J. Shang
Department of Socio-Economic System, Graduate School of Economics
University of Nagoya

01/07/2024

DATA SCIENCE — Estimation Statistics

1 Logistic Regression

ロジスティクス回帰式:

$$P = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1)}} \tag{1-1}$$

これを理解するために、まず線形回帰（単純回帰）式を見る:

$$y = \beta_0 + \beta_1 x \tag{1-2}$$

x:説明変数、独立変数、何かの原因となった変数

y:被説明変数、従属変数、目的変数、その原因を受けて発生した結果を表す変数

yの代わりに確率Pを取るとする。しかし、ここで問題がある。確率の範囲は[0,1]であることがわかっている。この問題を克服するために、Pの「オッズ」[1]の対数を取ることを考える。即ち、ロジット変換[2]を行う。

変換後の式は：

$$log(\frac{P}{1-P}) = \beta_0 + \beta_1 x \tag{1-3}$$

Pのオッズは：

$$\frac{P}{1-P} = exp^{\beta_0 + \beta_1 x} \tag{1-4}$$

pを解くと：

$$\begin{aligned} P &= \frac{exp^{\beta_0 + \beta_1 x}}{1 + exp^{\beta_0 + \beta_1 x}} \\ &= \frac{1}{1 + exp^{-(\beta_0 + \beta_1 x)}} \end{aligned} \tag{1-5}$$

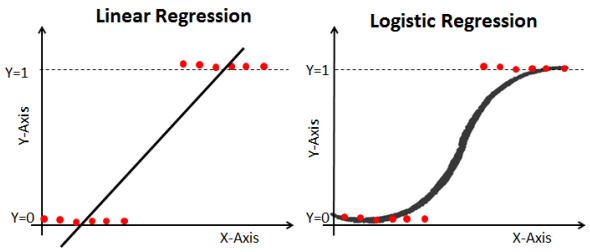

Figure 1: 線形回帰　vs　ロジスティクス回帰

[1]オッズ(odds)とは、成功率と失敗率の比

[2]ロジット変換とは：PをPのオッズの対数に変換すること、目的は値が0から1までの範囲しか取れないデータを$(-\infty, +\infty)$のデータに変換する:$logit(P) = log(\frac{p}{1-p})$

2 Simple linear regression単純線形回帰モデルの推定：最小二乗法

単純線形回帰モデルの推定回帰式：

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \tag{1-6}$$

Where:

$\hat{\beta_0}$: 定数項（或いは切片）の推定値/推定量
$\hat{\beta_1}$: 回帰係数の推定値/推定量
$\hat{y}_i$: 予測値（或いは理論値）

$y_i - \hat{y}_i$の差を$e_i = y_i - \hat{y}_i$と表し、（回帰）残差と呼ぶ[3]。図2を参照。

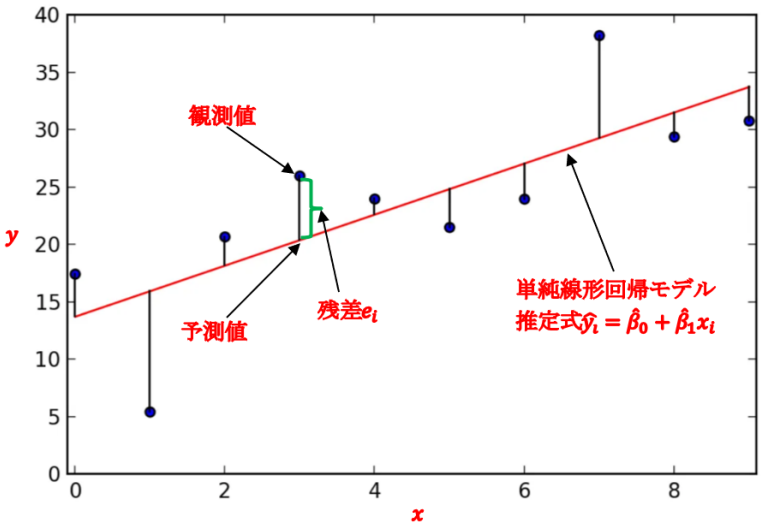

Figure 2:

未知パラメーターの推定量を推定するためには、できるだけ残差を小さくなるようにしたらいい。しかし、残差はプラスとマイナスの値が相殺されるので、残差二乗和(residual sum of squares: RSS)を最小にすることを考える：

$$RSS = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 \tag{1-7}$$

残差二乗和を最小になるように$\hat{\beta_0}, \hat{\beta_1}$を求める方法を最小二乗法(Least Squares Method)と呼び、このように求めた$\hat{\beta_0}, \hat{\beta_1}$を最小二乗推定量と呼ぶ。 $\hat{\beta_0}, \hat{\beta_1}$は下記のようにRSSを微分して0にして：

$$\begin{aligned} \frac{\alpha RSS}{\alpha \hat{\beta_0}} &= \sum_{i=1}^{n} (-2)(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) \\ \frac{\alpha RSS}{\alpha \hat{\beta_1}} &= \sum_{i=1}^{n} (-2x_i)(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) \end{aligned} \tag{1-8}$$

また、下記の標本の性質を用いて：

$$\begin{aligned} \bar{x} &= \frac{1}{n} \sum_{i=1}^{n} x_i \\ \bar{y} &= \frac{1}{n} \sum_{i=1}^{n} y_i \end{aligned} \tag{1-9}$$

[3]残差と誤差は違うことを注意。誤差とは観測された値y_iと真の値Yの差のこと、母集団：$Y = \beta_0 + \beta_1 X + \epsilon$とし、$\epsilon$は誤差項と呼ぶ。

式1-8と式1-9を$\hat{\beta_0}, \hat{\beta_1}$について解くと：

$$\begin{cases} \hat{\beta_1} & = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2} \\ \hat{\beta_0} & = \bar{y} - \hat{\beta_1}\bar{x} \end{cases} \tag{1-10}$$

最後、回帰式の当てはまりの良さを評価するため、基準の一つとして、決定係数R^2が用いられる：

$$\begin{aligned} R^2 &= \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \\ &= 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \\ &= 1 - \frac{\sum_{i=1}^{n}(e_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \end{aligned} \tag{1-11}$$

Where:

$$\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 : \text{は回帰モデルで説明できる変動}$$

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 : \text{はyの全変動}$$

3 確率

確率：事象Aが起こる確率は$n(A)/N$である。 $n(A)$とは事象Aの起こった回数、Nは標本の大きさ（サンプルサイズ）[4]

3.1 離散型確率変数と確率分布

サイコロの出目の確率分布：

Xの取る値(即ちXの実現値)	1	2	3	4	5	6	計
$P(X = x_i)$	1/6	1/6	1/6	1/6	1/6	1/6	1

Table 1: サイコロの出目の確率分布表

X：確率変数
x_i：Xの実現値

それぞれの実現値をとる確率を$p_1, p_2, ..., p_i, ...$とする、確率p_iは確率変数Xの取りうる値によって変わる。この意味で、p_iはXの取りうる値の関数として見ることができる：

$$P(X = x_i) = p_i = f(x_i), \quad i = 1, 2... \tag{1-12}$$

この$f(x_i)$を確率変数Xの**確率関数**(probability function).
確率関数に関して、確率は非負であり、確率の総和が1となることより、

$$\begin{aligned} p_i = f(x_i) \geq 0, \quad i = 1, 2... \\ \sum_i p_i = \sum_i f(x_i) = 1 \end{aligned} \tag{1-13}$$

Xが$f(x_r)$以下の値をとる確率を**分布関数**という：

$$F(x) = P(X \leq x_r) = \sum_{i}^{r} p_i = \sum_{i}^{r} f(x_i), \quad i = 1, 2... \tag{1-14}$$

分布関数の性質は：$F(-\infty) = 0, F(\infty) = 1$

[4]標本/サンプルの大きさと標本数の違いを注意！標本/サンプルの大きさとは、分析対象とされている母集団から一回の無作為抽出したサンプルのデータの個数のこと；その一方、標本数とは標本の個数、母集団から無作為抽出した回数のこと。

3.2 連続型確率変数と確率分布

Xが1,2,3...のような不連続な値をとる確率変数を**離散型確率変数**という。
これに対して、確率変数の実現値が連続した値（任意の実数値）をとる場合、**連続型確率変数**といい、その確率分布を**連続型確率分布**という。

しかし、Xが連続型であるとき、Xの確率分布をTable1のような対応表で表すことができない。この時、確率を分配する規則は、連続曲線によって表される。この曲線を確率変数Xの**確率密度関数**、或いは**密度関数**という、正規分布を例として：

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} exp(-\frac{(x-\mu)^2}{2\sigma^2}) \tag{1-15}$$

Where:

μ : 平均(mean)

σ^2 : 分散(variance)

σ : 標準偏差(standard deviation)

$exp(x) = e^x$ 指数関数、eは自然対数の底、2.71828...

離散型と同じで、Xがx_r以下の値をとる確率を**分布関数**といい、離散型確率変数の場合は合計で計算、連続型確率変数の場合は積分で計算。

正規分布の分布関数は：

$$\begin{aligned} F(x) = P(X \leq x_r) &= \int_{-\infty}^{x_r} f(x)dx \\ &= \int_{-\infty}^{x_r} \frac{1}{\sqrt{2\pi}\sigma} exp(-\frac{(x-\mu)^2}{2\sigma^2})dx \end{aligned} \tag{1-16}$$

このような平均μ、分散σ^2の正規分布を$N(\mu, \sigma^2)$と表し[5]、確率変数Xが正規分布$N(\mu, \sigma^2)$に従うことを

$$X \sim N(\mu, \sigma^2) \tag{1-17}$$

と表す。

正規分布の確率密度関数(式1-15)を図3で表せば、図3のように、赤色の曲線を正規曲線という。下記の性質がある：

1. 正規曲線の値は正であり、正規曲線の下側の面積の総和は1となる。
2. 正規曲線は平均$x = \mu$で左右対称
3. 正規分布の平均値、中央値、最頻値はすべてμと等しい

正規分布の中で、特に平均が0、分散が1の正規分布を標準正規分布といい、$N(0, 1)$と表す。
標準正規分布の確率密度関数、或いは密度関数は：

$$f(x) = \frac{1}{\sqrt{2\pi}} exp(-\frac{x^2}{2}) \tag{1-18}$$

Where:

$exp(x) = e^x$ 指数関数、eは自然対数の底、2.71828...

標準正規分布の分布関数は：

$$\begin{aligned} f(x) = P(X \leq x) &= \int_{-\infty}^{x} f(t)dt \\ &= \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} exp(-\frac{t}{2})dt \end{aligned} \tag{1-19}$$

[5] NはNormal Distribution正規分布の頭文字

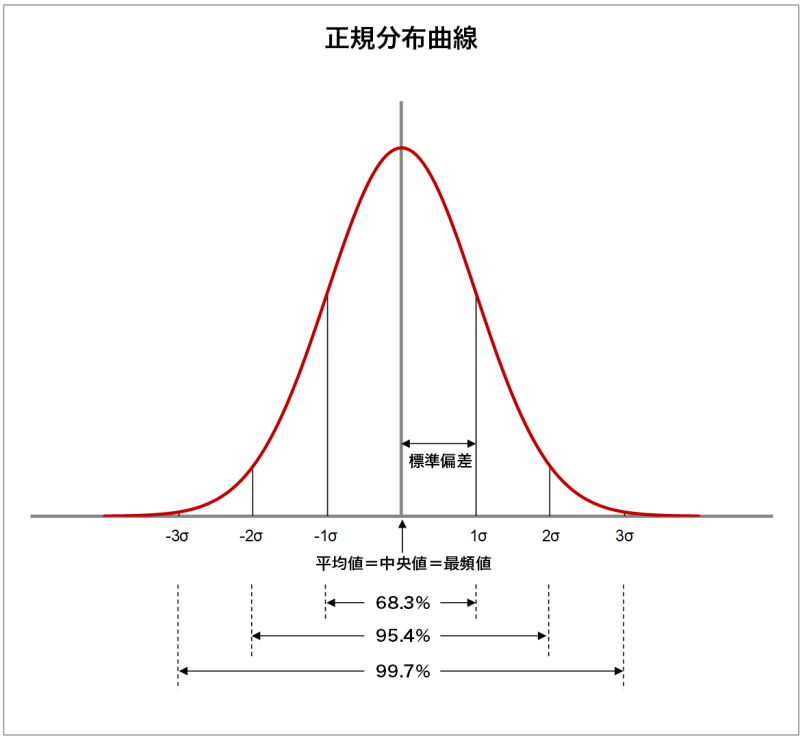

Figure 3:

任意の正規分布$N(\mu, \sigma^2)$は下記の標準化を行うことによって、標準正規分布$N(0,1)$に変換することができる。$X \sim N(\mu, \sigma^2)$とすると、

$$Z = \frac{X - \mu}{\sigma} \sim N(0,1) \tag{1-20}$$

ZはXを標準化した確率変数、標準正規分布に従う。

正規分布に従う母集団から無作為抽出した標本の標本平均$\bar{X}$は平均μ、分散σ^2/nの正規分布に従う：

$$\bar{X} \sim N(\mu, \sigma^2/n) \tag{1-21}$$

Where:

μ : 母集団平均(mean)

σ^2 : 母集団分散(variance)

σ^2/n : 標本の分散

$\sigma/\sqrt{n}$: 標本の標準偏差

その標準化した変数$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$も標準正規分布に従う$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$

3.3 正規分布表の使い方

今、確率変数Zが標準正規分布に従っているとする。Zが1.96より大きくなる確率$P(Z > 1.96)$を求めるにはどうすればいいのか。
一般に、連続型確率分布の確率の計算は、確率密度関数の積分を行う必要があるが、ここで簡単化するため、正規分布表を活用しましょう。
正規分布表には、$N(0,1)$の**上側確率**が計算されている。**上側確率**とは、確率変数Zがある値（z）よりも大きくなる確率$P(Z > z)$のこと(図4参照)。お手元の付表1_正規分布表をご覧ください。正規分布表の左端はzの小数点第1位までの数値が与えられており、上端にはzの小数点第2位の数値が与えられている。確率$P(Z > 1.96)$を求める場合、左端の1.9と上端の0.06が交差するところ、即ち、0.025が求める確率となる。図5を参照。

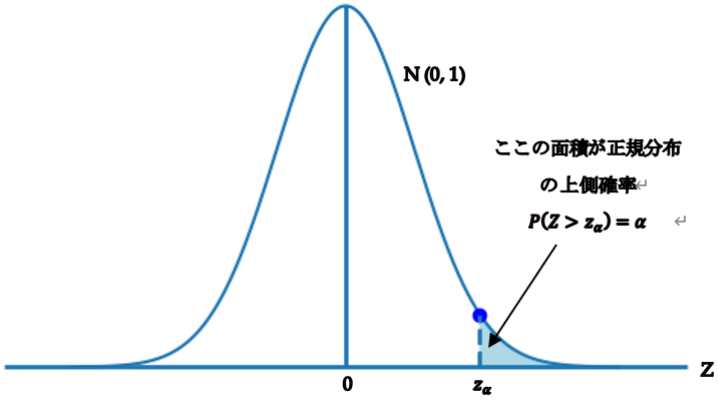

Figure 4:

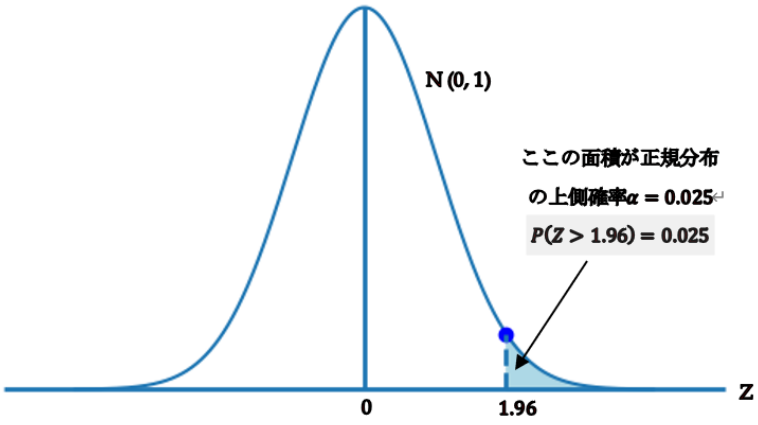

Figure 5:

3.4 母集団平均の区間推定：母集団分散が既知の場合

母集団平均が２つの値の間、即ち、ある区間にあるとして推定することは**区間推定**という。大きさnの無作為標本$X_1, X_2, ..., X_n$の標本平均$\bar{X}$の標本分布は$N(\mu, \sigma^2(\bar{X})$である。式1-21を使うと、

$$Z_n = \frac{\bar{X} - \mu}{\sigma(\bar{X})} \sim N(0, 1) \tag{1-22}$$

Where:

$$\sigma(\bar{X}) = \sigma/\sqrt{n}$$

図6の黄色塗りつぶした部分は確率$1 - \alpha$とする。母平均μの**信頼係数**が$1 - \alpha$の時の**信頼区間**を求めるのは、下記の式を満たす$z_{\alpha/2}$の値を見つけたらいい。

$$P(|Z_n| < z_{\alpha/2}) = P(|\frac{\bar{X} - \mu}{\sigma(\bar{X})}| < z_{\alpha/2}) = 1 - \alpha \tag{1-23}$$

同時に、$z_{\alpha/2}$は標準正規分布の上側確率が$\alpha/2$となる点でもある。即ち、正規分布表から、上側確率が$\alpha/2$となる$z_{\alpha/2}$を見つかったらいい。

$\alpha = 0.05$の時、$z_{\alpha/2} = 1.96$であり、$\alpha = 0.10$の時、$z_{\alpha/2} = 1.645$である。

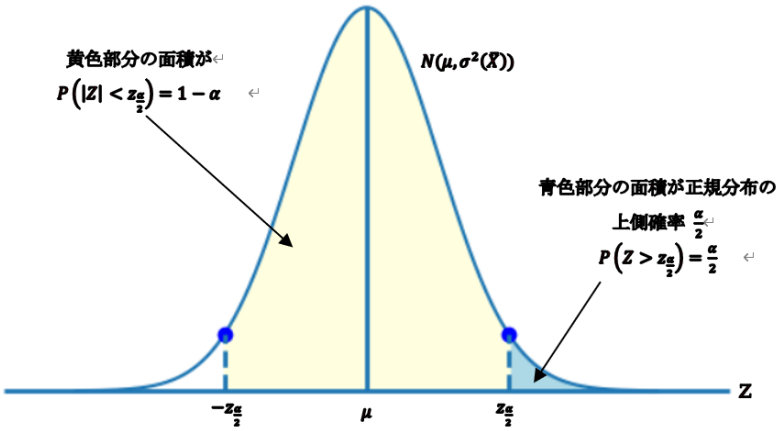

Figure 6:

母平均μの**信頼係数**が$1-\alpha$の時の**信頼区間**は式1-23で求める。式1-23をμについて解くと、下記のようになる：

$$P(\bar{X} - z_{\alpha/2}\sigma(\bar{X}) < \mu < \bar{X} + z_{\alpha/2}\sigma(\bar{X})) = 1 - \alpha \tag{1-24}$$

これは母平均μが区間$(\bar{X} - z_{\alpha/2}\sigma(\bar{X}), \bar{X} + z_{\alpha/2}\sigma(\bar{X}))$に含まれる確率が$1-\alpha$であることを示している。

確率変数$\bar{X}$をその実現値$\bar{x}$で置き換えた区間：$(\bar{x} - z_{\alpha/2}\sigma(\bar{x}), \bar{x} + z_{\alpha/2}\sigma(\bar{x}))$を**母平均$\mu$の信頼係数が$1-\alpha$であるときの信頼区間**という。信頼区間の上限と下限を信頼限界という。

例題1：正規母集団$N(\mu, 2^2)$から大きさ16の標本をとって標本平均を計算したら、$\bar{x} = 3.2$であった。μの信頼係数が0.95(95%と表している時もある)の信頼区間を求めよ。
解信頼係数が0.95、即ち$\alpha = 0.05$に対する$z_{\alpha/2} = 1.96$である。$n = 16, \sigma = 2$であるので、$\bar{X}$の標準偏差は：

$$\sigma(\bar{x}) = \frac{\sigma}{\sqrt{n}} = \frac{2}{\sqrt{16}} = 0.5$$

となる。信頼限界は

$$\bar{x} \pm z_{\alpha/2}\sigma(\bar{x}) = 3.2 \pm 1.96 \times 0.5 \tag{1-25}$$

から計算すると、2.22と4.18となる。μの信頼係数が0.95の信頼区間は$(2.22, 4.18)$である。

3.5 母集団平均の区間推定：母集団分散が未知の場合

母集団分散が未知の場合、一つの標本から**標本(不偏)分散**が下記の式で求められる：

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}) \tag{1-26}$$

Where:

n : 標本(サンプル)の大きさ

証明はここで示さないが、結論だけを提示する。$Z = (\bar{X} - \mu)/(\sigma/\sqrt{n})$の$\sigma$を$S$で置き換えた統計量はt統計量といい：

$$T_n = \frac{\bar{X} - \mu}{S/\sqrt{n}} \tag{1-27}$$

は自由度$k = n - 1$のt分布に従う。よって、t分布表から：

$$P(|T_n| < t_{\alpha/2}(k)) = P(|\frac{\bar{X} - \mu}{S/\sqrt{n}}| < t_{\alpha/2}(k)) = 1 - \alpha \tag{1-28}$$

を満たすt分布の上側100αパーセント点$t_{\alpha/2}(k)$の値を見つけることができる。例えば、$n = 11$、即ち、$k = 10$、$\alpha = 0.05$の時、t分布表から$t_{\alpha/2}(k) = 2.228$である。

式1-28をμについて解くと：

$$P(\bar{X} - t_{\alpha/2}(k)\frac{S}{\sqrt{n}} < \mu < \bar{X} + t_{\alpha/2}(k)\frac{S}{\sqrt{n}}) = 1 - \alpha \tag{1-29}$$

となる。信頼係数確率変数$(\bar{X}, S)$をその実現値$(\bar{x}, s)$で置き換えた区間：$(\bar{x} - t_{\alpha/2}(k)\frac{s}{\sqrt{n}}, \bar{x} + t_{\alpha/2}(k)\frac{s}{\sqrt{n}})$を**母平均$\mu$の信頼係数が$1 - \alpha$であるときの信頼区間**という。

例題2：正規母集団$N(\mu, \sigma^2)$から大きさ9の標本をとって標本平均と標本標準偏差を計算したら、それぞれ$\bar{x} = 3.2, s = 2.1$であった。μの信頼係数が0.95(95％と表している時もある)の信頼区間を求めよ。
解信頼係数が0.95、即ち$\alpha = 0.05$、$n = 9, k = 8, \bar{x} = 3.2, s = 2.1$なので、$t_{0.025}(8) = 2.306$である。信頼係数0.95の信頼限界は：

$$\bar{x} \pm t_{\alpha/2}(k)\frac{s}{\sqrt{n}} = 3.2 \pm 2.306 \times \frac{2.1}{\sqrt{9}} \tag{1-30}$$

から計算できる。μの信頼係数が0.95の信頼区間は$(1.586, 4.814)$である。

3.6 仮説検定の考え方

例題3：ある乗用車の燃費は従来車では平均17km/L、標準偏差は2km/Lの正規分布に従うという。改良車が開発されて、16台の走行テストを行ったところ、平均18km/Lということが分かった。では、改良車の燃費は従来車より良くなったと言えるか？ただし、改良車の燃費は正規分布に近似できるものとし、改良車の標準偏差も従来車と同じで2km/Lであるものとする。

分析と仮説検定の説明：題目より従来車~ $N(17, 2^2)$、改良車という母集団から無作為抽出した標本は標本分布$N(\mu, 2^2)$に従う、母集団の平均値のμは知らないことを注意する。

抽出した16台の平均値は18km/Lで、17km/Lより大きいが、改良したと見えるが、これは本当に改良車の母集団を代表できると言えるか？たまたま燃費が高い車ばかりを抽出したのではないか？このことを統計的に判断したい時は仮説検定が使われる。

仮説検定では、まず**帰無仮説(null hypothesis)**、**対立仮説(alternative hypothesis)**を立てる。

一般的に、**捨てたい仮説を帰無仮説にする**ことが多い。これに対して、**証明したい仮説を「対立仮説」**と呼ぶ。基本的な考え方は、「もし帰無仮説が正しいとしたら、今回取得できたデータが得られる確率は、どれぐらい小さいのか」を計算する。その確率が、一定の水準(例えば、0.05)未満であれば、帰無仮説は誤りと判断され、対立仮説が採択される。

仮説検定を行って、帰無仮説を捨てると、帰無仮説を**棄却(reject)**といい、帰無仮説を捨てずに採用することを**採択(accept)**という。

例題に戻る。改良車と従来車を比較すると、三つのパータンしか考えられない。従来車のの燃費値が17km/Lであるので、

$$\begin{cases} \text{もし改良車母集団の燃費が従来車と同じならば、} \mu = 17 \\ \text{もし改良車母集団の燃費が従来車より良くなったら、} \mu > 17 \\ \text{もし改良車母集団の燃費が従来車より悪くなったら、} \mu < 17 \end{cases}$$

ここで、走行テスト結果が18km/L、従来車の平均17km/Lより大きい、「改良車の燃費は従来車より良くなったと言えるか」という題目から、$\mu < 17$を考える必要がない、従って、帰無仮説と対立仮説を立てる：

$$\begin{cases} \text{帰無仮説：} H_0 : \mu = 17 \text{(改良車の燃費は従来車と同じ)} \\ \text{対立仮説：} H_0 : \mu > 17 \text{(改良車の燃費は従来車より良くなった)} \end{cases}$$

X_iを走行テストで使われた各車の燃費を表す確率変数とする。帰無仮説が正しいという条件のもとでは、$\mu = 17$、分散は既知2^2、$X_i \sim N(17, 2^2), i = 1, 2, ..., 16$、標本の大きさは16、標本平均の分布は：

$$\bar{X} = \frac{1}{16}\sum_{i=1}^{16} X_i \sim N(17, \frac{2^2}{16}) \tag{1-31}$$

走行テストの結果が18km/L、即ち、$\bar{X}$の実現値が18。
今、式1-31で与えられた確率変数$\bar{X}$が18よりも大きい確率を計算すれば、帰無仮説が正しい時、18という実現値がどの程度で起こりうるかを知ることができる。標準化して確率を計算すると：

$$\begin{aligned} P(\bar{X} \geq 18) &= P(\frac{\bar{X} - 17}{2/\sqrt{16}} \geq \frac{18 - 17}{2/\sqrt{16}}) \\ &= P(Z \geq 2) \\ &= 0.0228 = 2.28\%(\text{ 正規分布表より}) \end{aligned}$$

この確率をp値という。p値が0.0228ということは、$H_0 : \mu = 17$が正しいもとで、改良車16台の走行テストを100繰り返したら、$\bar{X}$の実現値が18以上になるのは2回程度しかない珍しい事象であることを意味する。
p値が小さければ小さいほど、帰無仮説H_0は正しいという可能性は小さく、対立仮説H_1が正しいと判断される。この判断の基準となる確率の値を有意水準といい、通常αと表す。慣例として、$\alpha = 0.01(1\%), 0.05(5\%), 0.10(10\%)$がよく使われる。$\alpha = 0.01(1\%)$の意味は、帰無仮説のもとで起こる事象が100回の繰り返し実験で5回しか起こらないことで、珍しい事象だと判断して帰無仮説を棄却する(対立仮説を採択する)ことを意味する。

例題3の仮説検定の結論：

$$\begin{cases} \text{有意水準の0.05では、帰無仮説は棄却され、対立仮説を採択する。即ち、改良車の燃費は従来車より良くなったと言える。} \\ \text{有意水準の0.01では、帰無仮説は棄却できず採択されることになる。即ち、改良車の燃費は従来車より良くなったと言えない。} \end{cases}$$

今までは$P(\bar{X} \geq 18)$を直接計算する方法もあるが、より簡単のほうは、まず事前決まった有意水準αの棄却点を見つかる、標準化した検定統計値を計算する、もし検定統計値のほうが棄却点より大きいならば、検定統計値は棄却域に入るので、帰無仮説は有意水準αで棄却される。

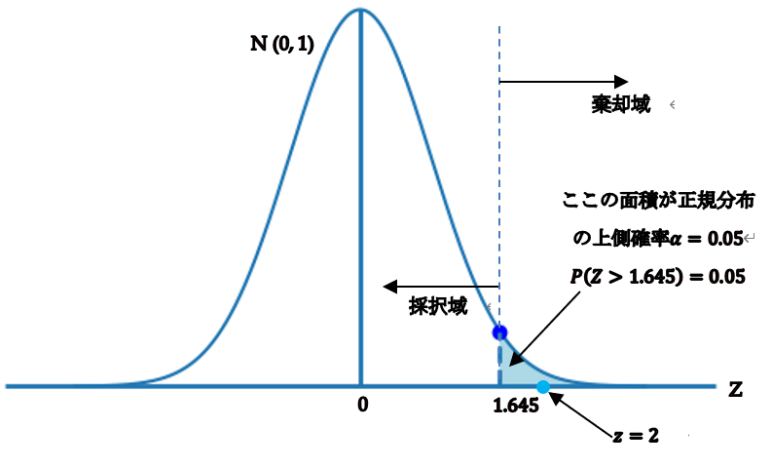

Figure 7:

右片側検定、左片側検定、両側検定：

$$\begin{cases} H_1 : \mu > \mu_0 (\text{右片側検定}) \\ H_1 : \mu < \mu_0 (\text{左片側検定}) \\ H_1 : \mu \neq \mu_0 (\text{両側検定}) \end{cases}$$

例題3の解:右片側検定である。棄却域は図7を参照。検定統計値は

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{18 - 17}{2/\sqrt{16}} = 2$$

有意水準の$\alpha = 0.05$とすると、棄却点は$z_\alpha = 1.645$であり、検定統計値のほうは棄却点より大きい、棄却域に入る。従って、帰無仮説は有意水準0.05で棄却される。

3.7 条件付き確率

条件付き確率を説明する前に、まず確率の加法定理と乗法定理を示す：

$$\text{確率加法定理：} P(A \cup B) = P(A) + P(B) - P(A \cap B)$$
$$\text{確率乗法定理：} P(A \cap B) = P(B) \times P(A|B)$$

条件付き確率とは、ある事象Aが起こったという条件のもとでの事象Bの起こる確率$P(B|A)$のことを「Aを与えた時のBの条件付き確率」という。 $P(B|A)$は$P_A(B)$とも表記する。「P, B given A」と読む。

$$P(B|A) = \frac{n(A \cap B)/N}{n(A)/N} = \frac{P(A \cap B)}{P(A)} \tag{1-32}$$

サイコロの例：

$$\begin{cases} \text{事象A:偶数の目が出る確率は} P(A) = 1/2 \\ \text{事象B:4以上(4を含め)の目が出る確率は} P(B) = 1/2 \end{cases}$$

P(B | A)とは事象A偶数の目が出たという条件のもとで、事象Bそれが4以上の目である確率のこと。
まず、AとB両方満たす確率を計算すれば、$\Rightarrow P(A \cap B) = \frac{2(\text{出た目が4か6か})}{6} = 1/3$となる。

$$P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{1/3}{1/2} = 2/3$$

例：下記の情報からある学生が工学部の学生でしかも数学好きと答えた確率は？

	P(EC)経済学部	P(EN)工学部	P(SC)理学部
在籍人数	200	500	300
P(S)数学が好き	30	60	70
P(K)数学が嫌い	70	40	30
調査対象の人数の合計	100	100	100

Table 2: 数学が好きかどうかの調査データ

まず、答えた学生は工学部に所属する確率を計算すれば、$P(EN) = 0.5$
次に、工学部の学生の中に数学が好きと答えた確率を計算すれば、$P(S|EN) = 0.6$
乗法定理を使うと、$P(S \cap EN) = P(EN) \times P(S|EN) = 0.5 \times 0.6 = 0.3$

3.8 ベイズ定理

ベイズ定理とは、$P(B|A)$の逆確率である$P(A|B)$は下記の式で求められる。

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)} \tag{1-33}$$

Where:

$P(A|B)$: 事後確率
$P(B|A)$: 尤度
$P(B)$: 事前確率

例題**4**：下記の条件で、煙が上がったのを見た時、その原因が火事である確率は？

$$\begin{cases} \text{事象A:火事が起こる確率は} P(fire) = 0.01 \\ \text{事象B:煙が上がることを見る確率は} P(smoke) = 0.1 \\ \text{火事の90\%で煙が上がる：} P(smoke|fire) = 0.9 \end{cases}$$

解

$$\begin{aligned} P(fire|smoke) &= \frac{P(smoke|fire) \times P(fire)}{P(smoke)} \\ &= \frac{0.9 \times 0.01}{0.1} \\ &= 0.09 \end{aligned}$$

例題**5**：下記の条件で、無作為に選んだメールが『キャンペーン』という単語を含んでいたという条件のもとで、それが迷惑メールである確率は？**(spam**は迷惑メール**)**

$$\begin{cases} \text{事象A:メールが迷惑メールの確率は} P(spam) = 0.2 \\ \text{事象B:キャンペーンという単語を含める確率は二つの場合ある：} \\ \begin{cases} \text{迷惑メール(0.2)の条件の下で『キャンペーン』という単語を含める確率は0.3:} P(campaign|spam) = 0.3 \\ \text{一般メール(0.8)の条件の下で『キャンペーン』という単語を含める確率は0.04:} P(campaign|NotSpanEmail) = 0.04 \end{cases} \end{cases}$$

解

$$\begin{aligned} P(campaign) &= P(campaign|spam) \times P(spam) + P(campaign|NotSpanEmail) \times P(NotSpanEmail) \\ &= 0.2 \times 0.3 + 0.8 \times 0.04\Gamma \\ &= 0.092 \\ P(spam|campaign) &= \frac{P(campaign|spam) \times P(spam)}{P(campaign)} \\ &= \frac{0.3 \times 0.2}{0.092} \\ &\approx 0.652 \end{aligned}$$

Submitted by J. Shang on 01/07/2024.