データリテラシー

第12回　仮説検定

メディア情報コース
平居　悠（ひらい　ゆたか）

到達目標

- データサイエンス・AIが社会でどのように活用されているのかを理解する。
- データを適切に読み解き理解する力をつける。
- データを適切な可視化手法で他者に説明できるようになる。
- 表計算ソフトを使ってデータを集計・加工できるようになる。

前回

第１回　社会におけるデータ・AI利活用
第２回　表計算ソフトの基本操作
第３回　セルの参照と集計
第４回　度数分布表とヒストグラム
第５回　代表値と散らばりの尺度
第６回　データのグラフによる表現とデータ比較
第７回　データの分布と標本抽出
第８回　推定
第９回　分割表とクロス集計
第１０回　２変量データと相関
第１１回　時系列データ
第１２回　仮説検定
第１３回　回帰分析

前回の目標

時系列データを分析できるようになる。

前回学んだこと

1. 時系列データ
2. 移動平均
3. 階差
4. 自己相関と指数

時系列データ

身の回りの現象は時間的に変動するものが多い。その現象の時間的変動を継続観測して得られるデータを**時系列データ**という。

時系列データ

時系列データは観測される時間と順序に意味がある。

時系列データ

時系列データから取り出す情報として、長期的傾向と短期的傾向がある。時系列データの長期的傾向は**トレンド**という。

時系列のトレンドを見たいときは、短期的な変動は邪魔（ノイズ）になり、逆に短期的な変動を見たい時は、トレンドが邪魔になる。

移動平均

時系列データが短期的に大きくなったり小さくなったりを繰り返す振動をしている場合、各時点の前後のいくつかの観測値で平均を取ってトレンドの情報を取り出すことができる。これを**移動平均**という。

移動平均

ある時点の観測値y_iの前後$2k+1$個の観測値 (前のk個と後ろのk個)との移動平均は以下の通り。

$$z_i = \frac{y_{i-k} + y_{i-k+1} + \cdots + y_{i-1} + y_i + y_{i+1} + \cdots + y_{i+k-1} + y_{i+k}}{2k+1}$$

$$= \frac{1}{2k+1} \sum_{j=-k}^{k} y_{i+j}$$

階差

ある時点の観測値y_iとその一つ前の観測値y_{i-1}の差を**階差**といい、Δy_iで表す。

$$\Delta y_i = y_i - y_{i-1}$$

階差によってトレンドを取り除き、短期的傾向を見ることができる。

自己相関

同じ時系列データの中で、異なる時点との相関を**自己相関**という。自己相関を表す指標を**自己相関係数**という。各時点とh後の時点との自己相関係数r_hは次ページのように定義される。

自己相関係数

$$r_h = \frac{\dfrac{1}{n-h}\displaystyle\sum_{i=1}^{n-h}(y_i - \bar{y})(y_{i+h} - \bar{y})}{S_y^2}$$

ここで$\bar{y}$は平均、S_y^2は分散である。

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \qquad S_y^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

自己相関係数

自己相関係数は相関係数の一種なので、
相関係数の性質を持つ。

$r_h > 0$：
各時点の傾向がh後の時点でも同じ傾向

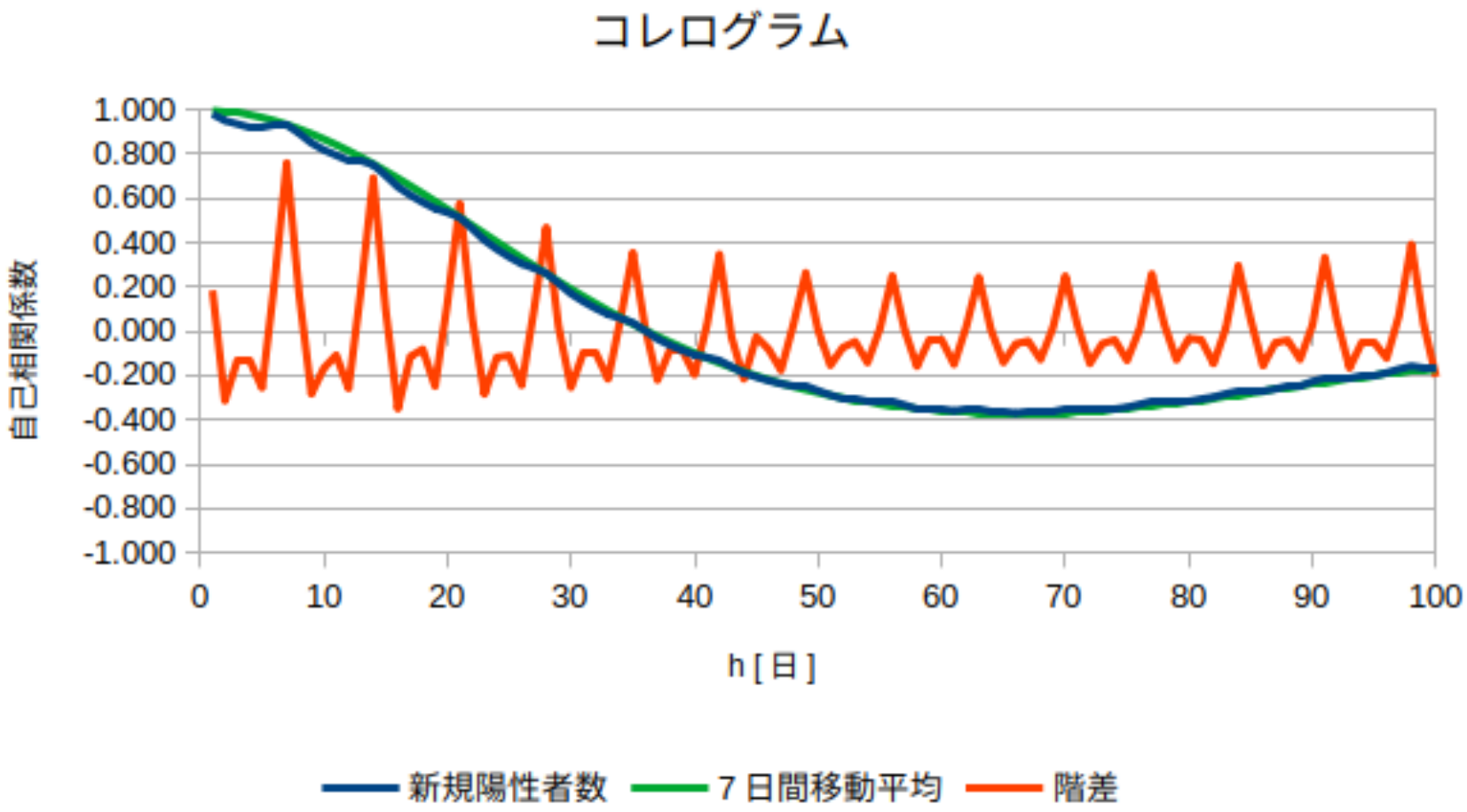
$r_h < 0$：
傾向が逆転

コレログラム

hを変化させることでr_hも変化する。横軸にh、縦軸にr_hをとって描いたグラフを**コレログラム**という。

コレログラムの例

新型コロナウイルス新規陽性者数のデータのコレログラム



コレログラム

hが大きいほど、より後で観測された観測値との相関になるので、相関は弱くなるのが一般的である。

しかし、周期性を持つ時系列データでは、その周期ごとに相関は強くなる。

コレログラムを見ることで周期がどれくらいか知ることができる。

指数

複数の時系列データを比較する際には、ある時点を基準の**100**として、そこからの変化率で見ることがある。この変化率を**指数**という。

物価指数

物価の変動を財・サービスの量と価格を元に計算される指数。消費者物価指数や企業物価指数などがある。

消費者物価指数 (CPI: Consumer Price Index)

全国の世帯が日常購入する商品やサービスの物価変動を表す指数。

ある基準となる時点（基準時）に家計で購入した商品やサービスと同じものと、比較する時点（比較時）に買い揃えるとしたらどれだけお金がかかるかという考えで計算されている（ラスパイレス方式）。

消費者物価指数 (CPI: Consumer Price Index)

比較時の消費者物価指数は、基準時の物価を100として表される。ラスパイレス方式の計算式は以下のようになる。

$$I_t = \frac{\sum_{i=1}^{n} p_{t,i} q_{0,i}}{\sum_{i=1}^{n} p_{0,i} q_{0,i}} \times 100$$

変数	意味
I_t	比較時の消費者物価指数
i	品目
$p_{t,i}$	比較時の品目の価格
$p_{0,i}$	基準時の品目の価格
$q_{0,i}$	基準時の品目の購入数量

前回の問い

- 移動平均を取ると何がわかるか？
  - トレンドの情報を取り出すことができる。
- 階差で見れる傾向は何か？
  - トレンドを取り除き、短期的傾向を見ることができる。
- 自己相関とは何か？
  - 同じ時系列データの中で、異なる時点との相関。

タイピング練習スケジュール

第１回　　ホームポジション
第２回　　ホームポジション
第３回　　ローマ字
第４回　　英語初級
第５回　　日本国憲法（trr試験）
第６回　　ホームポジション
第７回　　ローマ字
第８回　　英語初級
第９回　　日本国憲法（trr試験）
第１０回　ホームポジション（ボーナス練習）
第１１回　ローマ字
第１２回　英語初級
第１３回　日本国憲法（trr試験）

タイピング練習 (英語初級)

1. ブラウザを起動し、https://www.koeki-prj.org/trr/ に繋ぐ。
2. 学籍番号（Cは大文字、省略なし８桁）を入力する。
3. Koeki MAILに届いたパスコードをPasscode: 欄に入力する。

ホームポジション


左手でタイプするキー
右手でタイプするキー
小指
薬指
中指
人差指
人差指
中指
薬指
小指
漢字
Back space
Tab
Enter
Caps Lock
Shift
Shift
Ctrl
Fn
Alt
Space
Alt
Fn
Ctrl
左手の人差指から小指までのホームポジション
両手の親指のホームポジション
右手の人差指から小指までのホームポジション

今回

第１回　社会におけるデータ・AI利活用
第２回　表計算ソフトの基本操作
第３回　セルの参照と集計
第４回　度数分布表とヒストグラム
第５回　代表値と散らばりの尺度
第６回　データのグラフによる表現とデータ比較
第７回　データの分布と標本抽出
第８回　推定
第９回　分割表とクロス集計
第１０回　２変量データと相関
第１１回　時系列データ
第１２回　仮説検定
第１３回　回帰分析

今回の目標

母集団についての仮説が正しいかを標本から検証できるようになる。

今回学ぶこと

1. 仮説検定の考え方
2. 母平均の仮説検定
3. p値

今回の問い

- 仮説検定とは何か？
- 仮説検定はどのような手順で行えば良いか？
- p値とは何か？

今回学ぶこと

1. 仮説検定の考え方
2. 母平均の仮説検定
3. p値

仮説検定とは

母集団について仮定された命題（仮説）を標本に基づき統計的に検証すること。

仮説検定

例えば、ある母集団について

「母平均μは3より大きい」($\mu > 3$)

という仮説が正しいかを標本から検証したいとする。

仮説検定

μの推定に用いた統計量は標本平均$\bar{X}$なので、

「$\bar{X} > 3$であれば$\mu > 3$という仮説は正しい」

と単純に結論づけたいと思うかもしれない。

仮説検定

もし$\mu > 3$という仮説が正しいとしても、$\bar{X}$ の値は標本によってばらつきがあるため、$\bar{X}$が3より大きくなることもあれば、運悪くたまたま3以下となることもあるかもしれない。

仮説検定

このような仮説を直接検証するのは難しいため、仮説検定では数学の**背理法**に近い方法で検証を行う。

背理法：ある命題を直接証明するのではなく、その命題の結論を否定する仮定をして推論を行うと矛盾が生じることからもとの命題を証明する方法。

仮説検定の方法

- 検証したい仮説（**対立仮説**）とそれに反する仮説（**帰無仮説**）を設定する。
- 「もし帰無仮説が正しいなら起こりにくい結果」が標本から得られたならば、それは帰無仮説が正しくないためであると判断して帰無仮説を**棄却**する（捨て去る）証拠とする。
- 帰無仮説を棄却することはそれに反する対立仮説を**採択**することになり、これで対立仮説を検証できたことになる。

仮説検定の方法

もし帰無仮説が正しいなら起こりにくい結果が得られる確率を**有意水準**αといい、この起こりにくい結果の範囲を**棄却域**Rという。

有意水準 α は検証する人が恣意的に設定する確率で、0.1 (10%), 0.05 (5%), 0.01 (1%)がよく使われる。

仮説検定の方法

標本から得られた統計量が棄却域 R に入った場合は、帰無仮説を棄却して、対立仮説を採択するという判断をする。

棄却域はαによって決まる。もし帰無仮説が正しいならば、αが小さいほど標本の統計量が棄却域に入る可能性は低くなるが、そんな小さい確率でも棄却域に入ったのであれば、帰無仮説が正しいという仮定が間違っていた証拠として良いことになる。

仮説検定の方法

標本の統計量が棄却域Rに入らなかった場合は、帰無仮説を棄却できないが、帰無仮説を採択することにはならない。

帰無仮説は結果がどうであれ、最後は無に帰する運命にある。検証しようとしているのはあくまで対立仮説である。

仮説検定における誤り

仮説検定は確率的に判断するので、標本による判断が誤りである可能性はゼロではなく、次の二通りの誤りが起こりうる。

第一種の過誤

本当は帰無仮説が正しいのに帰無仮説を棄却してしまう。

第二種の過誤

本当は帰無仮説が正しくないのに
帰無仮説を棄却できない。

第一種の過誤と第二種の過誤

第一種の過誤が起こる確率が有意水準αである。αが小さいほど第一種の過誤が起こりにくいが、逆に第二種の過誤が起こりやすくなってしまう。そのため、αが小さいほどよいということにはならない。

第一種の過誤と第二種の過誤

第一種の過誤と第二種の過誤の両方を起こりにくくしたい場合は、αを小さくし、かつ、標本サイズnを大きくすることが必要。

今回の問い

- **仮説検定とは何か？**
  - **母集団についての仮説を標本に基づき統計的に検証すること。**
- 仮説検定はどのような手順で行えば良いか？
- p値とは何か？

今回学ぶこと

1. 仮説検定の考え方
2. **母平均の仮説検定**
3. p値

母平均の仮説検定

母平均μの仮説検定の手順はほぼ決まっており、母平均以外の仮説検定についても仮説や統計量と標本分布が異なるだけで手順は同じ。

母平均の仮説検定

母平均μの仮説検定に使われる統計量は標本平均$\bar{X}$である。

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

仮説検定の手順

①仮説の設定

帰無仮説H_0
と
対立仮説H_1
を設定する。

①仮説の設定

対立仮説H_1によって

- **右片側検定**
- **左片側検定**
- **両側検定**

の三通りある。

①仮説の設定

右片側検定

- 帰無仮説$H_0: \mu = \mu_0$
- 対立仮説$H_1: \mu > \mu_0$

ここで、μ_0は検証する人が設定する定数。

①仮説の設定

左片側検定

- 帰無仮説$H_0: \mu = \mu_0$
- 対立仮説$H_1: \mu < \mu_0$

①仮説の設定

両側検定

- 帰無仮説$H_0: \mu = \mu_0$
- 対立仮説$H_1: \mu \neq \mu_0$

①仮説の設定

検証したいのは対立仮説 H_1 であり、検証したいことによって、この三通りの中からどれか一つ選んで検定を行う。

②検定統計量と標本分布の決定

仮説検定に用いる統計量を**検定統計量**という。母平均の仮説検定では、検定統計量には標本平均$\bar{X}$を用いる。

②検定統計量と標本分布の決定

中心極限定理によれば、$\bar{X}$は（近似的に）$\mathbf{N}(\mu, \sigma^2/n)$に従う。

②検定統計量と標本分布の決定

$\bar{X}$から新たに検定統計量Zを定義する。母分散が既知か未知かでZの定義は少し異なる。

②検定統計量と標本分布の決定

母分散σ^2が既知の場合の検定統計量

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} = (\bar{X} - \mu_0)\frac{\sqrt{n}}{\sigma}$$

②検定統計量と標本分布の決定

母分散σ^2が未知でnが大きい$(n \geq 30)$場合の検定統計量

$$Z = \frac{\bar{X} - \mu_0}{s/\sqrt{n}} = (\bar{X} - \mu_0)\frac{\sqrt{n}}{s}$$

母分散σ^2が未知の場合は、代わりに不偏分散s^2を使う。

②検定統計量と標本分布の決定

$\bar{X}$は近似的にN$(\mu, \sigma^2/n)$に従うので
Zは近似的にN(0, 1)に従う。

③有意水準と棄却域の決定

有意水準をαとすると、
棄却域Rは次のようになる。

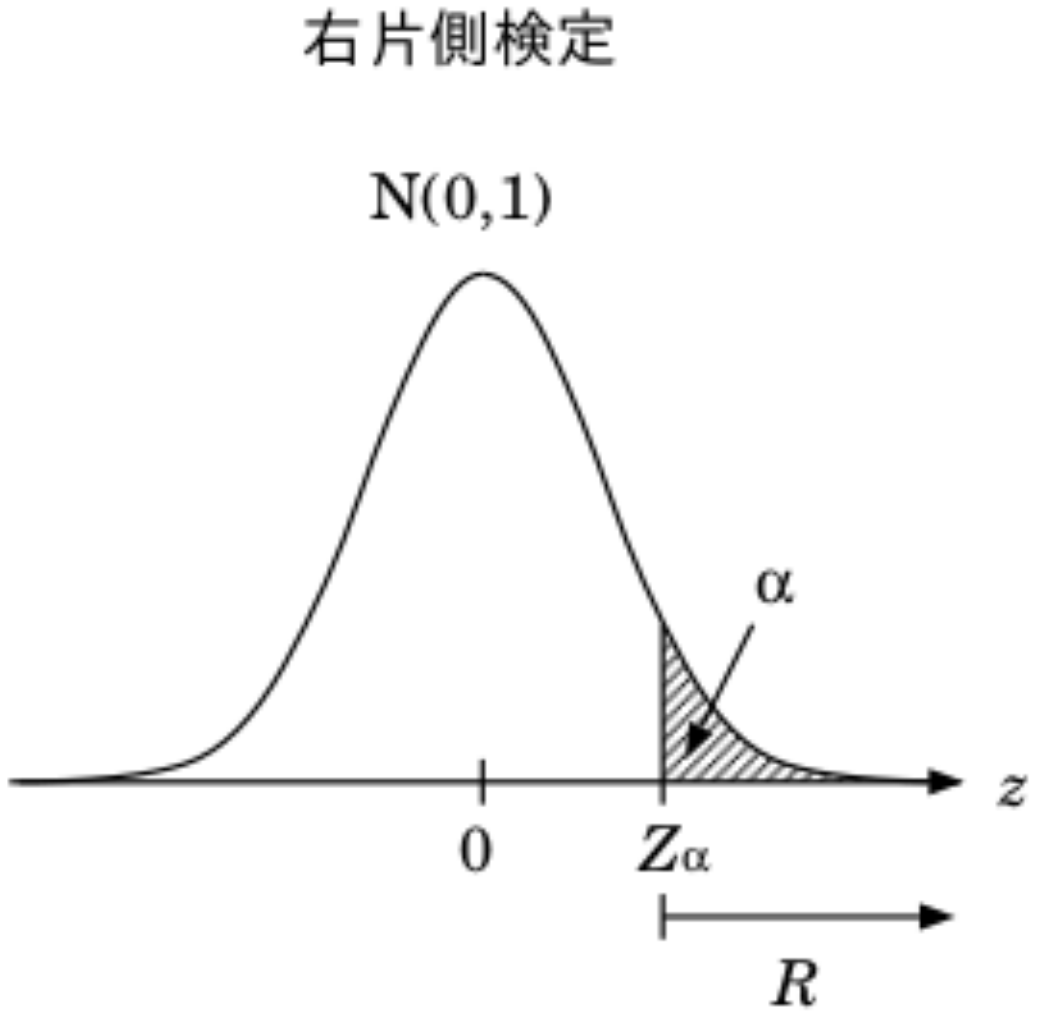
③有意水準と棄却域の決定

右片側検定の場合

$$R = \{Z; Z > Z_\alpha\}$$

棄却域はZがZ_αより大きい範囲

ここで、Z_αは上側確率αのパーセント点

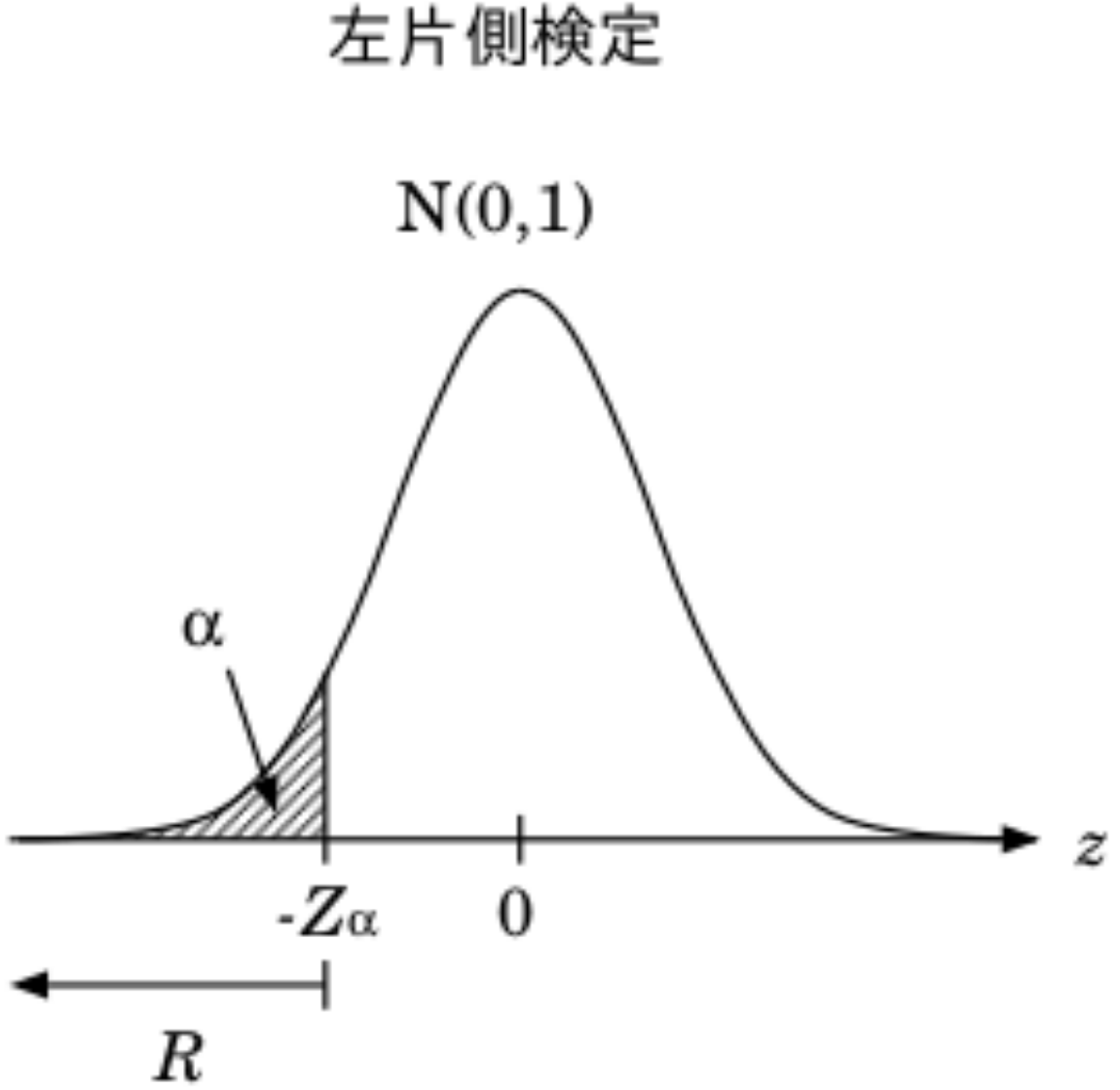


③有意水準と棄却域の決定

左片側検定の場合

$R = \{Z; Z < -Z_\alpha\}$

棄却域はZが$-Z_\alpha$より小さい範囲



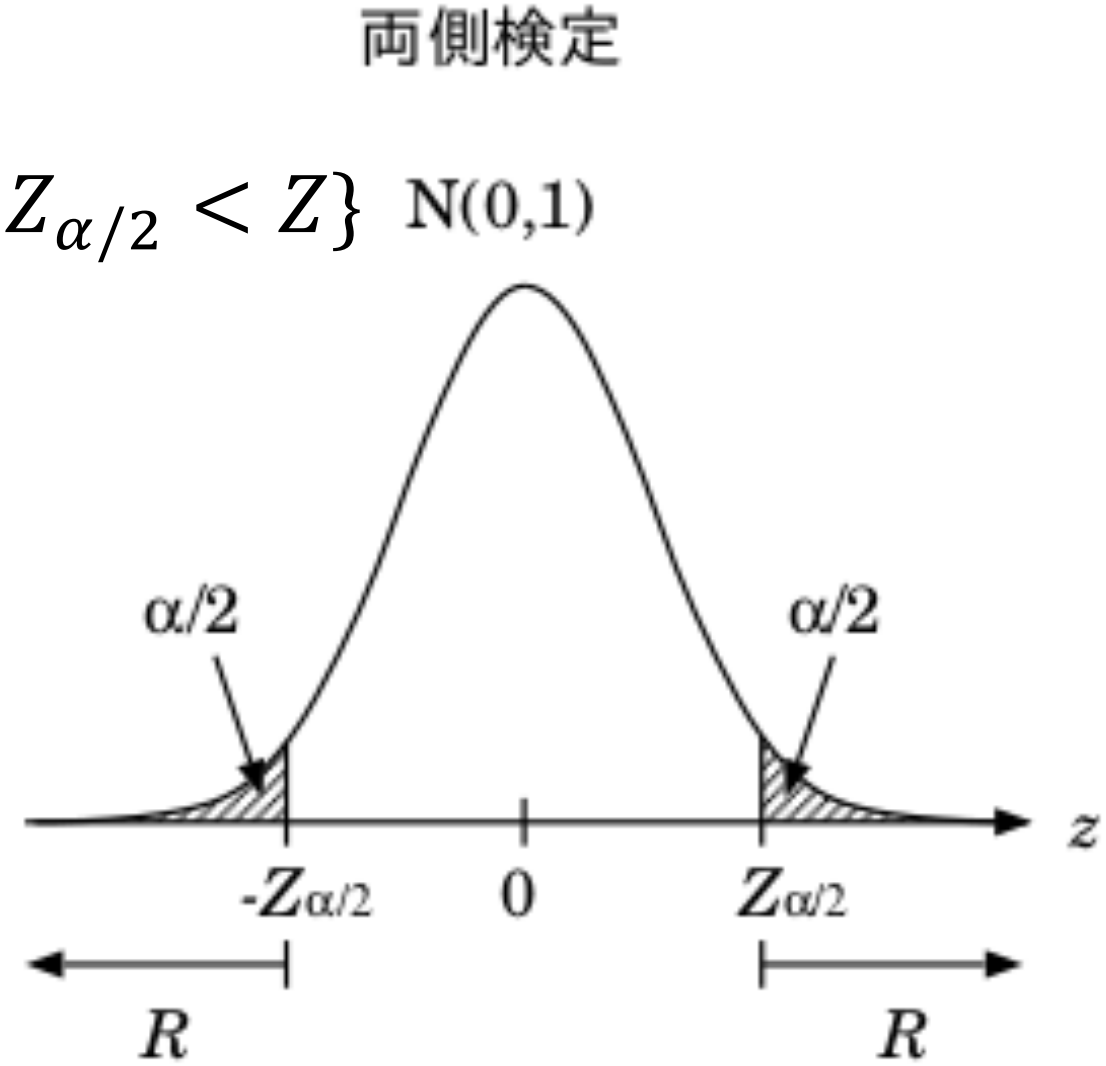
③有意水準と棄却域の決定

両側検定の場合

$R = \{Z; |Z| > Z_{\alpha/2}\} = \{Z; Z < -Z_{\alpha/2}, Z_{\alpha/2} < Z\}$

棄却域はZが$-Z_{\alpha/2}$より小さいか$Z_{\alpha/2}$より大きい範囲

ここで、$Z_{\alpha/2}$は上側確率$\alpha/2$のパーセント点



④検定の実施

標本から求めた検定統計量が棄却域Rに入るかどうかを調べる。

④検定の実施

棄却域Rに入った場合

帰無仮説H_0を棄却して、対立仮説H_1を採択する。つまり、目的通り対立仮説が検証されたことになる。

④検定の実施

棄却域Rに入らなかった場合

帰無仮説H_0を棄却できないので、対立仮説H_1は採択されない。つまり、目的通り対立仮説を検証するに足る証拠が得られなかったことになる（対立仮説が否定された訳ではない）。

④検定の実施

以上のような検定は検定統計量にZを用いるので**Z検定**という。

NORM.INV関数

N(0, 1)のパーセント点Z_αは
NORM.INV関数で求められる。

関数	説明
NORM.INV(1-α, 0, 1)	右側（上側）確率 α における $\mathrm{N}(0,1)$ のパーセント点 Z_α を返す。

演習：仮説検定

dataset1.odsはある母集団から抽出した500人の標本データである。この中の体重のデータから標本平均と不偏標準偏差を計算すると以下の値になる。

- 標本平均$\bar{X} = 56.66$ [kg]
- 不偏標準偏差$s = 9.50$ [kg]

演習：仮説検定

母分散σ^2が未知として、この標本データから

「体重の母平均μは56 kgより大きい」
$(\mu > 56)$

といえるかを有意水準5% $(\alpha = 0.05)$で検定する。

①仮説の設定

検証したいのは、母平均$\mu > 56$といえるかである。そこで、次の仮説を設定する。

- 帰無仮説$H_0: \mu = 56$
- 対立仮説$H_1: \mu > 56$

これは**右片側検定**である。

②検定統計量と標本分布の決定

母分散σ^2が未知なので検定統計量は次を使う。

$$Z = (\bar{X} - 56)\frac{\sqrt{n}}{s}$$

$n = 500$と大きいのでZは近似的にN(0,1)に従う。

③有意水準と棄却域の決定

有意水準$\alpha = 0.05$の右片側検定なので、$Z_\alpha = Z_{0.05}$を求める。

LibreOffice Calcの空いているセルに「**=NORM.INV(1-0.05, 0, 1)**」と入力しよう。

③有意水準と棄却域の決定

この結果、$Z_{0.05} \simeq 1.64$であることがわかる。したがって、右片側検定の棄却域Rは次のようになる。

$$R = \{Z; Z > 1.64\}$$

④検定の実施

標本から$n = 500, \bar{X} = 56.66, s = 9.50$なので、②のZの式に代入して計算する。

$$Z = (56.66 - 56) \times \frac{\sqrt{500}}{9.50} \simeq 1.55$$

LibreOffice Calcの空いているセルに「**=(56.66-56)*SQRT(500)/9.50**」と入力してZを計算してみよう。SQRT関数は平方根（ルート）を計算する関数である。

④検定の実施

$Z \simeq 1.55 < 1.64$なので、③で求めた棄却域$R = \{Z; Z > 1.64\}$に入らない。したがって、帰無仮説は棄却されないので、対立仮説は採択されない。

つまり、有意水準5%で「体重の母平均μは56 kgより大きい」と言えない。

④検定の実施

標本平均$\bar{X} = 56.66$は明らかに56 kgより大きいのにも関わらず、体重の母平均は56 kgより大きいとは簡単に言えないことに注意してほしい。

今回の問い

- 仮説検定とは何か？
  - 母集団についての仮説を標本に基づき統計的に検証すること。
- **仮説検定はどのような手順で行えば良いか？**
  - **仮説の設定→検定統計量と標本分布の決定→有意水準と棄却域の決定→検定の実施の順に行う。**
- p値とは何か？

今回学ぶこと

1. 仮説検定の考え方
2. 母平均の仮説検定
3. **p値**

p値

前述の手順では、有意水準αからパーセント点Z_αを求めて棄却域Rを決定したが、αがどのくらいなら帰無仮説が棄却されるかを知るためには少々手間である。

p値

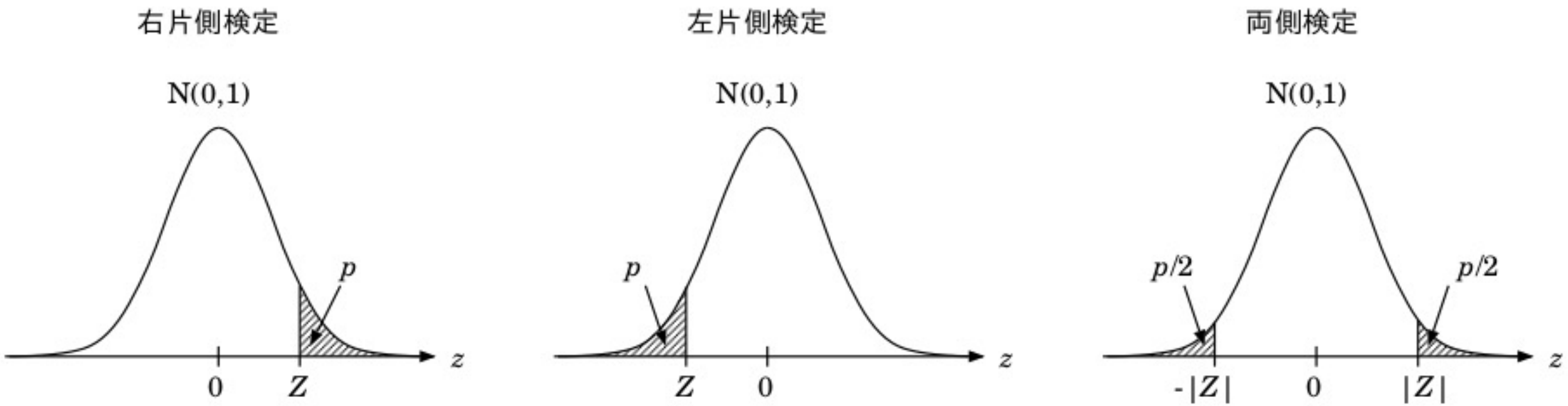
標本から得られた検定統計量の値から（仮に**帰無仮説が正しいとしたときの**）**標本分布の確率**を求めることが一般に行われる。

この確率を**p値**という。標本から得られた検定統計量の値がパーセント点となる有意水準ともいえる。

p値

帰無仮説を棄却するかどうかを、棄却域の代わりにp値を使って判断することがよく行われる。

Z検定の場合のp値



図中のZは標本から得られた検定統計量の値、$|Z|$はZの絶対値である。

p値の意味

p値より大きい有意水準αで帰無仮説は棄却される。

p値の意味

例えば、p値が$p = 0.012$なら帰無仮説は有意水準5% $(\alpha = 0.05)$で棄却されるが、1% $(\alpha = 0.01)$では棄却されない。

$$0.01 < 0.012 < 0.05$$

p値の定義は片側検定か両側検定かによって異なるので注意が必要。

Z.TEST関数

LibreOffice CalcやMicrosoft Excelでは、p値はZ.TEST関数で求められる。

右片側検定の場合

関数	説明
Z.TEST(データ範囲, μ_0)	データ範囲の標本平均から Z 検定による右片側検定の p 値を求める。母標準偏差 σ が未知の場合に用いる。
Z.TEST(データ範囲, μ_0, σ)	データ範囲の標本平均から Z 検定による右片側検定の p 値を求める。母標準偏差 σ が既知の場合に用いる。

左片側検定の場合

関数	説明
1-Z.TEST(データ範囲, μ_0)	データ範囲の標本平均から Z 検定による左片側検定の p 値を求める。母標準偏差 σ が未知の場合に用いる。
1-Z.TEST(データ範囲, μ_0, σ)	データ範囲の標本平均から Z 検定による左片側検定の p 値を求める。母標準偏差 σ が既知の場合に用いる。

両側検定の場合

関数	説明
MIN(Z.TEST(データ範囲, μ_0), 1-Z.TEST(データ範囲, μ_0))*2	データ範囲の標本平均から Z 検定による両側検定の p 値を求める。 母標準偏差 σ が未知の場合に用いる。
MIN(Z.TEST(データ範囲, μ_0, σ), 1-Z.TEST(データ範囲, μ_0, σ))*2	データ範囲の標本平均から Z 検定による両側検定の p 値を求める。 母標準偏差 σ が既知の場合に用いる。

演習：p値の計算

dataset1.odsでp値を計算する。p値を求めるにはZ.TEST関数を使う。

右片側検定なので、LibreOffice Calcの空いているセルに「**=Z.TEST(E2:E501, 56)**」と入力しよう。

演習：p値の計算

その結果、$p \simeq 0.0596$が得られる。

$$0.05 < 0.0596 < 0.1$$

より、帰無仮説は有意水準5% $(\alpha = 0.05)$なら棄却されないが、10% $(\alpha = 0.1)$なら棄却される。

つまり、有意水準10%なら「体重の母平均μは56 kgより大きい」といえる。

今回の問い

- 仮説検定とは何か？
  - 母集団についての仮説を標本に基づき統計的に検証すること。
- 仮説検定はどのような手順で行えば良いか？
  - 仮説の設定→検定統計量と標本分布の決定→有意水準と棄却域の決定→検定の実施の順に行う。
- **p値とは何か？**
  - **帰無仮説が正しいと仮定したときに、観測されたデータ以上に極端な結果が得られる確率。**

今回学んだこと

1. 仮説検定の考え方
2. 母平均の仮説検定
3. p値

今回の目標

母集団についての仮説が
正しいかを標本から検証
できるようになる。

課題

dataset1.odsにおいて、500人の身長のデータを母集団とみなす。この母集団から50人の標本を無作為抽出して、その標本平均から

「身長の母平均μは160 cmではない」($\mu \neq 160$)

といえるかを仮説検定するために、p値を求めなさい。これは**両側検定**になる。

課題

50人の標本抽出（復元抽出）は以下のようにして行う。

1. dataset1.odsに新たなシートを追加し、セルA1からE501をコピー&ペーストする。
2. セルG1に「標本番号」、H1に「身長[cm]」と入力する。
3. セルG2に「**=RANDBETWEEN(1, 500)**」と入力する。
4. セルH2に「**=VLOOKUP(G2, A$2:E$501, 4, 0)**」と入力する。
5. セルG2をG3からG51までコピー&ペーストする。
6. セルH2をH3からH51までコピー&ペーストする。

これで列Hに50人の身長のデータが標本抽出される。

課題

標本のp値は空いているセルに計算するが、p値と分かるようにしておく。

作成したファイルはODF表計算ドキュメント(拡張子.ods)で提出すること。

提出先：yutaka.hirai@koeki-u.ac.jp

件名：データリテラシー第12回課題 [学籍番号]

締め切り：1月8日 (金)

次回

第１回　社会におけるデータ・AI利活用
第２回　表計算ソフトの基本操作
第３回　セルの参照と集計
第４回　度数分布表とヒストグラム
第５回　代表値と散らばりの尺度
第６回　データのグラフによる表現とデータ比較
第７回　データの分布と標本抽出
第８回　推定
第９回　分割表とクロス集計
第１０回　２変量データと相関
第１１回　時系列データ
第１２回　仮説検定
第１３回　回帰分析