統計解析論　1

データの性質
第1章

1－0　記号とその対応

- 観測値，測定値（データ値）
  - 例：京都市左京区の土曜日以降の予想最低気温
  - 数値：{8，8，6，6，10}
- 記号

$$\{x_1, x_2, x_3, x_4, x_5\}$$

  - x_i：i番目の観測値
  - n：　標本サイズ＝最後の番号（例では5）
  - x：観測値の変数を一般的に表す（この例では「京都市左京区の土曜日以降の予想最低気温」）

2

1－1　平均と分散（1）

- 標本平均：観測値の中心をあらわす指標

$$\frac{\text{標本値の合計}}{\text{標本数}} = \frac{\sum_{i=1}^{n} x_i}{n} \qquad \frac{\sum_{i=1}^{\text{□の標本サイズ}} \square_i}{\text{□の標本サイズ}}$$

  - 記号：$\bar{x}$　　$\overline{\square}$
  - データの値から計算した平均
  - 「平均」は「分布の平均」「母平均」
    - 分布の平均＝期待値

3

1－1　平均と分散（2）

- 標本分散：観測値のちらばりを表す指標第1号

$$\frac{(\text{観測値} - \text{標本平均})^2\text{の合計}}{\text{割る数}} = \frac{1}{\text{割る数}} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

  - データを取得した集団が・・・ $\dfrac{1}{\text{割る数}} \sum_{i=1}^{\text{□の標本サイズ}} \left(\square_i - \overline{\square}\right)^2$
  - 母集団と一致する場合
    - 標本サイズ＝母集団サイズで割る
  - 母集団の一部の場合
    - （標本サイズ－1）で割る
  - 記号：s_x^2, s_{xx}，　$s_{\square}^2$　：□の分散

4

1－1　平均と分散（3）

- 標本標準偏差
  - 観測値のちらばりを表す指標第2号

$$\sqrt{\text{標本分散}} = \sqrt{s_x^2} = s_x = \sqrt{s_{xx}}$$

- □σ区間（□シグマ区間）

$$\left[\bar{x} - \square s_x, \bar{x} + \square s_x\right]$$　の区間のこと

  - たとえば，□＝1．5なら，1．5シグマ区間で

$$\left[\bar{x} - 1.5 s_x, \bar{x} + 1.5 s_x\right]$$

5

1－1　平均と分散（4）

- チェビシェフの不等式
  - どんなデータにも当てはまる

$$\frac{\square\text{シグマ区間に入る標本の数}}{\text{標本サイズ}} \geq 1 - \frac{1}{\square^2}$$

  - 例えば，1．5シグマ区間に入る標本の割合は

$$1 - \frac{1}{1.5^2} = \frac{5}{9}$$　以上

  - 逆に考えると $\dfrac{\square\text{シグマ区間外の標本の数}}{\text{標本サイズ}} < \dfrac{1}{\square^2}$

6

1－1　平均と分散(5)

- チェビシェフの不等式に関する注意

$$\frac{(\bar{x}-k s_x, \bar{x}+k s_x)\text{区間に入る標本の数}}{\text{標本サイズ}} > 1-\frac{1}{k^2}$$

$$\frac{(\bar{x}-k s_x, \bar{x}+k s_x)\text{区間外の標本の数}}{\text{標本サイズ}} \leq \frac{1}{k^2}$$

  - 法則
    - 区間が閉区間なら等号付き
    - 区間が開区間なら等号なし

1－1　平均と分散(6)

- 標準化した□　$\dfrac{\square_i - \overline{\square}}{s_{\square}}$
- 標準化した□の標本平均は0
- 標準化した□の標本分散は1

1－2　相関(1)

- 二つの変数の間の関連を指標化する
- 記号

i	x_i	X	Y	y_i
1	x_1	7.27	14.50	y_1
2	x_2	6.66	17.95	y_2
3	x_3	4.36	11.00	y_3
20	x_{20}	1.86	7.89	y_{20}

1－2　相関(2)

- 標本共分散:二つの変数の関連の指標
  - ＋，－，0だけが問題
    - 変数のちらばり方に影響を受けるから
  - 共分散が正
    - xが大きくなれば，yも大きくなる
    - 正の相関
  - 共分散が負
    - xが大きくなれば，yは小さくなる
    - 負の相関
  - 共分散が0
    - xとyには関係がない
    - 無相関

1－2　相関(3)

- 標本共分散の定義

$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})$$

$$s_{\square\triangle} = \frac{1}{\text{標本サイズ}-1}\sum_{i=1}^{\text{標本サイズ}}(\square_i-\overline{\square})(\triangle_i-\overline{\triangle})$$

- 割る数に関しては，
  - データを取得した集団が･･･
  - 母集団と一致する場合
    - 標本サイズ＝母集団サイズで割る
  - 母集団の一部の場合
    - (標本サイズ－1)で割る

1－2　相関(4)

- 標本相関係数

$$r_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{s_x}\cdot\frac{y_i-\bar{y}}{s_y}\right)$$

$$= \frac{1}{n-1}\sum_{i=1}^{n}(\text{標準化した}x\cdot\text{標準化した}y)$$

$$= \text{標準化した}x\text{と標準化した}y\text{との共分散}$$

  - 標準化されていると言うことはx，yの散らばり具合に影響されない

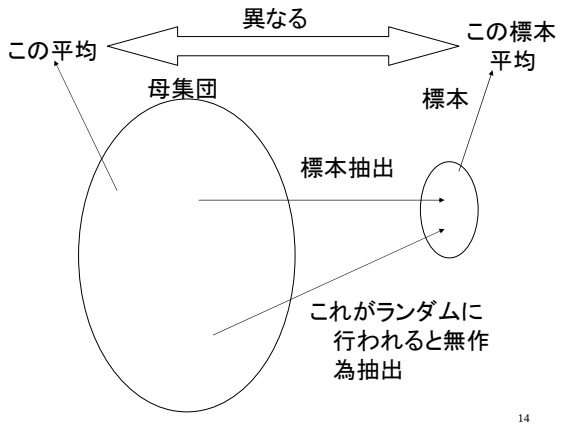
1－2　相関(5)

- 標本相関係数の性質

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

$$-1 \le r_{xy} \le 1$$

13



14

1－3．検定(1)

- 検定は何を問題にしているか
  - パラメータの推定値がある値と異なっているときその異なり方が，本当の値がその値と異なっているといえるほど十分大きいのか？
- 有意
  - 本当にその値と異なっているといえるほど十分大きい＝有意
  - 逆に小さいと偶然，誤差項の影響で異なっているだけと考える．

15

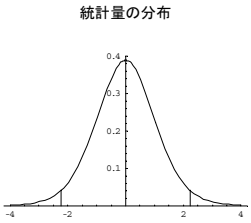
1－3．検定(2)

- 帰無仮説と対立仮説
- 有意性を考えるときは否定をまず考える．
  - 実際には異なってはいないと仮定するところから出発する．(言いたいことの否定を仮定することが多い)
  - この仮説を**帰無仮説**と呼ぶ．回帰係数に関しては，「パラメータ＝ある値」が**帰無仮説**である．
- 帰無仮説が反証されたときにはその否定を正しいと見なす(採択する)．
  - これ以後は帰無仮説がデータによる反証を受ける．データが反証するに十分なものなら，その(一般的には)否定＝**対立仮説**を採択する．帰無仮説は棄却される．
  - 対立仮説は一般的には「パラメータ値≠ある値」となる

16

1－3．検定(3)

- 有意水準(その1)
- 「本当にある値と異なっている」といえるほど十分大きいとは？その基準＝反証の基準
- 差ではなく検定統計量で比較する(この場合t統計量)
- t統計量の分布は帰無仮説が正しければ，自由度n-Kのt分布である．
- 十分異なっている場合はt統計量の値は自由度n-Kのt分布のもとであまり起きそうもない値となる．
- K(自由度と呼ぶ)は問題によってかわる

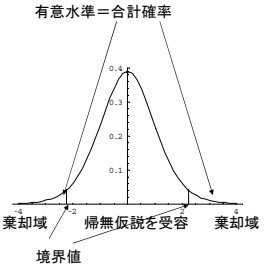


17

1－3．検定(4)

帰無仮説が正しいときの統計量の分布

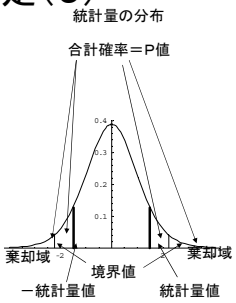
- 有意水準(その2)
- 棄却域
  - 起きそうもない値と考える範囲(図の縦線の外側)
  - この線の外側の値では対立仮説を受け入れる．
- 有意水準
  - 外側の確率
  - 有意水準としては1%，5%が使われる．
- 境界値
  - 帰無仮説を受け入れる境界



18

1－3．検定(5)

- P値
- 統計量の値を境界値とした検定を仮に考える．
- その検定の有意水準をP値と呼ぶ
- P値＜有意水準なら，帰無仮説が棄却され，対立仮説が採択される．
- P値≧有意水準なら，帰無仮説が受容される．



19

1－3．検定(6)

- 片側検定と両側検定
- 対立仮説（示したい仮説）が「パラメータ＜ある値」，「パラメータ>ある値」のとき＝片側検定
  - 帰無仮説は「β j≦ある値」か「β j≧ある値」
  - このときt統計量の分布の計算は「β j＝ある値」で計算しても良いことがわかっている．
  - 従って，棄却域は対立仮説の側にだけ＝片側に取られる．
  - 同じ有意水準なら片側検定の境界値の絶対値は小さくなる
- 対立仮説が「パラメータ≠ある値」のとき＝両側検定
  - 片側検定のα％有意水準での境界値は，２α％有意水準の両側検定での境界値

20

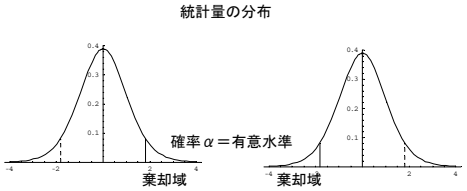
1－3．検定(7)

$H_a(H_1$対立仮説$): \beta >$ ある値　　$H_a(H_1$対立仮説$): \beta <$ ある値

H_0(帰無仮説$): \beta \leq$ ある値　　H_0(帰無仮説$): \beta \geq$ ある値

または　　または

$\beta =$ ある値　　$\beta =$ ある値



21

1－3．検定(8)

- 両側検定
  - 検定統計量の絶対値＞境界値の絶対値
    - 帰無仮説を棄却，対立仮説を採択
  - さもなくば帰無仮説は棄却できず，帰無仮説を受容
- 片側検定
  - 対立仮説が「パラメータ＞ある値」の場合
    - 検定統計量＞境界値
      - 帰無仮説を棄却，対立仮説を採択
  - 対立仮説が「パラメータ＞ある値」の場合
    - 検定統計量＜境界値
      - 帰無仮説を棄却，対立仮説を採択

22

1－3．検定(9)

- 検定の手順
  - 帰無仮説，対立仮説の設定
    - 両側検定か片側検定かがきまる
  - 有意水準の決定
    - どの程度の差を有意な差とみなすか？
  - 境界値の決定
    - 帰無仮説が正しいときの検定統計量の分布
    - 分布表を使う
  - 検定統計量の計算
    - 計算式に従う
  - 検定統計量と境界値の比較
    - 帰無仮説を棄却し対立仮説を採択する
    - 帰無仮説は棄却できず，帰無仮説を受容する

23

1－3．検定(10)

- 検定の実際
- 帰無仮説が「パラメータ＝0以外の値」の場合
  - t統計量を計算し直す．
  - P値を計算し直す．
  - 実例をよく見ておいてください．
- 片側検定の場合
  - Excelの分析ツールの出力では，t検定のP値は両側検定で計算される．
  - 片側検定でのP値はExcelで出力されたP値の半分である．したがって，P値／2を有意水準と比較

24

1－3．検定（11）

- 母平均に関する検定（1）
  - 帰無仮説
    - 母平均（分布の平均）はある値と等しい
      - このタイプの検定は「平均に関する検定」の手順を踏む
  - 対立仮説を見極めよう
    - 言いたいことは，「母平均がその値と等しくない（A1）」ことか？それとも「母平均がその値より大(A2)」「母平均がその値より小(A2)」なのか
    - 前者なら，両側検定，後者なら片側検定

25

1－3．検定（12）

- 母平均に関する検定（2）
  - 有意水準を決める．通常5%
  - 境界値をさだめる
    - (A1)なら自由度が（標本サイズ－1）のt分布よりもとめ，教科書293の表では，列を上記の自由度の列，行はp=有意水準/2となる行の交差するところの値
    - (A2)なら自由度が（標本サイズ－1）のt分布よりもとめ，教科書293の表では，列を上記の自由度の列，行はp=有意水準となる行の交差するところの値
    - (A3)なら自由度が（標本サイズ－1）のt分布よりもとめ，教科書293の表では，列を上記の自由度の列，行はp=有意水準となる行の交差するところの値×－1

26

1－3．検定（13）

- 母平均に関する検定（3）
  - 検定統計量 $$t = \frac{\bar{x} - \text{ある値}}{s_x / \sqrt{n}}$$
  - 検定結果の判定
    - (A1)
      - 検定統計量の絶対値＞境界値の絶対値の時，帰無仮説を棄却，対立仮説を採択．さもなくば，帰無仮説を受容
    - (A2)
      - 検定統計量＞境界値の時，帰無仮説を棄却，対立仮説を採択
    - (A3)
      - 検定統計量＜境界値の時，帰無仮説を棄却，対立仮説を採択

27

1－3．検定（14）

- （対応のない）2標本の平均の検定（1）
  - 帰無仮説
    - 両方の標本の属す母集団の母平均（分布の平均）は等しい
  - 両方の母集団の分散（母分散）は等しいと仮定
  - 対立仮説を見極めよう
    - 言いたいことは，「両母集団の母平均が等しくない（A1）」ことか？それとも「ある母平均がある母平均より大(A2)」なのか
    - 前者なら，両側検定，後者なら片側検定

28

1－3．検定（15）

- （対応のない）2標本の平均の検定（2）
  - 有意水準を決める．通常5%
  - 境界値をさだめる
    - (A1)なら自由度が（標本サイズの合計－2）のt分布よりもとめ，教科書293の表では，列を上記の自由度の列，行はp=有意水準/2となる行の交差するところの値
    - (A2)なら自由度が（標本サイズの合計－2）のt分布よりもとめ，教科書293の表では，列を上記の自由度の列，行はp=有意水準となる行の交差するところの値

29

1－3．検定（16）

- （対応のない）2標本の平均の検定（2）
  - 検定統計量 $$t = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{(n_x - 1)s_x^2 + (n_y - 1)s_y^2}{n_x + n_y - 2}} \sqrt{\frac{1}{n_x} + \frac{1}{n_y}}}$$
  - 検定結果の判定
    - (A1)
      - 検定統計量の絶対値＞境界値の絶対値の時，帰無仮説を棄却，対立仮説を採択．さもなくば，帰無仮説を受容
    - (A2)
      - 検定統計量＞境界値の時，帰無仮説を棄却，対立仮説を採択

30

1－3．検定（17）

- 分散が異なる場合
  - 境界値
    - p．293の付表2でp=有意水準/2（両側検定），有意水準（片側検定）の列をみる
  - 検定統計量（帰無仮説下で正規分布）

$$\frac{\bar{x}-\bar{y}}{\sqrt{s_x^2/n_x+s_y^2/n_y}}$$

$n_x = x$集団の標本数，$n_y = y$集団の標本数

$$s_x^2=\sum_{i=1}^{n_x}(x_i-\bar{x})^2\Big/n_x,\ s_y^2=\sum_{i=1}^{n_y}(y_i-\bar{y})^2\Big/n_y$$

1－3．検定（18）

- 等分散の検定（1）
  - 帰無仮説：$\sigma_A^2=\sigma_B^2$
  - 対立仮説：$\sigma_A^2\neq\sigma_B^2$
  - 統計量：（こっそり母集団は正規分布と仮定）

$$\frac{\frac{1}{n_A-1}\sum_{A集団}^{n}(X_i-\bar{X}_A)^2}{\frac{1}{n_B-1}\sum_{B集団}^{n}(X_i-\bar{X}_B)^2}\cong\frac{\frac{\chi^2(n_A-1)}{n_A-1}}{\frac{\chi^2(n_B-1)}{n_B-1}}=F(n_A-1,n_B-1)$$

  - 両側検定を用いて，帰無仮説が棄却されれば，異分散の場合の検定，受容されれば等分散の検定

1－3．検定（19）

- 等分散の検定（2）
  - 境界値の計算法
    - 検定統計量を計算し，もし1より小さかったならば逆数をとる
    - これと，$F(n_A-1,n_B-1)$ の97.5%点を比較する
    - もし，検定統計量の値が大きければ帰無仮説を棄却
    - そうでなければ，帰無仮説を受容
  - 等分散性が棄却されたなら，分散が異なる場合の検定を利用する

1－3．検定（20）

- 相関係数の検定
  - 帰無仮説：相関はない（母相関係数が0）
  - 対立仮説：相関がある（母相関係数が0ではない）
  - 検定統計量 $\sqrt{\frac{(n-2)r^2}{1-r^2}}$
  - 境界値の決め方
    - 自由度n－2のt分布の分布表からきめる