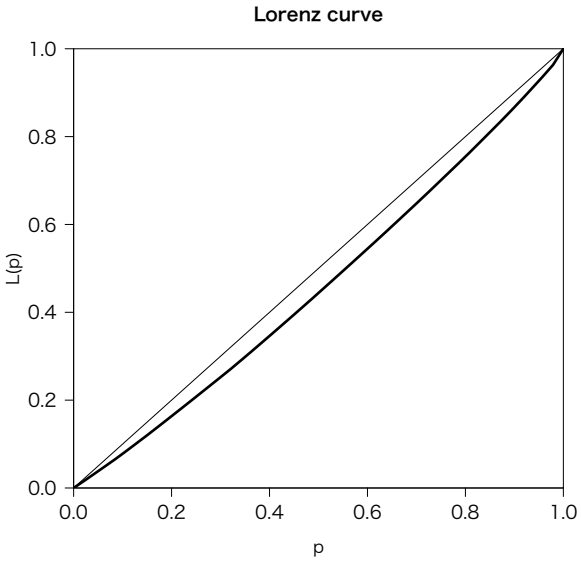
$$G = \sum_{i=1}^{n}\sum_{j=1}^{n}|x_i - x_j|/2n^2\bar{x}$$

- ジニ係数の値が 1 に近づくほど、地域間の不平等（格差）の度合いが高いことを示し、 0 に近いほど地域感の兵藤の度合いが高いことを示す尺度として用いられる。

ローレンツ曲線

- ジニ係数に示される地域間の不平等の状況を可視化する方法としてローレンツ曲線がある

- 全く地域間の格差が存在しない場合には、ローレンツ曲線は45°線と一致し、45°線とローレンツ曲線とで囲まれた弓形の部分の面積を二倍した値が、ジニ係数の値と一致する



変動係数

- 変動係数C_vは、地域属性の標本標準偏差sと標本平均$\bar{x}$を用いて、次式のように計算できる

$$C_v = \frac{s}{\bar{x}} = \frac{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2/n}}{\sum_{i=1}^{n} x_i/n}$$

- 変動係数が大きいほど、地域格差が大きいことを示す尺度として用いられる

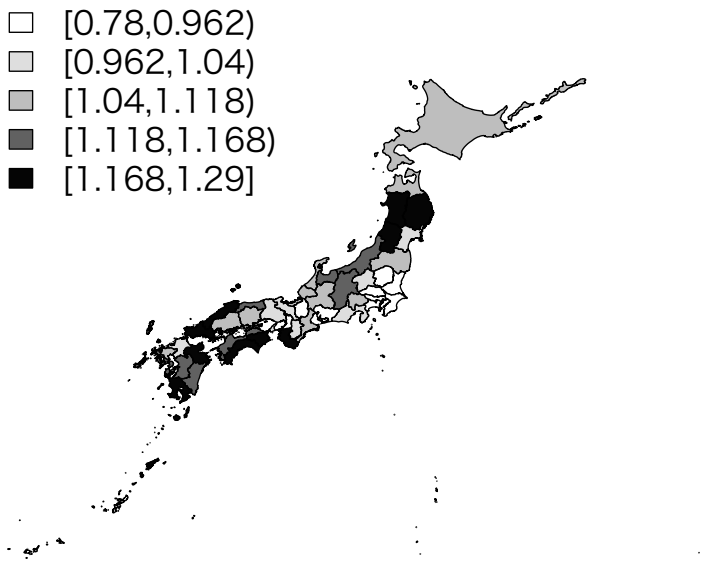
地域特化係数(Coefficient of Localization)

- 地域iの産業部門kに属する従業人口数をx_{ik}とする
- このとき、産業部門kの地域特化係数CLは次式で表される

$$CL = \frac{x_{ik}}{\sum_{k=1}^{K} x_{ik}} \Big/ \frac{\sum_{i=1}^{n} x_{ik}}{\sum_{i=1}^{n}\sum_{k=1}^{K} x_{ik}}$$

- 地域特化係数が 1 より大きければ、地域iは産業部門kに特化していると言える

地域特化係数の分布（高齢者分布の例）



演習課題２：地域特化係数の計算

- 以下の表はA〜Eの５地域の産業別人口数を示したものである
- 各地域の地域特化係数を計算しなさい（単位：万人）
- 産業別人口の平均値をそれぞれ求めなさい

地域	第一次産業	第二次産業	第三次産業
A	15	15	40
B	20	25	35
C	25	40	80
D	30	10	20
E	10	30	40

e-stat

https://www.e-stat.go.jp/SG1/estat/eStatTopPortal.do

Rのインストール

- 次回以降、統計ソフトRを使うので、インストールしたPCを持参すること
- SFC-SFSから授業演習用データをダウンロードしておくこと